令和元年度特許庁委託事業

令和元年度

最新の OCR 技術による公開特許公報等のテキストデータ作成等を通じた機械翻訳サービスの充実化に向けた調査事業

調査報告書

令和 2 年 3 月 18 日

凸版印刷株式会社

目次

本事業の概要

第 1 章　事業の目的

本事業では、高精度なテキストデータを特許庁が保有していない古い国内の特許及び実用新案の各種公報のイメージデータを対象に、そのフォーマット等の特徴を分析したうえで、最新の既存 OCR ソフト及び人工知能技術を活用した OCR 技術（以下、AI-OCR）等によって当該イメージデータのテキスト化を実施する場合の精度等を調査・分析し、得られたテキストデータを利用して機械翻訳サービスを充実化するにあたっての課題等を整理することを目的とする。

第 2 章　調査用データの準備

調査用データとしては、特許庁から貸与を受けた約 1,547 万件の特許及び実用新案の各種公報のイメージデータから、文書種別・年代の分散を考慮して、8,098 件をサンプルデータとして抽出し、それをもとにフォーマット調査を実施した。さらに、既存 OCR ソフト及び AI-OCR による文字認識等の実験を行うために、実験用データとして、サンプルデータから 2,019 件を抽出し、そのイメージデータに対応するテキストデータ等を作成した。実験用データは、1,797 件を AI 用の教師データ、残りの 222 件を検証用データとした。

第 3 章　公開特許公報等のフォーマット等の調査

第 2 章で抽出したサンプルデータを目視で分析し、各種公報における、制度変更等による大きなフォーマットの変更（表組、版面構成、段組み、縦書き・横書き・新旧仮名遣い等）が生じた年代を特定した。

また、上記目視による分析中に、OCR ソフトによるテキスト化の阻害要因となり得る、多面付けや多段組み、文字のかすれやつぶれ、文字領域の傾き等の要素が、各種公報の種別及び発行年代ごとに異なって存在していることが判明した。

そして、上記フォーマット及び要素等に基づいて、各種公報を、同種の OCR ソフトによりテキスト化が行い得る単位として、公報種別及び発行年代別に、細かく 32、大きく 3 に区分した。

第 4 章　OCR ソフトの調査

第 3 章における区分毎に、第 1 章で作成した検証用データのイメージデータを既存 OCR ソフトによってテキスト化した場合における、文字認識、レイアウト認識（文字領域の検出）、レイアウト解析等について評価を行った。

文字認識精度及び処理能力については、区分毎の最適な既存 OCR ソフトが判明し、区分によっては、既存の OCR ソフトで十分な精度が達成できることが判明した。

また、レイアウト認識及びレイアウト解析については、「発明の名称」等の書誌情報項目

を既存の OCR ソフトで認識することは困難であることが判明した。

第 5 章　テキスト化精度向上策の検討・評価

既存 OCR ソフトを利用した文字認識精度の向上策として、①多数決処理、②レイアウト認識と文字認識の最適組み合わせ、及び、③文脈補正の 3 つの手段について検証した。

そして、いずれの手段を用いても劇的なテキスト化精度の向上は見込めず、テキスト化精度の向上には、第 3 章における区分毎に最適な既存 OCR ソフトを採用することが最も効果的であることが判明した。

第 6 章　人工知能技術を活用した OCR の検証

第 1 章で作成した検証用データのイメージデータに対して AI-OCR によるテキスト化を行った結果、文字認識精度（行単位）が平均 93.6%となり、既存 OCR ソフトによるテキスト化を行った場合（最も精度が高いソフトで平均 91.4%）よりも高い精度となった。

また、AI－OCR の文字認識精度は、本事業で使用した教師データ数の 7 倍程度の数の教師データを追加で準備し、AI に学習させれば、99%の認識精度に達すると推定された。

さらに、現状の文字認識精度を低下させている 6 の原因について、その対応策を検討した。これら対応策の実証により更なる認識精度の向上が期待できる。

また、文字認識の精度向上には、前処理として画像補正をすることが効果的であり、特に、人工知能を活用た高解像度化及びノイズ除去が最も効果的であることが判明した。

第 7 章　総合分析

第 1 章から第 6 章の結果を受け、今後の各種公報のイメージデータをテキスト化する場合におけるテキスト化の精度向上策を、第 3 章で区分した 32 及び 3 の区分毎に検討した。

区分毎にテキスト化を実施する場合における、既存 OCR ソフトおよび AI-OCR の使い分けや多数決処理等によるテキスト化の精度向上策を、第 3 章で区分した 32 及び 3 の区分毎に検討した。

既存 OCR ソフトは、極端に文字認識精度が悪い区分が散在したのに対し、AI-OCR は、ほぼすべての区分において安定して高精度の文字認識精度を発揮した。特に旧字旧仮名遣いの古い年代（1920～1940 年代）又は画質が悪い区分の各種公報のイメージデータについては、AI-OCR が既存 OCR ソフトに比べて、高い文字認識精度を示した。

そして、各種公報のイメージデータのテキスト化に際して AI-OCR と既存 OCR ソフトを使い分ける際には、第 3 章で区分した 32 の区分のうち、旧字旧仮名遣いの要素が存在する区分及び新字新仮名遣いの要素を有し且つ画質が悪い区分は AI-OCR で、新字新仮名遣いでかつ画質が良いデータは既存 OCR で処理する方法が効果的と判明した。

また、既存 OCR ソフト 2 種と AI-OCR を併用し、多数決処理を行うことにより、文字認識精度がさらに向上する区分が存在することも判明した。

結果として、各種公報のイメージデータをテキスト化するにあたっては、上述した 32 又は 3 区分毎に、AI-OCR 又は既存 OCR ソフトを適切に使い分ける方法が最も効果的であること、AI－OCR は、各区分に対する文字認識精度の差異が小さく、教師データの増強や各種改善策の実施によって更なる精度向上が期待できることが判明した。

そして、各種公報のイメージデータをテキスト化するにあたって、[Ⅰ] 旧字旧仮名遣い又は[II] 新字新仮名遣い且つ画質が悪い要素を有する区分には、AI－OCR を適用し、[III] 新字新仮名遣い且つ画質が良い区分には、既存 OCR ソフトを適用する方法が効果的であると判明した。また、今後、高精度のテキストデータが必要な約 1,500 万件（約 1 億画像ファイル）の各種公報のイメージデータを、人手による修正をせず、当該方法によってテキスト化する場合のコストおよび想定認識精度を、レイアウト解析（ラベル抽出と認識）を含めて概算で算出した。

また、本事業の目的である機械翻訳における使用について考察すると、高精度な機械翻訳を可能とするためには、平均的な文字認識率の向上に加えて、以下の自動翻訳特有の課題が想定される。

1）行のつながりを正しく推定すること。

OCR のレイアウト認識の際、行矩形の認識に成功したとしても、行矩形間のつながり方の推定に失敗すると、正しい機械翻訳の結果は得られない。

2）文・文節の区切りを正しく識別すること。

OCR は単純な記号の識別に失敗する傾向があるが、句読点の識別に失敗すると、文・文節の区切りの判断ができず、正しい翻訳翻訳の結果は得られない。

上記事例のように、少数の誤認識箇所が翻訳精度に大きな影響を及ぼすようなケースを分析し、重点的にチューニングすることができれば、特許庁がテキストデータを保有していない又は高精度なテキストデータを保有していない特許及び実用新案の各種公報について、その高精度な機械翻訳文の提供をすることが可能であると考えられる。

第１章　事業の目的

1．本事業の目的

本事業では、高精度なテキストデータを特許庁が保有していない古い国内の特許及び実用新案の各種公報のイメージデータを対象に、そのフォーマット等の特徴を分析したうえで、最新の既存 OCR ソフト及び人工知能技術を活用した OCR 技術（以下、AI-OCR）等によって当該イメージデータのテキスト化を実施する場合の精度等を調査・分析し、得られたテキストデータを利用して機械翻訳サービスを充実化するにあたっての課題等を整理することを目的とする。

2．本事業実施の背景

特許庁および独立行政法人工業所有権情報・研修館は、令和元年 5 月から、刷新された特許情報プラットフォーム（J-PlatPat）を通じて、最新の AI 技術を活用したニューラル機械翻訳によって翻訳した、国内の審査書類や各種公報等の機械翻訳文を、外部ユーザーに提供している。

外部ユーザーに十分な翻訳品質の機械翻訳文を提供するためには、翻訳対象としての高精度なテキストデータが必要である。しかし、特許及び実用新案の各種公報について、特許庁は、昭和 45 年以前に発行された公報のテキストデータを保有していない。また、昭和 46 年から平成 4 年に発行された公報について、特許庁は、OCR ソフトによってテキスト化したテキストデータを保有しているが、その多くは、人手による修正が行われていないテキスト化精度が低いものである。

したがって、テキストデータを保有していない又は高精度なテキストデータを保有していない各種公報についても、十分な翻訳品質の機械翻訳文を外部提供するためには、そのテキストデータが必要である。

加えて、各種公報のテキストデータは、特許庁審査官や外部ユーザーがテキスト検索によって効率的かつ網羅的に先行技術調査を行う環境整備という観点からも必要である。

しかし、テキストデータを保有していない又は高精度なテキストデータを保有していない特許及び実用新案の各種公報は、1,547 万件にのぼることから、そのイメージデータのテキスト化の順序及び方法が大きな課題である。

他方、近年、ディープラーニングや機械学習を通じた人工知能技術を活用することにより、既存 OCR ソフトに比べて飛躍的にテキスト化の精度を向上させた AI-OCR という技術が実用化されつつある。

AI-OCR は、人工知能による画像補正やレイアウト解析、文脈補正、文字認識精度の向上等を行うことで、既存 OCR ソフトでは高精度なテキスト化ができなかったイメージデータについても、高精度なテキスト化を可能とし得るものであるから、各種公報のイメージデータのテキスト化に有用であると期待される。しかし、AI-OCR については、現時点で達成可能なテキスト化の精度や AI-OCR に適したイメージデータの属性・性質等、未知な点が多

い。

そこで、本事業では、各種公報のイメージデータのテキスト化における、AI-OCR の有用性についても、検証項目とする。

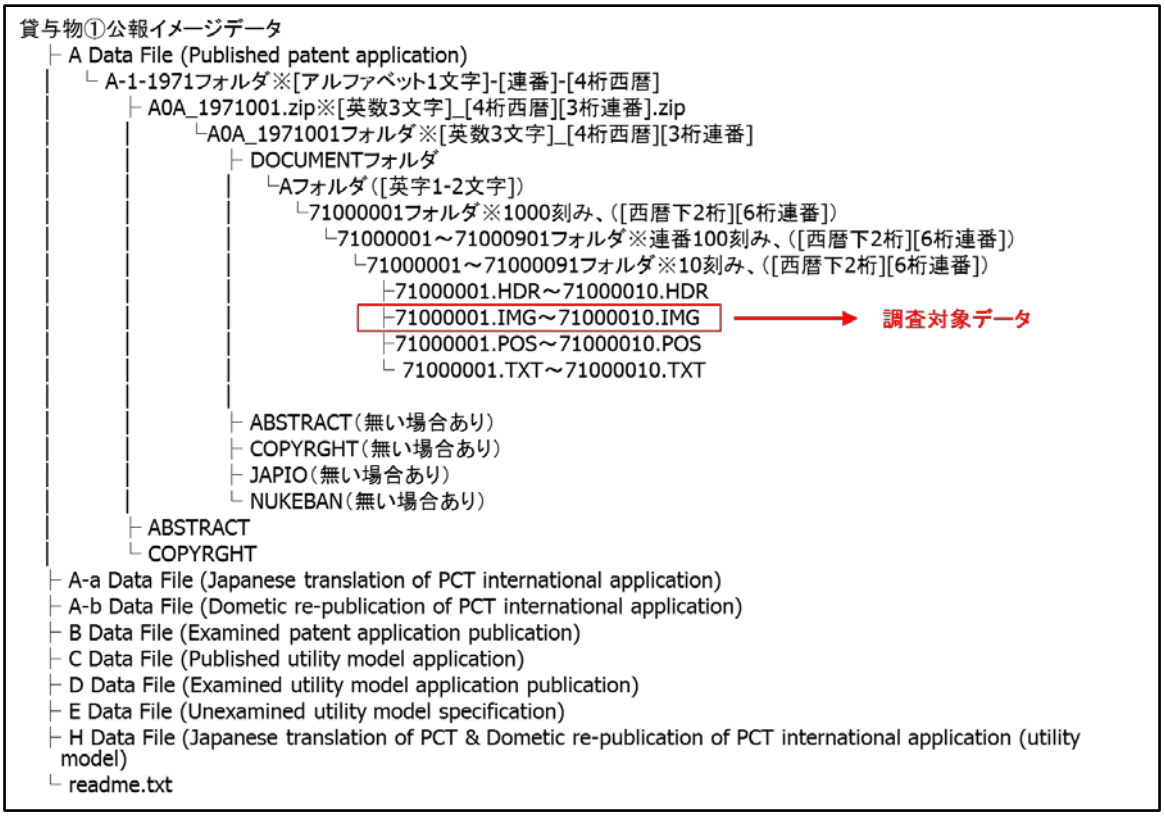
第 2 章　調査用データの準備

特許庁から貸与を受けた貸与データから、目視によるフォーマット調査用のサンプルデータ、並びに、OCR ソフト及び AI-OCR による評価、検証用の実験用データ等を抽出した。以下、貸与データの構成及び実験用データの要件と内容を示す。

1. 貸与データの構成

本事業において、特許庁から特許及び実用新案の各種公報のデータとして、約 1,547 万件の貸与を受けた。この全データサイズは、3.15 テラバイトに及ぶ。当該データに含まれる画像形式は、マルチページ TIFF、モノクロ二階調であり、CCITT T.6＋ZIP 形式で圧縮されていた。

図 2-1　貸与データのフォルダ構成



貸与データは、公報種別がアルファベットで区分され、さらに西暦の作成年次が 4 桁で記載され、そのあと公報の番号が記載されるという形で整理されていた。例えば、公報種別について、公開特許公報は A、公表特許は A-a と区分されている。

2. サンプルデータの抽出

目視によるフォーマット調査用のサンプルデータとして、貸与データ約 1,547 万件より 8,098 件を抽出した。

公報の仕様が変更された推定される年代を考慮し、まず種別ごとに分類し、さらに公報種別毎に、フォーマットの変遷を考慮し、各フォーマットに対して十分な調査が可能なようにサンプルデータを抽出した。

※参考資料：特許庁編『工業所有権制度百年史』

表 2-1　サンプルデータ

A	公開特許	1971-1992	1,001 件
A-a	公表特許	1979-1995	701 件
A-b	再公表特許	1979-1995	700 件
B	公告特許	1922-1993	1,754 件
C	公開実用新案	1971-1992	1,001 件
H-a	公表実用新案	1979-1995	166 件
H-b	再公表実用新案	1981-1992	19 件
D	公告実用新案	1922-1993	1,754 件
E	公開実用新案全文	1971-1992	1,002 件

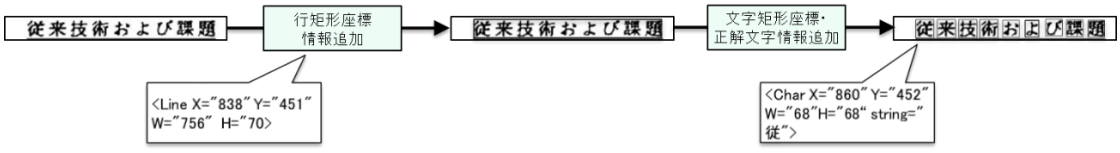
8,098 件

3. 実験用データの抽出

OCR ソフト及び AI-OCR による評価や検証用の実験用データとして、サンプルデータ 8,098 件から 2,019 件を抽出した。

抽出した実験用データには OCR ソフトの調査用の正解データ、及び AI-OCR 用の教師データとして使用するため、1 行ごとに座標情報、1 文字ごとに座標情報と正解文字情報の正解データを手入力により付与した。このような実験用データの例は別紙 4 を参照されたい。

図 2-2　実験用データの座標情報及び正解文字情報の付与



4. 検証用データの抽出

実験用データ 2,019 件のうち 1,797 件については AI-OCR を生成する際の教師データとし、残りの 222 件を OCR ソフト及び AI-OCR の検証用データとした。

サンプルデータ、実験用データおよび検証用データの種別内訳は以下の通りである。

表 2-2　サンプルデータ、実験用データおよび検証用データの種別内訳

識別区分	種別	年代	サンプルデータ	実験用データ	検証用データ
A	公開特許	1971-1992	1,001	207	28
A-a	公表特許	1979-1995	701	107	15
A-b	再公表特許	1979-1995	700	105	13
B	公告特許	1922-1993	1,754	504	58
C	公開実用新案	1971-1992	1,001	201	29
H-a	公表実用新案	1979-1995	166	166	15
H-b	再公表実用新案	1981-1992	19	19	1
D	公告実用新案	1922-1993	1,754	502	47
E	公開実用新案全文	1971-1992	1,002	208	16
合計	－	－	8,098	2,019	222

サンプルデータ 8,098 件は、第 3 章の「公開特許公報等のフォーマット等の調査」に使用した。実験用データ 2,019 件は、そのうち 1,797 件を第 6 章で述べる AI-OCR 用の教師データとして使用し、残り 222 件を、検証用データとして第 4 章～第 6 章の OCR ソフト調査および AI-OCR 調査で使用した。

第 3 章　公開特許公報等のフォーマット等の調査

1. 調査方針

平成 4 年度以前のテキストデータを保有していない、またはテキスト化精度が低い公開特許公報等について、貸与データから抽出した 8,098 件のサンプルデータに対して、目視による公開種別や発行年代によるフォーマットや活字字体等の調査を実施した。

合わせて、サンプルデータに対して、 OCR ソフトによるテキスト化阻害要因となり得る要素に関しても調査を実施した。

2. 調査対象

公開種別や発行年代によるフォーマット調査対象は下記の通りとした。

表 3-1　調査対象

A	公開特許	1971-1992	1,001 件
A-a	公表特許	1979-1995	701 件
A-b	再公表特許	1979-1995	700 件
B	公告特許	1922-1993	1,754 件
C	公開実用新案	1971-1992	1,001 件
H-a	公表実用新案	1979-1995	166 件
H-b	再公表実用新案	1981-1992	19 件
D	公告実用新案	1922-1993	1,754 件
E	公開実用新案全文	1971-1992	1,002 件

8,098 件

3. フォーマットの変遷に関する調査

(1) 調査手順

① 公報種別ごとに年代にしたがって公報の画像を目視で確認することでフォーマットが切り替わったと思われる発行日を推定し、その推定発行日によって公報種別ごとに発行年代を分類したものを「小区分」とする。また、特に大きなフォーマットの変更と思われる発行日で公報種別ごとに発行年代を分類したものを「区分」とする。

② 上記「区分」ごとに、下記 a)に示すフォーマット・活字字体の特徴を調査した。

③ 下記 b-1)に示すように OCR ソフトによるテキスト化を行う際にテキスト化精度に大きな影響を与えると考えられる要因および下記 b-2)に示すようにその他テキスト化精度に影響を与えると考えられる要因を設定し、上記「区分」ごとに、その要因を有している公報の割合を目視で調査した。

a）フォーマット・活字字体の特徴

図 3-1　4 面付（多面付）

1 枚の紙面に 4 段の文章がある。

図 3-2　2 段組

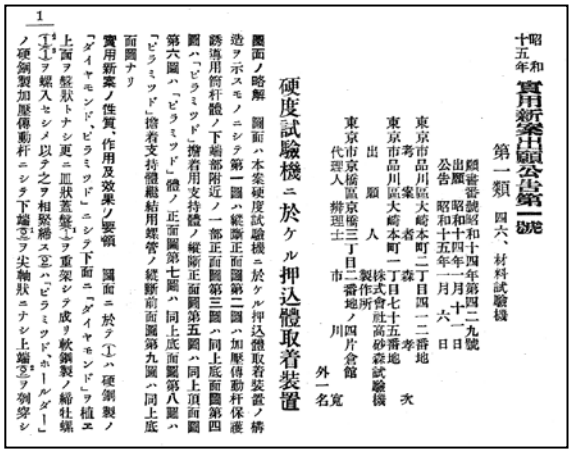
1 枚の紙面に本文が 2 段で記載されている。

特許公報　昭27-3

計算機

図 3-3　縦書き

天地方向に文字が記載されている。



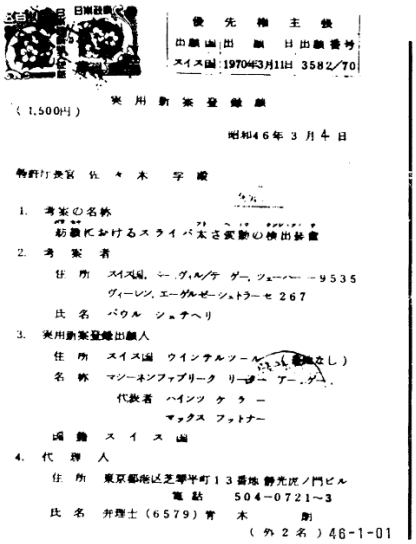
昭和十五年　實用新案出願公告第一號

第一類　四六、材料試験機

硬度試験機ニ於ケル押込體取着装置

図 3-4　（参考）横書き

左右方向に文字が記載されている。



実用新案登録願

昭和46年3月4日

図 3-5　印紙

印紙が貼られている。

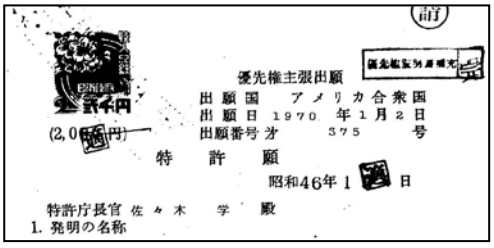
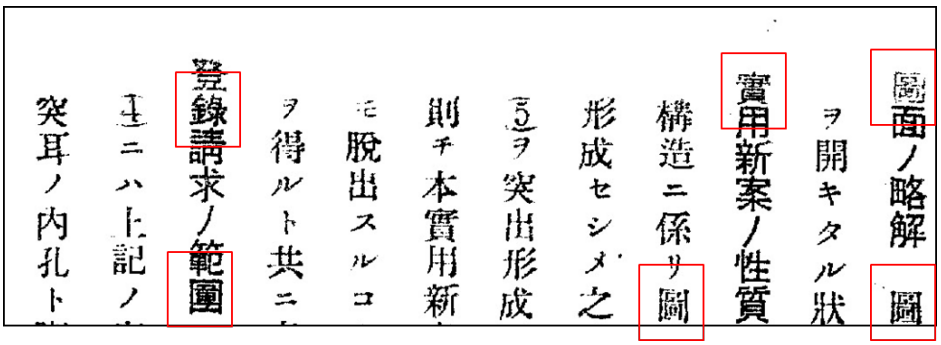


図 3-6　押印

判が押されている。

図 3-7　旧字旧仮名遣い

旧仮名遣いの文章。



b-1）テキスト化精度に大きな影響を与えると考えられる要因

図 3-8　文字のかすれ
文字がかすれている。

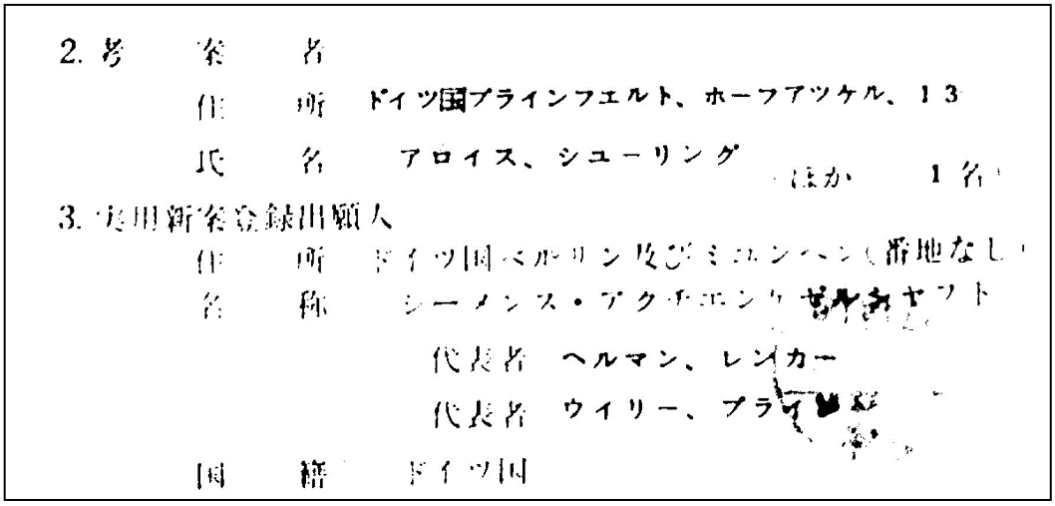
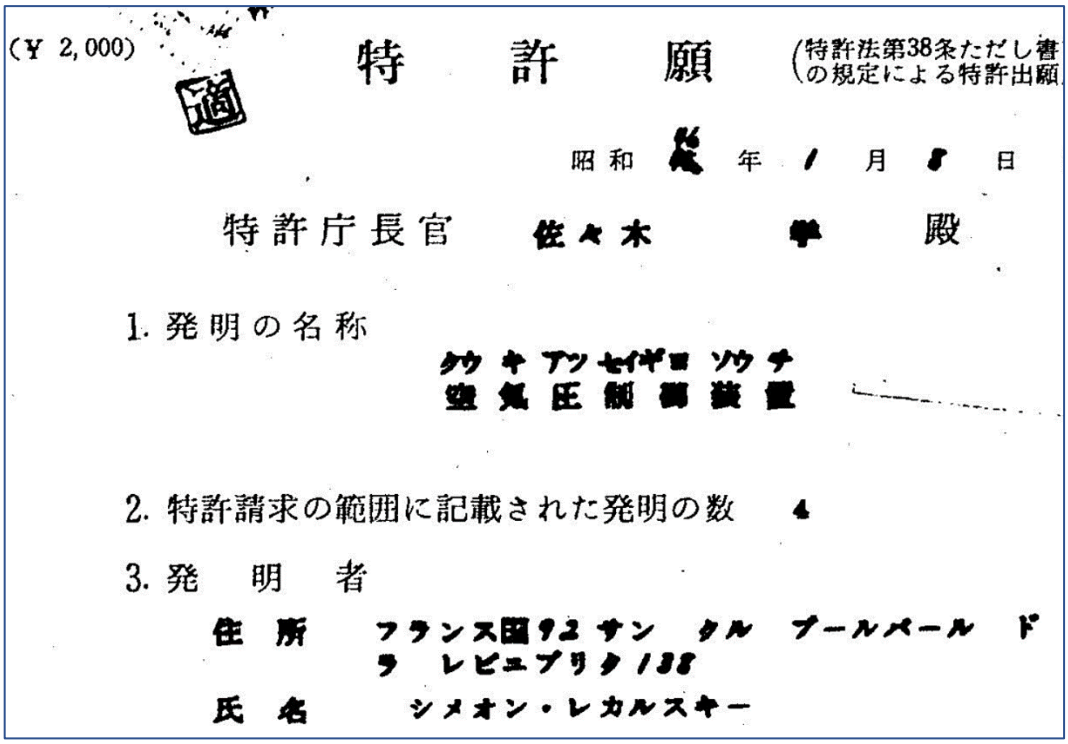


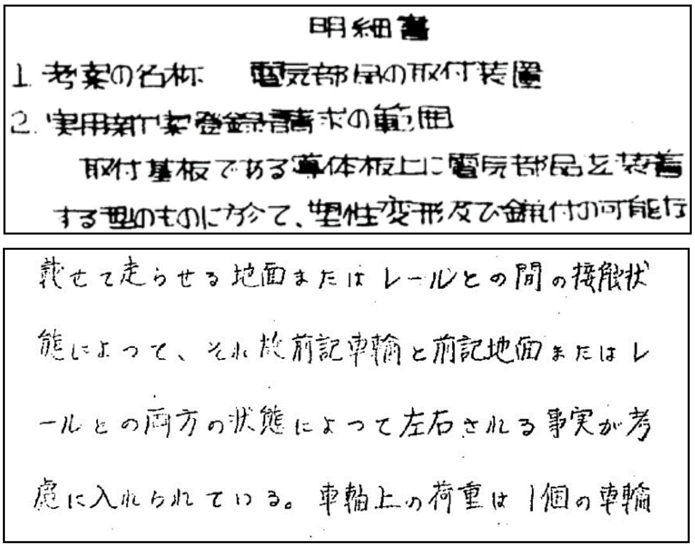
図 3-9　文字のつぶれ
文字がつぶれている。



b-2）その他テキスト化精度に影響を与えると考えられる要因

図 3-10　手書き

手書きで記載されている。



明細書

1. 考案の名称　電気部品の取付装置

2. 実用新案登録請求の範囲

取付基板である導体板上に電気部品を装着

する型のものに於て、塑性変形及び鋲付の可能な

載せて走らせる地面またはレールとの間の接触状

態によって、それ故前記車輪と前記地面またはレ

ールとの両方の状態によって左右される事実が考

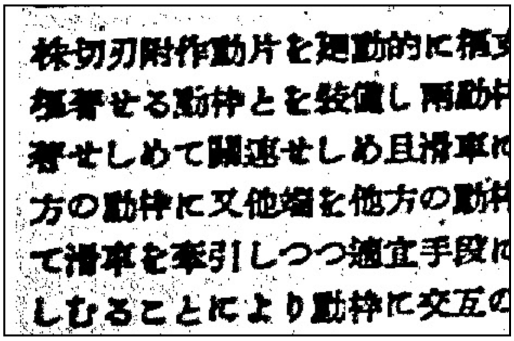
慮に入れられている。車軸上の荷重は1個の車輪

図 3-11　傾き

紙面が傾いている。

図 3-12　ノイズ

紙面にノイズがある。



株切刃附作動片を運動的に植立

係着せる動枠とを装備し両動枠

着せしめて関連せしめ且滑車に

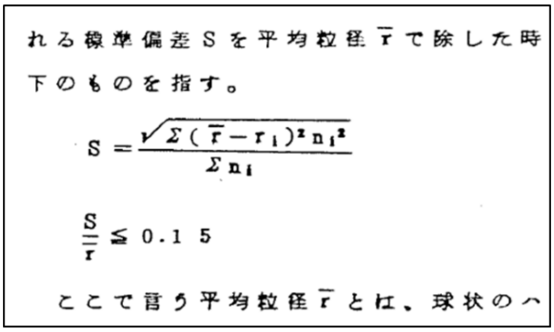
方の動枠に又他端を他方の動枠

て滑車を牽引しつつ適宜手段に

しむることにより動枠に交互の

図 3-13　数式

文中に数式がある。



れる標準偏差Sを平均粒径 $\bar{r}$ で除した時

下のものを指す。

$$S = \frac{\sqrt{\Sigma(\bar{r} - r_i)^2 n_i^2}}{\Sigma n_i}$$

$$\frac{S}{\bar{r}} \leq 0.15$$

ここで言う平均粒径 $\bar{r}$ とは、球状のへ

図 3-14　化学式

文中に化学式がある。

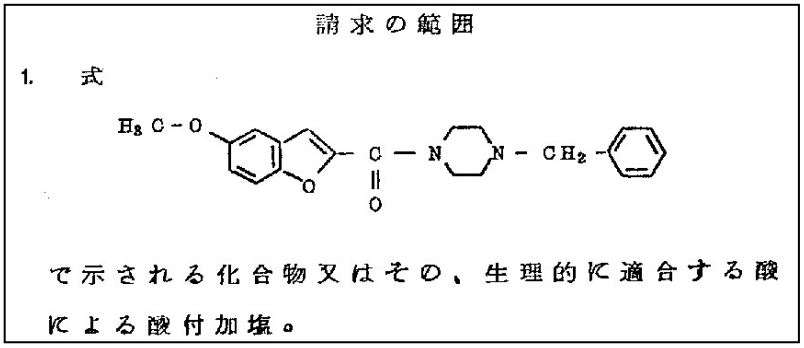


図 3-15　ルビ

文中にルビが振られている。

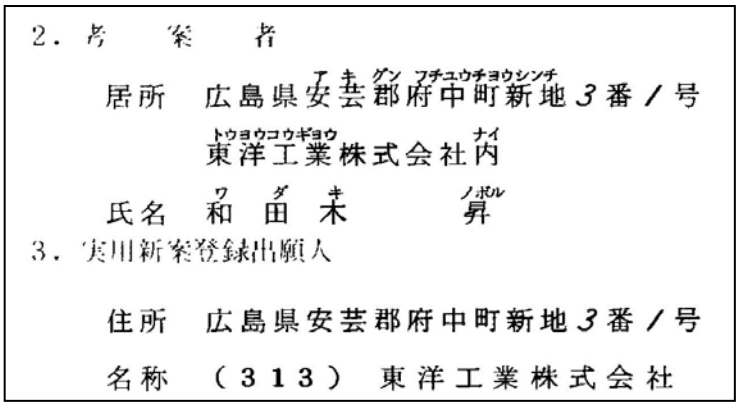


図 3-16　表組

表で表示されている。

第５表

試　料		青色光測定	緑色光測定	赤色光測定
9（比　較）	相対感度	100	100	100
	L.E.S.	2.59	2.66	2.68
10（比　較）	相対感度	97	101	109
	L.E.S.	2.58	2.67	2.72
11（比　較）	相対感度	100	100	115
	L.E.S.	2.57	2.65	2.73
12（本発明）	相対感度	101	100	115
	L.E.S.	2.57	2.65	2.88
13（本発明）	相対感度	100	99	116
	L.E.S.	2.58	2.68	2.95
14（本発明）	相対感度	100	101	115
	L.E.S.	2.60	2.68	2.96
15（本発明）	相対感度	112	115	116
	L.E.S.	2.92	3.01	3.01

第５表から明らかなように、本発明に係る試料１２～１５と比較して、多分散性ヘロゲン化銀乳剤を使用した比較試料９は、赤色光でのL.E.S.値が小さいものであることが確認された。又、単分散性ヘロゲン化銀乳剤を使用した比較試料１０及び１１は、多分散性ヘロゲン化銀乳剤を使用した比較試料９に比べると赤色光でのL.E.S.値は向上しているものの、本発明試料と比較すると赤色光でのL.E.S.値が小さいものであることが確認された。即ち、本発明試料は、露光域の拡大という点において優れたものであることが確認された。

図 3-17　ブロック囲み罫

紙面に囲み罫がひかれている。

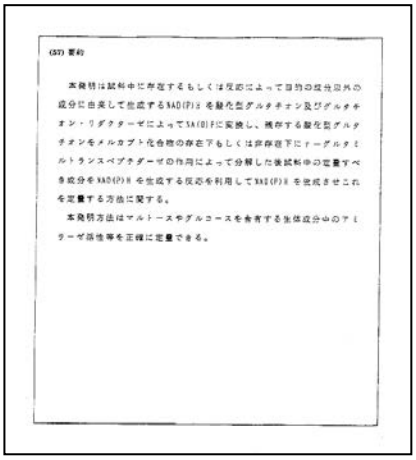
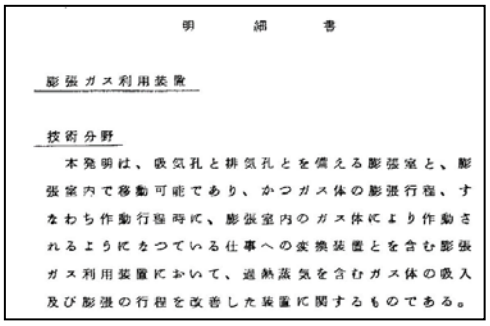


図 3-18　本文強調下線

文中に強調線が引かれている。



明　　細　　書

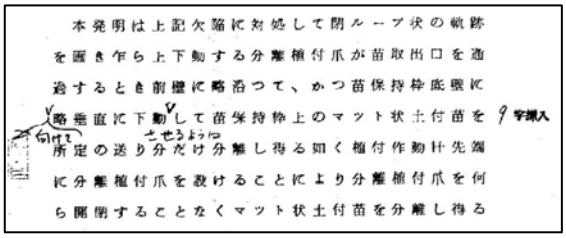
膨張ガス利用装置

技術分野

本発明は、吸気孔と排気孔とを備える膨張室と、膨張室内で移動可能であり、かつガス体の膨張行程、すなわち作動行程時に、膨張室内のガス体により作動されるようになっている仕事への変換装置とを含む膨張ガス利用装置において、過熱蒸気を含むガス体の吸入及び膨張の行程を改善した装置に関するものである。

図 3-19　手書書込み

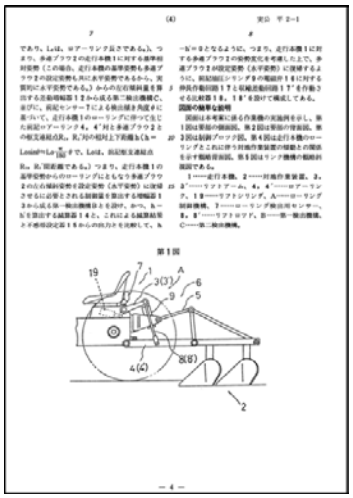
紙面に手書きで追加書き込みがある。



本発明は上記欠陥に対処して閉ループ状の軌跡を描き乍ら上下動する分離植付爪が苗取出口を通過するとき前壁に略沿つて、かつ苗保持枠底壁に略垂直に下動して苗保持枠上のマット状土付苗を所定の送り分だけ分離し得る如く植付作動杆先端に分離植付爪を設けることにより分離植付爪を何ら開閉することなくマット状土付苗を分離し得る

図 3-20　レイアウト（テキスト/図版混在）

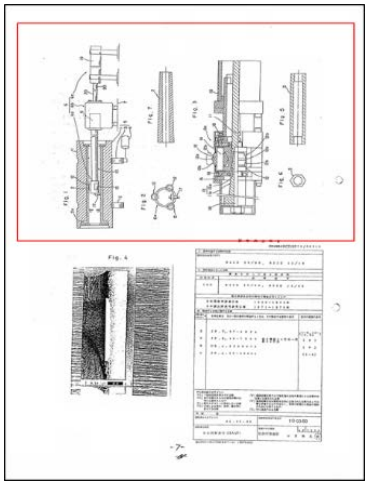
紙面にテキスト、図版、表組などが混在している。



第1図

図 3-21　90°回転

紙面が 90°回転している。



(2) 調査結果

「a) フォーマット・活字字体の特徴」及び「b-1) テキスト化精度に特に大きな影響を与えると考えられる要因」についての調査結果を以下に示す。

なお、各小区分に属する公報の例については、別紙 1 を参照されたい。

また、「b-2) その他テキスト化精度に大きな影響を与えると考えられる要因」についての調査結果は別紙 2 を参照のこと。

以下では、文字のつぶれ又は文字のかすれが、区分内の資料の 50%以上に出現する場合、「画質が悪い」に✔を付与している。

表 3-2　A）公開特許

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
A-①	1	フォーマットA-①-1	1971/7/16	1975/3/17	発行開始	350	○	×	—	○	○	—	55.2%	81.7%	✔
	2	フォーマットA-①-2	1975/3/18	1977/6/30	Int.Cl2.追加										
A-②	1	フォーマットA-②-1	1977/7/1	1979/12/27	書誌事項全体のレイアウトが変わる	350	○	×	—	×	×	—	7.4%	51.7%	✔
	2	フォーマットA-②-2	1980/1/5	1985/2/13	日本分類の記載が無くなる										
A-③	1	フォーマットA-③-1	1985/2/14	1992/7/30付近	発明の数、審査請求の位置が移動 書誌事項が1段表記になる	300	○	×	—	×	×	—	0.0%	6.7%	
A-④	1	フォーマットA-④-1	1992/7/30付近		現行公開公報に近い形になる	1	×	○	—	×	×	—	0.0%	0.0%	

表 3-3　A-a）公表特許

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
A-a-①	1	フォーマットA-a-①-1	1979/7/26	1979/12/27	発行開始	350	○	×	—	×	○	—	26.6%	22.9%	
	2	フォーマットA-a-①-2	1980/1/10	1983/7/7	日本分類の記載が無くなる										
	3	フォーマットA-a-①-3	1983/7/14	1984/12/27	「予備審査請求」の欄が登場										
A-a-②	1	フォーマットA-a-②-1	1985/1/10	1993/12/22	「審査請求」、「予備審査請求」の欄が移動 書誌事項が1段表記になる	350	○	×	—	×	×	—	12.0%	27.4%	
A-a-③	1	フォーマットA-a-③-1	1994/1/6		現行実用新案公報に近い形になる	1	○	×	—	×	×	—	0.0%	0.0%	

表 3-4　A-b）再公表特許

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
A-b-①	1	フォーマットA-b-①-1	1979/8/9	1979/12/6	発行開始	700	○	×	—	×	×	—	92.1%	99.9%	✔
	2	フォーマットA-b-①-2	1980/1/10	1983/7/7	日本分類の記載が無くなる										
	3	フォーマットA-b-①-3	1983/8/4		「審査請求」、「予備審査請求」の欄が登場										

表 3-5　B）公告特許

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
B-①	1	フォーマットB-①-1	1922/6/9	1922/11/17	発行開始	350	×	×	○	×	○	○	16.6%	0.3%	
	2	フォーマットB-①-2	1922/11/22	1923/3/7	「出願ノ要旨」が「發明ノ性質及ヒ目的ノ要領」と「特許請求ノ範圍（※「請」は右が「青」）」とに分かれる										
	3	フォーマットB-①-3	1923/3/9	1923/8/24	「願書番號」が「特許願番號」に変わる										
		公報無し	1923/8/25	1924/5/12	（関東大震災の影響）										
	4	フォーマットB-①-4	1924/5/13	1924/7/16	公告番号を初期化して公報発行再開 「願書番號」に戻る										
	5	フォーマットB-①-5	1924/7/18	1933/1/30	タイトル変更（「特許出願公告」の位置が変更） ※1927年～公告番号が年ごとになる										
	6	フォーマットB-①-6	1933/2/1	1934/2/9	ページ番号の位置が変わる										
	7	フォーマットB-①-7	1934/2/12	1936/3/30	「（特許局発行）」との記載が登場										
	8	フォーマットB-①-8	1936/4/1	1938/7/29	冊子全体の通番が登場										
B-②	1	フォーマットB-②-1	1938/8/1	1940/8/23	1段表記から2段表記に変わる	3	×	○	○	×	×	○	0.0%	0.0%	
	2	フォーマットB-②-2	1940/8/26	1940/11/30	「發明ノ性質及目的ノ要領」の記載が無くなる										
	3	フォーマットB-②-3	1940/12/5	1943/11/11	紙面のサイズが縮小する										
		公報無し	1943/11/12	1947/2/14	（戦時特例により出願公告制度が停止）										
	4	フォーマットB-②-4	1947/2/15	1947/12/26	公報発行再開 「（特許標準局発行）」との記載となる										
B-③	1	フォーマットB-③-1	1948/2/10	1949/12/23	1紙面に対し、複数文献が記載される	1	×	×	—	×	×	○	100.0%	0.0%	✔
B-④	1	フォーマットB-④-1	1950/1/9	1951/10/30	文献ごとの記載に戻る	350	×	○	—	×	×	△	29.4%	100.0%	✔
	2	フォーマットB-④-2	1951/11/4	1952/12/23	公報下部の書誌事項がなくなる 「日本國政府」→「特許庁」に表記が変更 「（全 〇 頁）」との記載が登場										
	3	フォーマットB-④-3	1953/1/6	1963/12/28	「発明の性質及目的の要領」の記載が無くなる										

B-⑤	1	フォーマット B-⑤-1	1964/1/9	1969/3/31	書誌事項が公報上部から本文に移動	700	×	○	—	×	×	—	0.0%	0.1%	
	2	フォーマット B-⑤-2	1969/4/2	1970/9/30	書誌事項に番号が導入される 欄の番号、発明の数の欄が登場										
	3	フォーマット B-⑤-3	1970/10/1	1972/12/27	Int.Cl.導入										
	4	フォーマット B-⑤-4	1973/1/5	1975/3/31	クレームに丸囲み文字が記載される										
	5	フォーマット B-⑤-5	1975/4/1	1977/3/31	庁内整理番号導入										
	6	フォーマット B-⑤-6	1977/4/1	1979/3/31	書誌事項が「特許公報」の下に移動する										
	7	フォーマット B-⑤-7	1979/4/2	1979/12/27	JP、Bの表示										
	8	フォーマット B-⑤-8	1980/1/5	1984/12/27	日本分類の記載が無くなる										
B-⑥	1	フォーマット B-⑥-1	1985/1/5	1993/12/24	書誌事項が1段表記になる	350	×	○	—	×	×	—	0.0%	0.0%	

表 3-6　C）公開実用新案

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
C-①	1	フォーマット C-①-1	1971/9/13	1972/12/27	発行開始	350	×	×	—	×	×	—	13.1%	2.0%	
	2	フォーマット C-①-2	1973/1/5	1975/3/24付近	クレームに丸囲み文字が記載される										
	3	フォーマット C-①-3	1975/3/24付近	1977/3/31	Int.Cl2.追加										
	4	フォーマット C-①-4	1977/4/1	1977/6/30	書誌事項が「公開実用新案公報」の下に移動する										
C-②	1	フォーマット C-②-1	1977/7/1	1979/3/31	「(全 ○ 頁)」との記載が登場 書誌事項が2段表記になる	350	×	×	—	×	×	—	4.0%	0.0%	
	2	フォーマット C-②-2	1979/4/2	1985/2/13	JP、Uの表示										
C-③	1	フォーマット C-③-1	1985/2/14	1989/7/14付近	審査請求の位置が移動 書誌事項が1段表記になる	300	×	×	—	×	×	—	3.7%	0.0%	
	2	フォーマット C-③-2	1989/7/14付近	1992/7/28付近	請求項の数の欄が登場										
C-④	1	フォーマット C-④-1	1992/7/28付近		現行実用新案公報に近い形に変わる	1	×	×	—	×	×	—	0.0%	0.0%	

表 3-7　H-a）公表実用新案

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
H-a-①	1	フォーマットH-a-①-1	1979/9/6	1979/10/25	発行開始	40	○	×	—	×	×	—	35.0%	100.0%	✔
	2	フォーマットH-a-①-2	1980/9/6	1983/6/23	日本分類の記載が無くなる										
	3	フォーマットH-a-①-3	1983/8/4	1983/12/20	「予備審査請求」の欄が登場										
H-a-②	1	フォーマットH-a-②-1	1985/1/17	1993/9/2	「審査請求」、「予備審査請求」の欄が移動 書誌事項が1段表記になる	111	○	×	—	×	×	—	25.5%	99.1%	✔
H-a-③	1	フォーマットH-a-③-1	1994/1/6		現行実用新案公報に近い形に変わる	15	○	×	—	×	×	—	0.0%	0.0%	

表 3-8　H-b）再公表実用新案

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
H-b-①	1	フォーマットH-b-①-1	1982/3/11	1982/3/11	発行開始	19	○	×	—	×	×	—	15.8%	31.6%	
	2	フォーマットH-b-①-2	1984/7/19		「審査請求」、「予備審査請求」の欄が記載される										

表 3-9　D）公告実用新案

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
D-①	1	フォーマットD-①-1	1922/6/1	1923/8/31	発行開始	350	×	×	○	×	○	○	37.7%	16.0%	
		公報無し	1923/9/1	1924/5/18	（関東大震災の影響）										
	2	フォーマットD-①-2	1924/5/19	1924/7/24	公告番号を初期化して公報発行再開										
	3	フォーマットD-①-3	1924/7/31	1933/1/31	タイトル変更（「実用新案出願公告」の位置が変更） ※1927年～公告番号が年ごとになる										
	4	フォーマットD-①-4	1933/2/2	1934/2/10	ページ番号の位置が変わる										
	5	フォーマットD-①-5	1934/2/13	1936/3/31	「（特許局発行）」との記載が登場										
	6	フォーマットD-①-6	1936/4/2	1938/1/29	冊子全体の通番が登場										
	7	フォーマットD-①-7	1938/2/1	1938/7/30	冊子全体の通番の位置が変わる										

D-②	1	フォーマット D-②-1	1938/8/2	1940/11/30	1段表記から2段表記に変わる	3	×	○	○	×	×	○	0.0%	0.0%	
	2	フォーマット D-②-2	1940/12/3	1943/11/13	紙面のサイズが縮小する										
		公報無し	1943/11/14	1947/2/27	（戦時特例により出願公告制度が停止）										
	3	フォーマット D-②-3	1947/2/28	1947/12/27	公報発行再開 「（特許標準局発行）」との記載となる										
D-③	1	フォーマット D-③-1	1948/2/10	1949/12/23	1紙面に対し、複数文献が記載される	1	×	○	—	×	×	○	0.0%	100.0%	✔
D-④	1	フォーマット D-④-1	1950/1/9	1951/10/30	文献ごとの記載に戻る	350	×	○	—	×	×	△	60.0%	14.3%	✔
	2	フォーマット D-④-2	1951/11/4	1963/12/28	公報下部の書誌事項がなくなる 「日本國政府」→「特許庁」に表記が変更 「(全 〇 頁)」との記載が登場										
D-⑤	1	フォーマット D-⑤-1	1964/1/9	1969/3/31	書誌事項が公報上部から本文に移動	700	×	○	—	×	×	—	0.6%	0.4%	
	2	フォーマット D-⑤-2	1969/4/2	1970/9/30	書誌事項に番号が導入される 欄の番号、発明の数の欄が登場										
	3	フォーマット D-⑤-3	1970/10/1	1972/12/27	Int.Cl.導入										
	4	フォーマット D-⑤-4	1973/1/5	1975/3/31	クレームに丸囲み文字が記載される										
	5	フォーマット D-⑤-5	1975/4/1	1977/3/31	庁内整理番号導入										
	6	フォーマット D-⑤-6	1977/4/1	1979/3/31	書誌事項が「実用新案公報」の下に移動										
	7	フォーマット D-⑤-7	1979/4/2	1979/12/27	JP、Yの表示										
	8	フォーマット D-⑤-8	1980/1/5	1984/12/27	日本分類の記載が無くなる										
D-⑥	1	フォーマット D-⑥-1	1985/1/5	1993/12/24	書誌事項が1段表記になる	350	×	○	—	×	×	—	0.0%	0.6%	

表 3-10　E）公開実用新案全文

区分No	小区分No	フォーマットNo	開始日	終了日	改訂箇所	サンプルデータ数	4面付	2段組	縦書き	印紙	押印	旧字旧仮名遣い	文字のかすれ	文字のつぶれ	画質が悪い
E-①	1	フォーマットE-①-1	1971/9/13	1983/3/31	発行開始	700	×	×	—	○	○	—	87.1%	86.7%	✔
E-②	1	フォーマットE-②-1	1983/4/1	1985/2/13	書誌事項の表記が公開公報と近い形式に変わる	1	×	×	—	×	×	—	0.0%	100.0%	✔
E-③	1	フォーマットE-③-1	1985/2/14	1989/7/14付近	審査請求の位置が移動 書誌事項が1段表記になる	300	×	×	—	×	×	—	18.7%	27.3%	
	2	フォーマットE-③-2	1989/7/14付近	1992/7/28付近	請求項の数の欄が登場										
E-④	1	フォーマットE-④-1	1992/7/28付近		現行実用新案公報に近い形に変わる	1	×	×	—	×	×	—	0.0%	100.0%	✔

＜旧字体から新字体への変遷について＞

公告特許の区分 B-④、及び公告実用新案の区分 D-④の期間で、本文の文字が徐々に旧字体から新字体に移行している。そこで、どの年代でこの以降が生じているのかを調査した。

表 3-11　区分別旧字体出現率

区分 No	年度	件数	出現率
B-④	1950	100	100%
	1951	100	100%
	1952	100	100%
	1953	50	0%
D-④	1951	100	100%
	1952	100	100%
	1953	50	0%

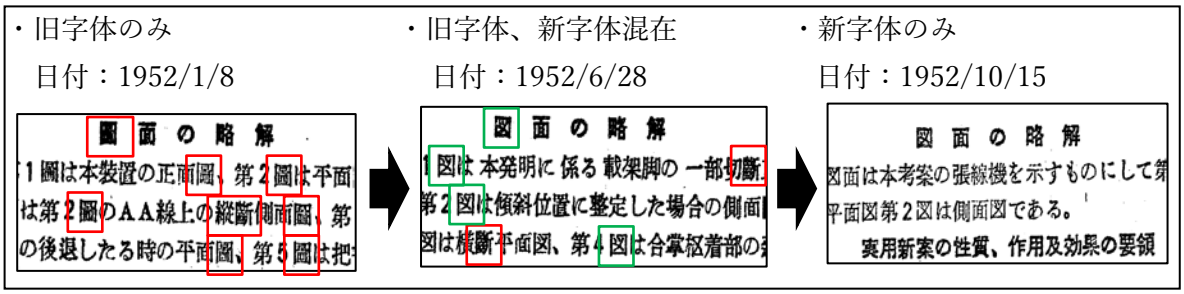
区分 B-④、D-④のサンプルデータはともに、1952 年の旧字体の出現率が 100%、1953 年の旧字体の出現率が 0%である。

したがって、旧字体から新字体への切り替えは 1952 年に実施されたと見なされる。

1952 年のデータを抽出して確認したところ、同年 5 月頃の資料から新字体が混在し始め、同年 8 月の資料では新字体のみで記載された資料が出現する。

同年 9 月以降も一部の文字が旧字体で記述された資料が出現するが、1953 年以降の資料では、旧字体で記載された資料は、ほぼ出現しない。

図 3-22　旧字体から新字体への変遷



4. 同種のエンジンによりテキスト化を扱うことができる公報種別・発行年代

3. で示したように、特に大きなフォーマットの変更と思われる分類である「区分」は 32 種類としたが、テキスト化精度に大きな影響を与える区分はさらに少ない数の区分に分類できるものと考えられる。そこで、次章以降では、テキスト化精度に大きな影響を与えるものと推測される区分は次の 3 区分となるものと仮説を立てて検証を行っていく。

① 旧字旧仮名遣い
② 画質が悪い資料（新字新仮名遣い）
③ 画質が良い資料（新字新仮名遣い）
※画質が悪い＝文字のかすれ又は文字のつぶれが区分内の 50%以上で出現と定義

これらの 3 つの区分とした理由は、以下の通りである。

- 既存 OCR が認識可能な漢字は JIS 第 2 水準の範囲である。一方、旧字には JIS 第 2 水準外の漢字が多数含まれているため、旧字のあるなしが認識に大きく影響する。
- サンプルによる予備調査の結果、文字のかすれ・つぶれにより、同一フォーマットであっても文字認識精度が大きく低下することが確認された。

同種のエンジンにおいてテキスト化を扱うことができると仮説を立てた 3 区分ごとに、該当する 32 種類の公報種別・発行年代の組み合わせを下記の通り示す。

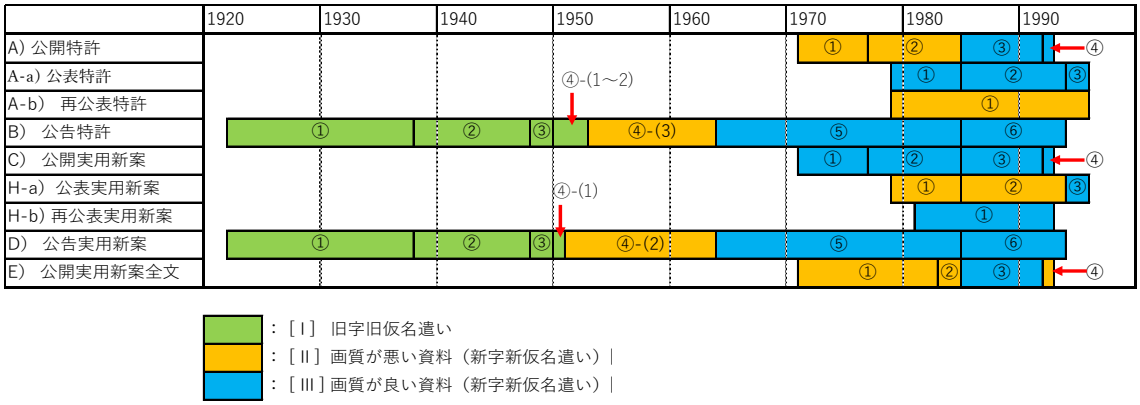
表 3-13　同種のエンジンによりテキスト化を扱うことができる公報種別

レイアウトの特徴	検証データ数	区分 No	種別	推定総件数
[I] 旧字旧仮名遣い	50	B-①②③, ④-（1～2） D-①②③, ④-（1）	公告特許 公告実用新案	51 万件
[II] 画質が悪い資料（新字新仮名遣い）	68	A-①② A-b-① B-④-（3） D-④-（2） E-①②④ H-a-①②	公開特許 再公表特許 公告特許 公告実用新案 公開実用新案全文 公表実用新案	473 万件
[III] 画質が良い資料（新字新仮名遣い）	104	A-③④ A-a-①②③ B-⑤⑥ C-①②③④ D-⑤⑥ E-③ H-a-③ H-b-①	公開特許 公表特許 公告特許 公開実用新案 公告実用新案 公開実用新案全文 公表実用新案、再公表実用新案	1,024 万件

※「A-b-①」再公表特許は、1995/10/5～18 の間で画質が良いデータに切り替わっている

上記 3 区分の年代の遷移を表したものを以下に示す。

表 3-14　公報種別区分遷移



5. 公報種別ごとの特徴

「3. フォーマットの変遷に関する調査」を踏まえて、公報種別ごとの特徴を以下に示す。

A	公開特許

1970 年代において、「多面付け」の資料が 90%を超え、1970 年代の資料は、ルビ、印紙、押印ありの割合も 90%前後と高い。一方で、1990 年代になってそれまで出現しなかった「多段組」の資料が出現し、その割合も 97%と高い。1980 年代までは 25%程度存在していた手書き資料は、1990 年以降は出現頻度が下がり、4%と減少した。全般的に数式や化学式が含まれる割合が全資料の数%から 20%程度に見られ、1970 年代の画質に、傾き、かすれ、つぶれ等の画質データの悪状況が 50%みられる。

A-a	公表特許

サンプリング調査対象の資料では 1980 年代は「多面付け」が 100%、1990 年代は「多段組み」が 100%で、書式変更が見られた。下線が引かれている率が 40%以上あり、全調査資料の中で最もこの率が高い。また、数式や化学式が含まれる割合が全般的に 10%程度に見られ、比較的高い傾向にある。本文と図版が混在している書類が 57%と多く見られる。

A-b	再公表特許

「公表特許」とよく似た傾向が見られる。サンプリング調査対象の資料では 1980 年代は「多面付け」が 100%、1990 年代は「多段組み」が 42%で、書式変更がみられた。下線が引かれている率が 20%前後あり、比較的よく見られる。また、数式や化学式が出現する

割合が 20%前後あり、比較的高い。1980 年代の画質で、かすれ、つぶれ等の画質データの悪状況が 90%以上に見られる。

B	公告特許

1930 年代の資料は、この「B 公告特許」と「D 公告実用新案」2 種のみであったが、いずれも縦書きのレイアウトであった。1930 年代には 100%、1950 年代になっても 85%が旧字旧かなによる資料であった。1930 年代から 1980 年代まで、100%「多段組み」の資料であり、多段組みのフォーマットである。1930 年代には、ほぼすべてブロック囲みの資料であったが、1950 年以降は 0%であり、ブロック囲みの書式はなくなった。1950 年以降、それまで出現しなかった化学式の出現が出始め、1950 年以降 15%程度に見られるようになった。1930 年代の資料にはノイズが 26%と多く発生している。1950 年代の画質に、つぶれ等の画質データの悪状況がみられる。

C	公開実用新案

1970 年代から 1990 年代まで、100%「多段組」の資料であり、多段組みのフォーマットである。手書き書類があるものが 70%を占め、数式や化学式は出現しない。多くの書類、90%前後で、傾きがみられる。

H	公表・再公表実用新案

1980 年代の書式としては多段組み 60%、下線が 40%程度出現していた。1970 年代から 1980 年半ばまでは、文字のつぶれ等の画質データの悪状況が 87%から 100%と多くみられる。

D	公告実用新案

広告特許以外に 1930 年代が存在するもう 1 種類の書類であり、100%が縦書きのレイアウト、旧字旧かな、多段組みであった。手書き書類があるものが 60-80%を占め、数式や化学式は出現しない。1950 年代の書類に傾きが多くみられる。1970 年以降に本文と図版が混在している書類が 20%程度みられる。

E	公開実用新案全文

印紙や押印のある書類が 100%近くを占めている。また 1970 年代から 1980 年代にかけて手書きが 72%ほど見られたが、1990 年代には手書きはなくなった。本文と図版が混在している書類が 20%程度みられる。

各年代、種別を通して共通するOCR技術によるテキストデータ作成時の留意点として、下記があげられる。

① 縦書きと横書きが、1947年から1948年の間に一斉に切り替わっている。
② 旧字、旧仮名遣いは1952年以前の文章全てにおいて確認されたが、1952年5月の文書より新字、新仮名遣いが混在し始め、8月の文書では新字、新仮名遣いのみの資料が確認された。その後も旧字、旧仮名遣い混じりの文書は確認されたが、1953年に入るとほぼ見られなくなった。
③ 公告特許、公開実用新案、公告実用新案全文は、1930年代を除き、全て多段組みレイアウトである。
④ 公開特許、公表特許、再公表特許、公表・再公表実用新案はほぼ全てが多面付レイアウトである。
⑤ 印紙は、1970年代までの公開特許、1980年代までの公開実用新案に限って極めて高い90%以上の出現率である。
⑥ 押印は、1970年代までの公開特許、1990年代までの公開実用新案で殆どに付されており、その後急激に出現率が減る。
⑦ 数式・化学式は、特許書面において相当程度（10%から30%程度）含まれる。
⑧ 本文中囲み罫や下線は、文書の種類及び年代を問わず出現頻度は低い。
⑨ ブロック囲み罫は、全体としては出現率が低いが、再公表特許においてのみ、極めて高い出現率となっている。
⑩ 手書き書き込みは、実用新案に集中して出現している。
⑪ 画質では、傾きが最も全体的な出現頻度が高く、ノイズは少ない。かすれ、つぶれは、特定の年代に集中して出現している（4.で詳述）。

今後のテキスト化の実施に工夫を要することが必要であるいくつかの点について、その原因および今後の方向性について検討を加える。

① 縦書き，横書き逆順文字，横書き逆順文字
1947年から1948年の間に一斉に、公報のフォーマット変更に伴って横書きから縦書きに切り替わり、横書き逆順文字についても1948年以降は見られなくなった。これらの改定に合わせてOCRソフトの設定等を明確に切り替えることが求められる。
② 旧字，旧仮名遣い
1952年より旧仮名遣いが混在しはじめ、1953年に入るとほぼ見られなくなったことから、移行期間を経て完全に切り替わったと想定される。このように変更した時期が明確なものについては，旧字の認識が得意な特定OCRソフトを選定する、または旧字の教師データを学習させたAI-OCRの利用等が望ましいと考えられる。
③ 多段組みレイアウトへの対応

公告特許，公開実用新案，公告実用新案全文において、1930 年代を除き全て多段組みであり、OCR ソフト側でのレイアウト認識の対応が求められる。

④ 多面付けレイアウトへの対応

公開特許，公表特許，再公表特許，公表・再公表実用新案はほぼ全てが多面付レイアウトであるため、レイアウト認識の対応が必要である。また，多数見られる 4 面付レイアウトで，面ごとに傾きがバラバラである場合には，既存の OCR ソフトが実装する傾き補正では対応できない。特に，4 面付レイアウトは，公開実用新案全文（E）以外のすべての文書種別において，多数出現しており，ページごとに傾きが異なっているものが相当程度存在していた。このため，OCR の前処理として，四分割処理および分割画面ごとの角度補正の仕組みの実装が有効であろう。

⑤ 印紙への対応

印紙は、1970 年代までの公開特許，1980 年代までの公開実用新案に限って極めて高い出現率であるため、古い年代資料で特に、印紙をレイアウト認識し、テキスト化対象外とする対応が求められる。

⑥ 押印への対応

押印は公開特許、公開実用新案で殆どに押印されており、押印部をレイアウト認識しテキスト化対象外とする対応が求められる。

⑦ 数式・化学式への対応

数式・化学式は相当程度（10%から 30%程度）含まれるため、テキスト化にあたってのルール設定と認識後の置換処理が必要である。

⑧ 本文中囲み罫や下線は、出現頻度が低いため優先順位は低い。

⑨ ブロック囲み罫

再公表特許においてのみ，ブロック囲み罫は極めて高い出現率のため，この書類のテキスト化の際の留意事項である。

⑩ 手書き書き込み

実用新案に集中して出現しており、この書類のテキスト化の際に対応が求められる。

⑪ 画質の乱れ

画質は、OCR によるテキスト化精度への影響が大きい。下記 a)に文字のかすれ・つぶれ、b)に画像の傾きに関してまとめる。

a）文字のかすれ・つぶれ

文字のかすれは，以下の種別，年代において多数確認された。

公開特許（A）	1970 年代
再公表特許（A-b）	1980 年代，1990 年代
公表特許（B）	1950 年代
公告実用新案全文（D）	1950 年代

公開実用新案全文（E）　　1970 年代，1980 年代

かすれの出現頻度は，年度との強い相関がみられ，さらなる詳細調査を行えば，どの種別の何年のファイルがかすれているかという点について，より具体的に特定していくことが可能である。こうした「かすれ」のあるイメージデータは，既存の OCR 技術では，実用精度でテキスト化することは不可能であるため，AI を用いた画像補正・文脈補正等の，新しい対応が求められる。

また，かすれが発生している種別・年代で，同時につぶれが発生しているパターンが多く見られた。この二つの事象は，必ずしも同時発生しているものではないが，併発しているものも複数存在した。

同様に，つぶれの場合も，カスレと同様に，既存の OCR 技術では，実用精度でテキスト化することは不可能であるため，AI を用いた画像補正・文脈補正等の対応が求められる。

b）画像の傾き

画像の傾きは，すべての種別・年度で確認された。

現状の OCR ソフトにおいても，一定の傾き補正機能は実装されているため，傾きの存在が，直ちに大幅な認識精度の低下を招くものではないが，特に，多数見られる 4 面付レイアウトで，面ごとに傾きがバラバラである場合には，既存の OCR ソフトが実装する傾き補正では対応できない。特に，4 面付レイアウトは，公開実用新案全文（E）以外のすべての文書種別において，多数出現しており，ページごとに傾きが異なっているものが相当程度存在していた。

このため，OCR の前処理として，四分割処理および分割画面ごとの角度補正の仕組みを実装する等の対策が求められるであろう。

6. 小括

平成 4 年度以前の公開特許公報等について、貸与データから抽出した 8,098 件のサンプルデータの調査により、種別ごとの発行日やフォーマットの遍歴、各資料種別の図版や下線やルビ、数式や化学式の混在状況が明らかとなった。

フォーマットの変遷による書誌情報やレイアウトの変更と合わせ、種別や年代ごとに OCR によるテキスト化阻害要因となり得る、多面付けや多段組み、文字のかすれやつぶれ、傾き等の要素も認められている。それらの特徴が、OCR の認識精度に大きな影響を与える場合は、それらの傾向の大小がデータのテキスト化に向けた優先順位の検討の重要な判断材料になり得る。

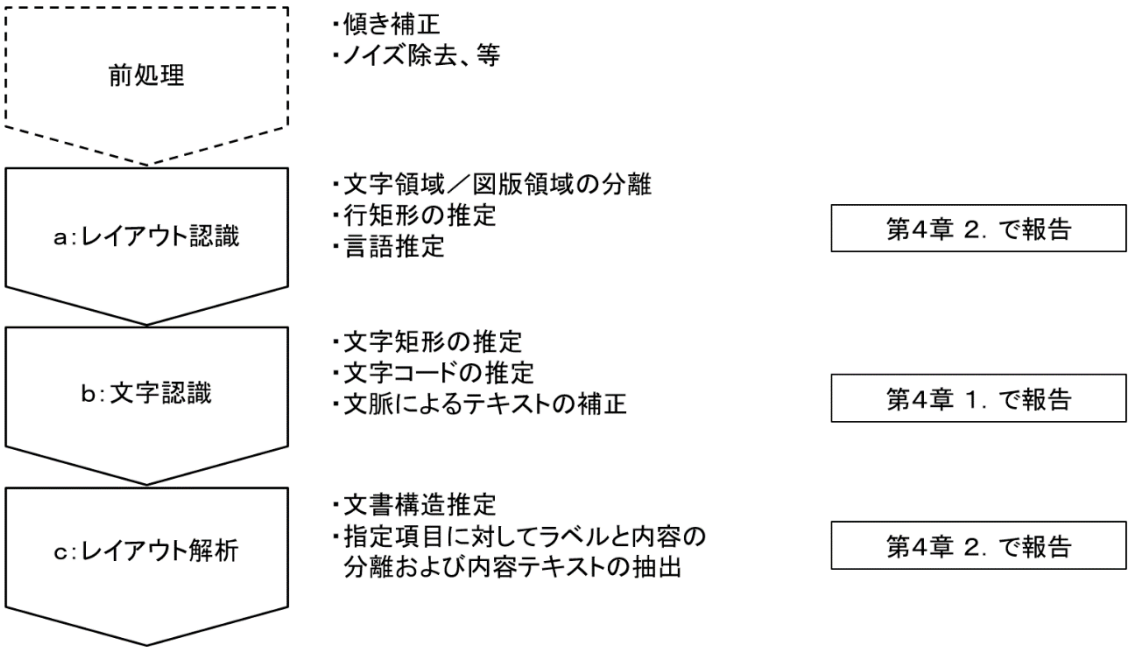
フォーマット調査の結果、現状技術では OCR による実用精度の確保が難しいと考えられるデータが多く見受けられ、テキスト化精度向上のための様々な追加施策が求められる結果となった。

第 4 章　既存 OCR ソフトの調査

日本語の公開特許等のイメージデータをテキストデータに変換するのに最も適していると認められる、複数の既存の無料または有料の OCR ソフトについて、文字認識精度、処理能力、コスト等についてツールごとに評価を行った。

一般に非定型文書対応の OCR ソフトは以下の 3 ステップで文字認識を処理している。

図 4-1　一般的な OCR ソフトの処理フロー



第 4 章では、まずレイアウト情報（文字領域および行矩形）が正しく与えられた場合の〈b：文字認識〉の精度に関して、「1. 文字認識精度の調査」にて調査結果を述べる。次に「2. レイアウト解析等の効率性・正確性」において、〈a：レイアウト認識〉と〈c：レイアウト解析〉の処理フローの中の位置づけおよび調査結果を述べる。

1. 文字認識精度の調査

(1) 調査概要

文字認識の対象となる文字の行矩形の座標情報（X 座標、Y 座標、幅、高さ）が正しく与えられた場合の文字認識精度（行内文字認識の精度）の調査を実施した。行矩形認識と行内文字認識を同時に実行した場合の、文書全体に対する認識精度に関しては、「2. レイアウト解析等の効率性・正確性」で述べる。

図 4-2　一般的な OCR ソフトの処理フロー

備考）

文書全体の文字認識精度と、レイアウト認識精度および行内文字認識の精度は、

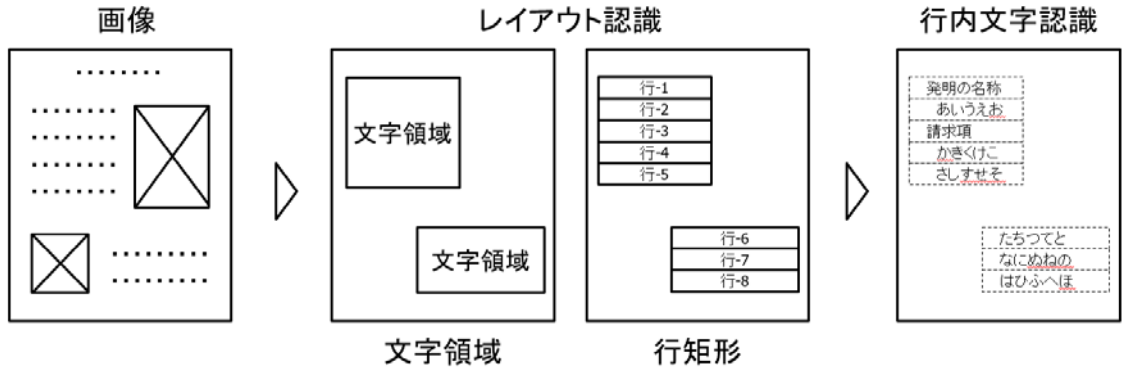
文書全体の文字認識精度　＝

レイアウト認識の精度（行矩形認識まで）　×　行内文字認識の精度

※レイアウト認識精度：90%、行内文字認識精度：99%の場合、全体の文字認識精度：89.1%

という式であらわされる。

このうち、レイアウト認識の精度は、公報における文字認識の対象（書誌情報、本文、図版中の文字、表組内の文字、等）の位置によって大きく変わり、また各種公報のレイアウト毎に当該各文字認識の対象の位置を特定する処理システムの開発によって大きく精度が向上するため、各既存 OCR ソフトの文字認識精度自体の評価には、「行内文字認識」を基礎的な指標として調査することが妥当と考えられる。



文字認識精度の調査対象は、第 2 章「4. 検証用データの抽出」で抽出した 222 件の検証用データを使用した。

使用した OCR ソフト（SDK 版）は以下の三種類である。

A：Google、ネバダ大学　「Tesseract　OCR」[1]

B：PANASONIC ソリューションテクノロジー　「活字認識ライブラリー」[2]

C：NTT データ NJK　「活字文書 OCR ライブラリ」[3]

※ A は無償ソフト、B および C は有償ソフトであり、それぞれ OCR ソフトとして主なものから選択した。

各 OCR ソフトによる文字認識精度を計測するため、予め各行の正解テキストを目視によって確定した。（総文字数：342,504 文字）また、文字認識精度は、行単位で正解テキストと OCR ソフトによる文字認識で得られたテキストとの間のレーベンシュタイン距離※を求め、以下の計算式で算出した。ここで、以下の式の行内文字認識率は文字が順序も含めてすべて正解であったときに 100%となり、レーベンシュタイン距離が行内文字数以上の場合には 0%となる。

式 4-1　文書全体の行内文字認識率の総和の計算式

$$\text{文書全体の行内文字認識率の総和} = \frac{\Sigma(\text{各行正解文字数} - \text{Min}(\text{各行正解文字数}, \text{レーベンシュタイン距離}))}{\Sigma(\text{各行正解文字数})}$$

※ レーベンシュタイン距離）1 文字の挿入・削除・置換により、正解文字列に修正するために必要な手順の最小回数。

また、文字の正誤判断基準は下記のとおりとした。

- 全角/半角スペースは正誤判断の対象外とし、正解テキストの文字数にも含めない。（例：「昭和五年」と「昭和　五年」は文字数 4 文字の同一の文字列として扱う）
- 英数字及び記号について、全角と半角を区別しない。（例：「A」と「A」、「：」と「:」は同じ文字として扱う）
- 英字の大文字と小文字は区別する。（例：「A」と「a」は区別する）
- 小書き文字と並字は区別する。（例：「あ」と「ぁ」は別の文字として扱う）
- 清音と濁音、半濁音は区別する。（例：「は」と「ば」、「ぱ」は全て別の文字として扱う）
- 図版及び図版内の文字は正誤判断の対象外とし、正解テキスト文字数とレーベンシュタイン距離を共に「0」とする。

[1] https://github.com/tesseract-ocr/tesseract

[2] https://www.panasonic.com/jp/business/its/ocr_sdk/textocr.html

[3] https://mediadrive.jp/products/library/katsuji_library/index.html

・ 正解テキスト内の「〓」（※目視でも判別不能な文字）は正誤判断の対象外とし、正解テキストの文字数にも含めない。

調査結果データは、画像名、行番号、正解テスト、正解文字数、既存 OCR ソフト A、B、C のそれぞれの文字認識結果、レーベンシュタイン距離で構成されている。その一部を下記に例示する。OCR ソフトの出力例については別紙 4 を参照されたい。

表 4-1　調査結果データ（一部）

画像名	行番号	正解 テキスト	正解文字数	認識 結果_A	レーベンシュタイン距離_A	認識 結果_B	レーベンシュタイン距離_B	認識 結果_C	レーベンシュタイン距離_C
H_83500005_01	1	'⑲日本国特許庁(jp)	11	'國日本国特許庁up)	3	'@日本国特許庁(jp)	1	'⑩日本国特許庁(jp)	1
H_83500005_01	2	'⑪実用新案出願公表	9	'@実用新案出願公表	1	'実用新案出願公表	1	'⑩実用新案出願公表	1
H_83500005_01	3	'⑫公表実用新案公報(u)	12	'@公表実用新案公報(u)	1	'公表実用新案公報(u)	1	'⑩公表実用新案公報町	4
H_83500005_01	4	'昭 58—500005	10	'日召 58 一 500005	3	'日召 58－500005	3	'昭 58-500005	1
H_83500005_01	5	'〓int.cl.3 識別記号庁内整理番号	18	'@耐〔cl3 識別記号庁内整理番号	5	'$ int.　cl3 識別記号庁内整理番号	2	'命 lnt.ci.3 識別記号庁内整理番号	2
H_83500005_01	6	'h02k21/087733—5h	16	'h02k2 -/087733一 5h	2	'ho2k21／087733－5h	3	'h02k21/087733-5h	1
H_83500005_01	7	'1/287509—5h	11	'-/287509~5h	2	'1／287509－5h	2	'1/287509-5h	1

(2) 調査結果

第 3 章で区分した 32 区分毎の文字認識精度として、各既存 OCR ソフトによる文字認識率を示す。「画質悪い」の項目に✔印がある区分は、文字のつぶれ・かすれが、当該区分に属するサンプルデータの 50%以上に目視で発見された区分である。

表 4-2　区分毎の各既存 OCR ソフトによる文字認識率

区分 No	実験用データ数	検証用データ数	縦組	旧字旧仮名	画質が悪い	OCR ソフト		
						A	B	C
A-①	211	13	—	—	✔	72.4%	84.7%	93.2%
A-②		13	—	—		67.3%	76.5%	88.1%
A-③		1	—	—		79.2%	94.1%	90.6%
A-④		1	—	—		85.2%	94.5%	91.8%
A-a-①	119	13	—	—		80.9%	91.4%	95.6%
A-a-②		1	—	—		76.9%	93.8%	90.4%
A-a-③		1	—	—		85.1%	96.7%	87.1%
A-b-①	100	13	—	—	✔	64.1%	86.4%	86.3%
B-①	503	14	○	○		40.2%	67.6%	70.3%
B-②		3	○	○		50.9%	54.6%	83.7%
B-③		1	—	○	✔	48.6%	71.7%	40.8%
B-④		13	—	○	✔	27.3%	64.0%	22.2%
B-⑤		26	—	—		86.1%	93.1%	97.0%
B-⑥		1	—	—		87.8%	94.3%	94.8%
C-①	200	14	—	—		80.4%	91.0%	94.7%
C-②		13	—	—		85.3%	92.6%	95.1%
C-③		1	—	—		82.4%	93.6%	93.9%
C-④		1	—	—		87.5%	95.5%	92.2%
H-a-①	166	10	—	—	✔	77.8%	91.6%	96.3%
H-a-②		4	—	—	✔	76.0%	90.8%	95.5%
H-a-③		1	—	—		75.1%	84.6%	98.0%
H-b-①	19	1	—	—		52.3%	86.3%	94.1%
D-①	500	1	○	○		47.7%	59.5%	83.8%
D-②		3	○	○		49.6%	45.1%	79.3%

区分 No	実験用 データ数	検証用 データ数	縦組	旧字 旧仮名	画質が 悪い	OCR ソフト		
						A	B	C
D-③		1	—	○	✔	19.6%	37.5%	49.1%
D-④		14	—	○	✔	58.9%	86.4%	73.5%
D-⑤		27	—	—		88.7%	95.9%	97.6%
D-⑥		1	—	—		88.1%	94.8%	94.2%
E-①	208	13	—	—	✔	59.5%	77.4%	87.4%
E-②		1	—	—	✔	66.8%	88.8%	90.1%
E-③		1	—	—		76.8%	95.3%	86.9%
E-④		1	—	—	✔	76.5%	92.6%	89.9%
					合計	81.6%	90.5%	91.4%

89.7% 最も精度がよいOCRツール

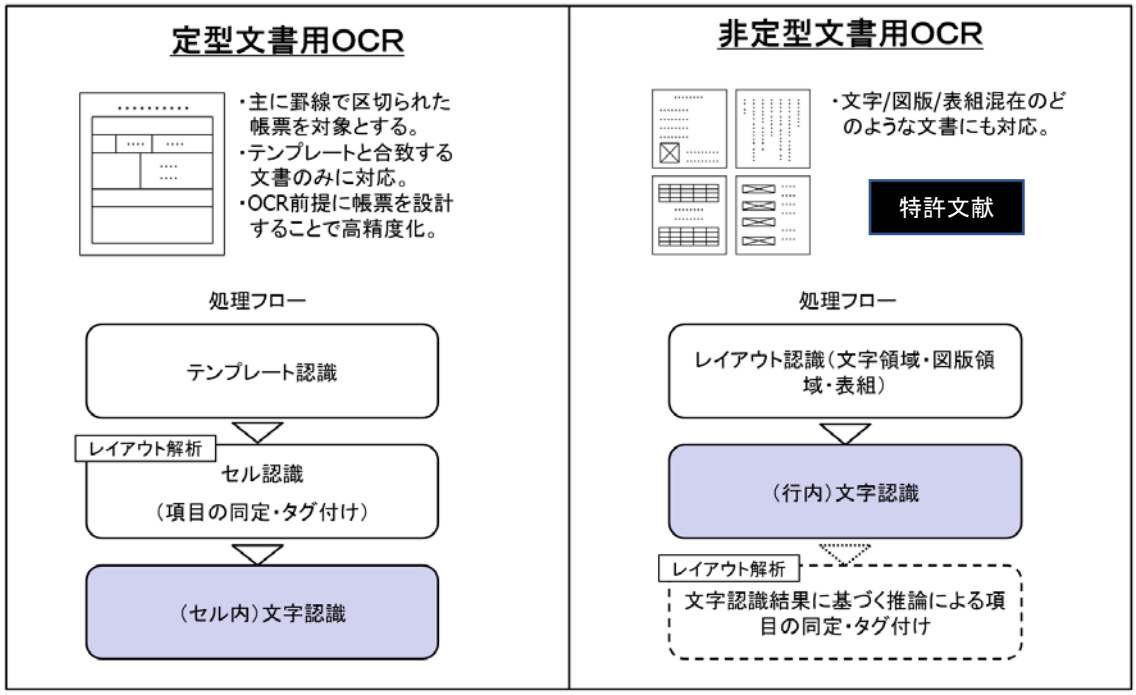
全体的な傾向として、OCR ソフト C の文字認識精度が高かったが、画質が悪い B-④、D-④の区分に関しては OCR ソフト B の文字認識精度が高かった。また B-③や D-③の区分（縦書き・旧字旧仮名遣い）は、全ての OCR ソフトにおいて文字認識精度が低く、最大で 70%程度に留まる。無償ソフトである OCR ソフト A は、他の B、C と比較して、全般的に文字認識精度が低かった。

2. レイアウト解析等の効率性・正確性

(1) 本事業におけるレイアウト解析の位置付け

一般に OCR ソフトは、帳票などの固定レイアウト文書を処理することに適した「定型文書用 OCR ソフト（帳票 OCR とも呼称される）」と自由レイアウト文書を処理することに適した「非定型文書用 OCR ソフト」に大別される。それぞれの特徴および一般的な処理フローを下図に示す。

図 4-3　定型文書と非定型文書の違い



文書構造を解析し特定の項目情報を抽出する際、定型文書用 OCR ソフトでは、テンプレートに従い罫線で囲まれたセルを同定することで、同時に項目情報まで同定する。一方、非定型文書用 OCR ソフトでは、文字認識処理を実行した後に読み取ったテキスト情報等から項目を推定する必要があるが、汎用的な項目同定機能を有する非定型文書用 OCR ソフトは存在せず、必要に応じて個別に開発する必要がある。

特許及び実用新案の各種公報は、構成するページの大部分は非定型文書に属し、書誌情報が記載されている箇所のみ、一部定型文書的な性質を有している。しかし書誌情報の記載箇所についても、罫線で明確に区切られていない、項目内容の長さが可変である等の理由により、既存の汎用的な定型文書用 OCR ソフトでは扱うことはできず、非定型文書用 OCR ソフトを使用し、項目同定機能を追加開発することで初めて各項目へのタグ付けを伴うテキスト化が可能になる。したがって、今回の調査事業では、各種公報を非定型文書として扱い、第 4 章冒頭で示した処理フローを前提とした調査を実施した。

本節では、前節で報告した文字認識前後のプロセスであるレイアウト認識（文字領域・図版領域の識別）とレイアウト解析（項目の同定）についての調査結果を報告する。

(2) 調査概要

a) レイアウト認識

各既存の OCR ソフトのレイアウト認識（文字領域と図版領域の識別）の精度を調査した。調査対象データとしては、第 2 章「4. 検証用データの抽出」で抽出した 222 件の検証用デ

ータを使用した。

使用した OCR ソフト（パッケージソフト版）は以下の 3 種類である。

B：パナソニックソリューションテクノロジー「読取革命 Ver.15」[4]　※1
C：NTT データ NJK　「e.Typist v15.0」[5]　※2
D：ABBYY　「FineReader15 Standard」[6]　※3
※1　文字認識精度検証に使用した OCR ソフト B と同一エンジン
※2　文字認識精度検証に使用した OCR ソフト C と同一エンジン
※3　文字認識精度検証に使用した OCR ソフト A は、機能に問題がありレイアウト認識率が測定不能だったため、市販の別の OCR ソフト D を使用した。

レイアウト認識の精度の調査は、222 件の検証用データについて、OCR ソフトの自動処理モードにより認識されたレイアウトと、目視で作成した正解のレイアウトとを比較し、レイアウト認識率を求めた。調査手法に関する概念図を以下に示す。OCR ソフトの出力例は別紙 4 を参照されたい。

[4] https://www.panasonic.com/jp/company/pstc/products/yomikaku.html
[5] https://mediadrive.jp/products/et/
[6] https://www.abbyy.com/ja-jp/finereader/

図 4-4　目視で作成した正解のレイアウトの例

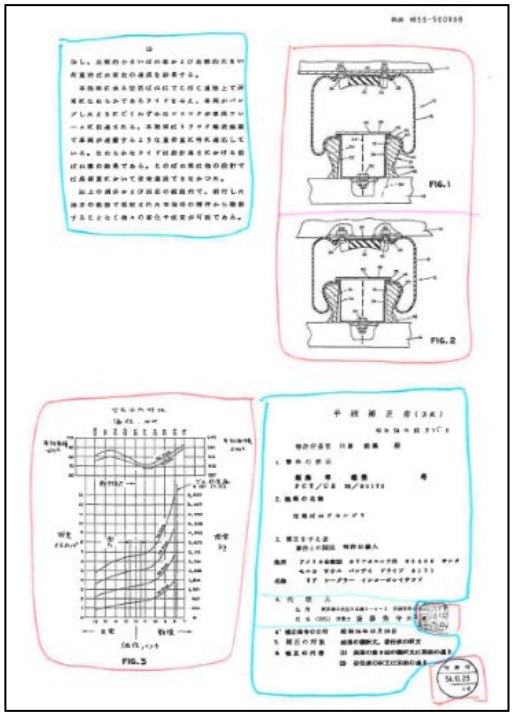


図 4-5　OCR ソフトの自動処理モードにより認識されたレイアウトの例

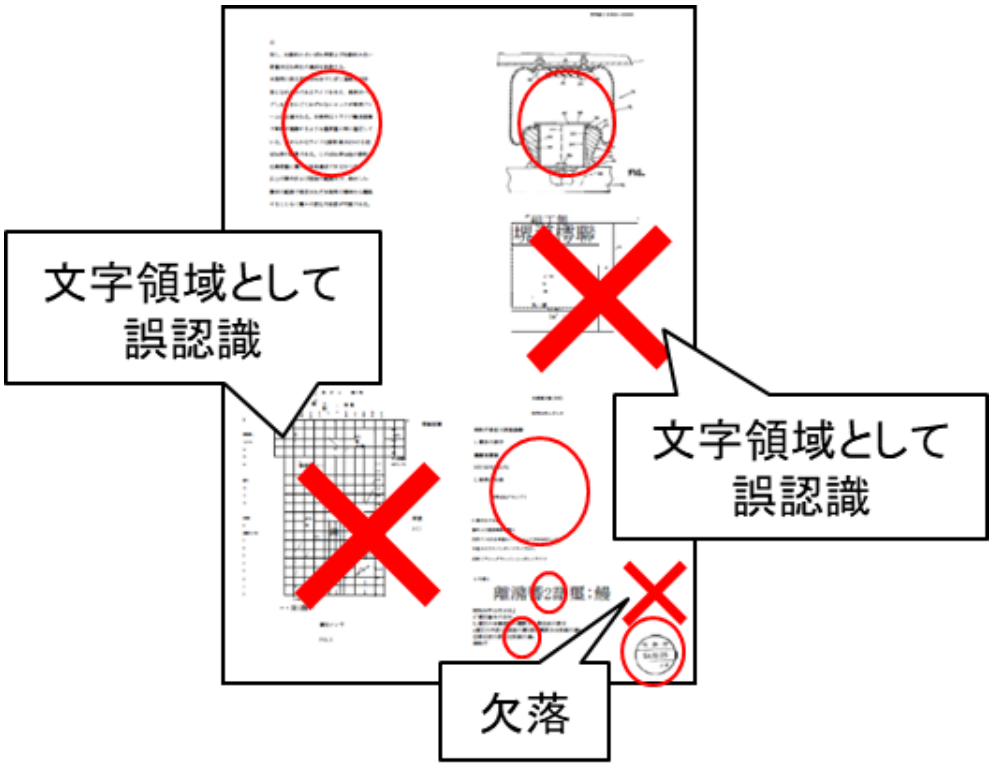
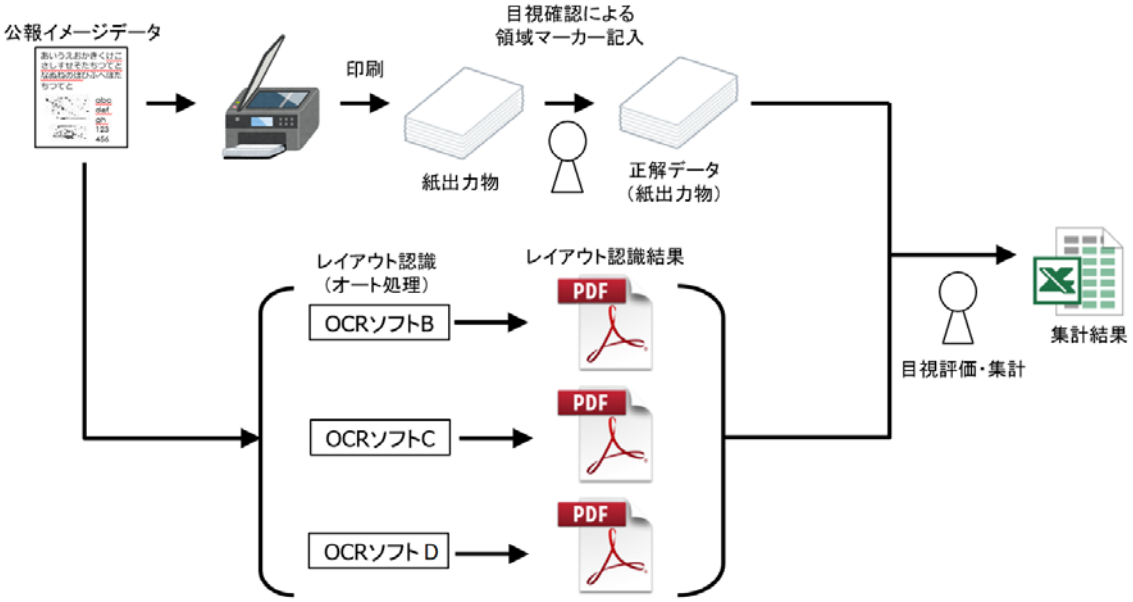


図 4-6　レイアウト認識率調査のワークフロー



b）レイアウト解析

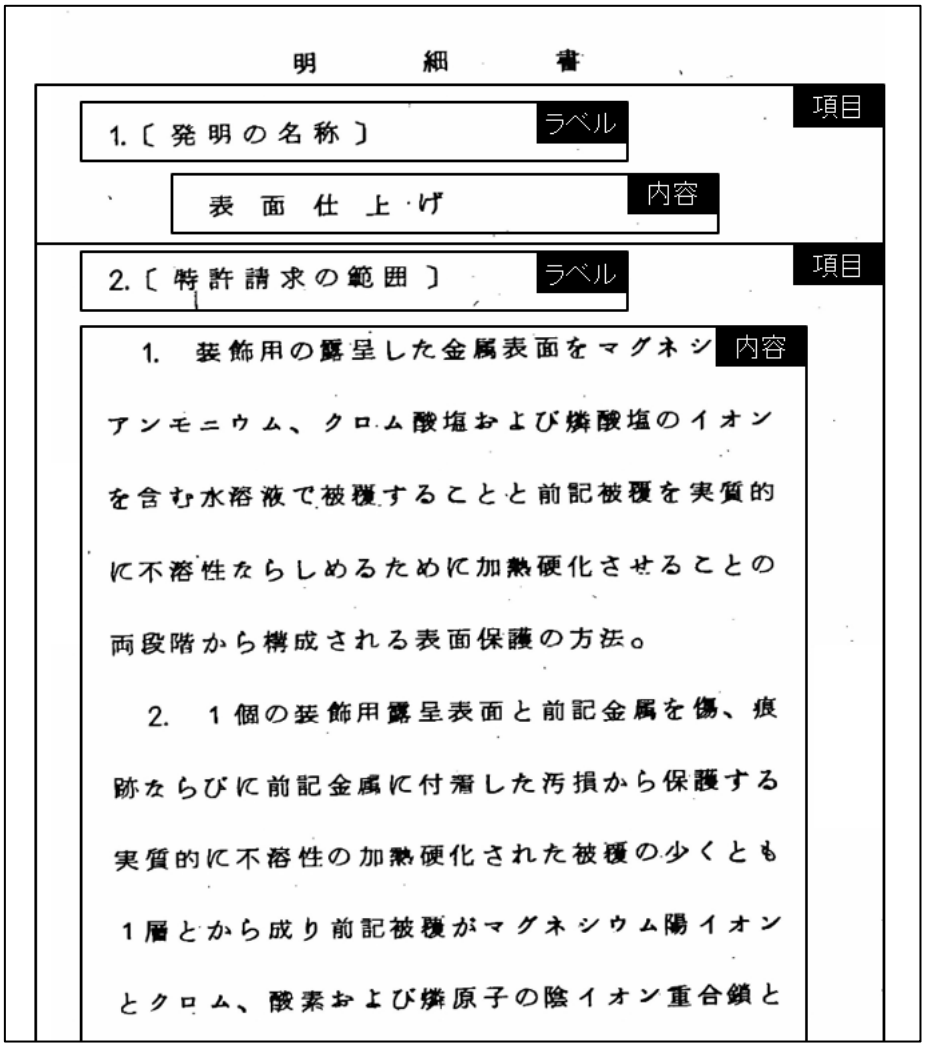
下図のように各項目に関して「ラベル」と「内容」を分離抽出するソフトの開発を想定し、基礎的な調査を実施した。各項目の「内容」に関する文字認識精度は、

「内容」の文字認識精度　＝

①「ラベル」の検出精度 × ②「内容」行の推定精度 × ③行内文字認識率

により算出できる。本節では、基礎情報となる①ラベルの検出精度について調査・報告した。(③の行内文字認識精度は第 4 章 1. 文字認識精度で調査している。②の「内容」行の推定精度は調査していないが、項目ごとの難易度・推定プログラム開発の手法によって大きく異なることが想定される。)

図 4-7　項目・ラベル・内容の関係



レイアウト解析の検証項目は、「発明の名称」「出願日」「出願番号」「要約」「特許請求の範囲」「発明の詳細な説明」の 6 項目とした。

これらの書誌項目名には、下記に示すとおり、同義語が多数存在するため、本調査にあたっては、予め定義した同義語のいずれかに完全一致する文字列が OCR ソフトによる文字認識によって検出されれば、その項目のラベルの同定に成功したとみなした。また、ラベルの文字列が複数検出された場合については、なんらかの処置により（例えば、ページの上部に出現した方を優先する、等）絞り込み可能と想定し、1 箇所のみ検出された場合と同等に取り扱った。

（書誌項目の同義語）

・発明の名称：考案の名称、名称、発明の名称、(54)※丸付数字

・出願日　　：出願、国際出願日、㉒、出願日、出願昭、出願平、出願大、出願△大

・出願番号　：特願、実願、實願、㉑実願、㉑特願、願書番號、特許願番號
・要約　　：考案の要約、要約、出願ノ要旨、發明ノ性質及ヒ目的ノ要領、發明ノ性質及目的ノ要領、發明の性質及目的の要領、発明の性質及目的の要領
・特許請求の範囲：特許請求ノ範圍、特許請求の範圍、請求の範囲、実用新案登録請求の範囲、登録請求の範囲、登錄請求の範圍、請求ノ範圍、請求の範圍、請求の範囲、請求の範圍
・発明の詳細な説明：發明の詳細なる說明、發明の詳細なる説明、発明の詳細なる說明、発明の詳細なる説明、發明の詳細な說明、發明の詳細な説明、発明の詳細な說明、発明の詳細な説明、發明の詳細なる説明、発明の開示、實用新案ノ性質、作用及ヒ效果ノ要領、實用新案ノ性質、作用及效果ノ要領、實用新案の性質、作用及效果の要領、實用新案の性質、作用及効果の要領、実用新案の性質、作用及效果の要領、実用新案の性質、作用及効果の要領、実用新案の説明、考案の開示、考案の詳細な説明

本レイアウト解析の調査では、区分毎のレイアウト解析結果の傾向を比較するため、すべての区分において、前節で全般的な文字認識精度が最も高かった OCR ソフト C を用いた。

なお、本事業で行うレイアウト解析では、公報中にその項目を認識できる上記の書誌の表記が存在していない場合には、レイアウト解析結果は 0%となる。どの区分でどのような項目の表記となっているのかについては、別紙 3 を参照のこと。

(3) 調査結果

a) レイアウト認識

まず、全体的な傾向を把握するため、各既存 OCR ソフトによりレイアウト認識を行った結果を、文字領域と図版領域それぞれについて以下に示す。

表 4-3　レイアウト認識結果

対象とする領域		OCR ソフト		
		B	C	D
縦組	文字領域	40.8%	80.1%	71.5%
	図版領域	72.3%	46.4%	89.8%
横組	文字領域	58.5%	89.2%	63.1%
	図版領域	86.4%	68.2%	95.5%
合計	文字領域	41.2%	80.3%	71.3%
	図版領域	72.5%	46.8%	89.9%

文字領域の認識は、OCR ソフト C が優れ、図版領域の認識は、OCR ソフト B と OCR ソフト D が優れているという結果が得られた。

次にテキスト化にあたって重要となる文字領域のレイアウト認識の精度について、32 区分毎に集計した、レイアウト認識率の結果を以下に示す。参考に、当該得られたレイアウト認識率を、前節の文字認識率と乗算して算出される、文書全体を既存 OCR ソフトにより自動処理した場合の文字認識率（レイアウト認識込み）の予測を併記した。

文書全体を自動処理した場合の予測精度

表 4-4　区分毎の文字領域のレイアウト認識率

区分 No	レイアウト認識率①			文字認識率②		参考 ： ①×②	
	B	C	D	B	C	B	C
A-①	51.4%	85.0%	70.4%	84.7%	93.2%	43.5%	79.2%
A-②	51.4%	85.0%	70.4%	76.5%	88.1%	39.3%	74.9%
A-③	26.6%	76.6%	72.8%	94.1%	90.6%	25.0%	69.4%
A-④	26.6%	76.6%	72.8%	94.5%	91.8%	25.1%	70.3%
A-a-①	26.6%	76.6%	72.8%	91.4%	95.6%	24.3%	73.2%
A-a-②	26.6%	76.6%	72.8%	93.8%	90.4%	24.9%	69.2%
A-a-③	26.6%	76.6%	72.8%	96.7%	87.1%	25.7%	66.6%
A-b-①	51.4%	85.0%	70.4%	86.4%	86.3%	44.4%	73.4%

区分 No	レイアウト認識率①			文字認識率②		参考 ： ①×②	
	B	C	D	B	C	B	C
B-①	58.5%	89.2%	63.1%	67.6%	70.3%	39.5%	62.7%
B-②	58.5%	89.2%	63.1%	54.6%	83.7%	31.9%	74.7%
B-③	52.3%	79.2%	65.1%	71.7%	40.8%	37.5%	32.3%
B-④	51.4%	85.0%	70.4%	64.0%	22.2%	32.9%	18.9%
B-⑤	26.6%	76.6%	72.8%	93.1%	97.0%	24.7%	74.3%
B-⑥	26.6%	76.6%	72.8%	94.3%	94.8%	25.0%	72.5%
C-①	26.6%	76.6%	72.8%	91.0%	94.7%	24.2%	72.5%
C-②	26.6%	76.6%	72.8%	92.6%	95.1%	24.6%	72.8%
C-③	26.6%	76.6%	72.8%	93.6%	93.9%	24.9%	71.9%
C-④	26.6%	76.6%	72.8%	95.5%	92.2%	25.4%	70.6%
H-a-①	51.4%	85.0%	70.4%	91.6%	96.3%	47.1%	81.9%
H-a-②	51.4%	85.0%	70.4%	90.8%	95.5%	46.7%	81.2%
H-a-③	26.6%	76.6%	72.8%	84.6%	98.0%	22.5%	75.0%
H-b-①	26.6%	76.6%	72.8%	86.3%	94.1%	22.9%	72.0%
D-①	58.5%	89.2%	63.1%	59.5%	83.8%	34.8%	74.7%
D-②	58.5%	89.2%	63.1%	45.1%	79.3%	26.4%	70.8%
D-③	52.3%	79.2%	65.1%	37.5%	49.1%	19.6%	38.9%
D-④	51.4%	85.0%	70.4%	86.4%	73.5%	44.4%	62.4%
D-⑤	26.6%	76.6%	72.8%	95.9%	97.6%	25.5%	74.7%
D-⑥	26.6%	76.6%	72.8%	94.8%	94.2%	25.2%	72.1%
E-①	51.4%	85.0%	70.4%	77.4%	87.4%	39.8%	74.3%
E-②	51.4%	85.0%	70.4%	88.8%	90.1%	45.7%	76.6%
E-③	26.6%	76.6%	72.8%	95.3%	86.9%	25.3%	66.5%
E-④	51.4%	85.0%	70.4%	92.6%	89.9%	47.6%	76.4%
平均	41.2%	80.3%	71.3%				

文字領域のレイアウト認識の精度は、全区分において OCR ソフト C の精度がよいことが判明した。

b) レイアウト解析

「発明の名称」「出願日」「出願番号」「要約」「特許請求の範囲」「発明の詳細な説明」の 6 項目について、ラベルの文字列の検出精度を 32 の区分毎に示す。

表 4-5　ラベル文字列検出精度

区分 No	項目のラベル文字列						平均
	発明の名称等	出願日等	出願番号等	要約等	特許請求の範囲等	発明の詳細な説明、等	
A-①	100%	31%	92%	92%	69%	54%	73%
A-②	92%	0%	46%	46%	69%	62%	53%
A-③	100%	0%	100%	100%	100%	50%	75%
A-④	100%	100%	100%	100%	100%	100%	100%
A-a-①	0%	0%	15%	15%	100%	15%	24%
A-a-②	100%	0%	100%	100%	100%	0%	67%
A-a-③	100%	0%	0%	100%	0%	100%	50%
A-b-①	92%	15%	100%	100%	8%	85%	67%
B-①	0%	0%	0%	0%	50%	0%	8%
B-②	0%	0%	0%	0%	33%	0%	6%
B-③	0%	0%	100%	100%	0%	0%	33%
B-④	0%	0%	38%	38%	0%	8%	14%
B-⑤	0%	0%	35%	38%	100%	88%	44%
B-⑥	100%	0%	100%	100%	100%	100%	83%
C-①	0%	0%	79%	0%	93%	0%	29%
C-②	0%	15%	85%	0%	92%	0%	32%
C-③	100%	0%	50%	0%	100%	0%	42%
C-④	100%	0%	0%	100%	100%	0%	50%
H-a-①	10%	0%	40%	10%	100%	60%	37%
H-a-②	100%	0%	100%	75%	100%	0%	63%

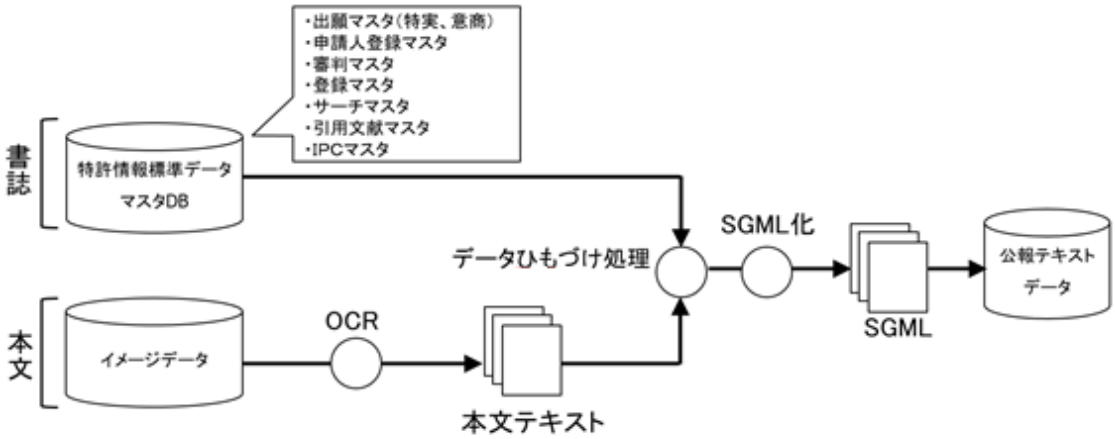
区分 No	項目のラベル文字列						平均
	発明の名称 等	出願日 等	出願番号 等	要約 等	特許請求の範囲 等	発明の詳細な説明、等	
H-a-③	100%	0%	0%	100%	100%	0%	50%
H-b-①	100%	0%	100%	0%	0%	100%	50%
D-①	0%	0%	0%	0%	0%	0%	0%
D-②	0%	0%	0%	0%	0%	0%	0%
D-③	0%	0%	0%	0%	0%	0%	0%
D-④	0%	0%	50%	0%	0%	0%	8%
D-⑤	0%	0%	37%	0%	100%	78%	36%
D-⑥	100%	0%	100%	0%	100%	100%	67%
E-①	85%	0%	15%	8%	77%	46%	38%
E-②	100%	0%	100%	0%	100%	100%	67%
E-③	100%	0%	100%	0%	100%	100%	67%
E-④	100%	0%	0%	100%	100%	100%	67%
平均	56%	5%	53%	41%	65%	42%	44%

結果として、各区分のいずれも年代が新しいものほど検出率がよいこと、各区分の公報の種別及び発行年代に関わらず「出願日」の検出が非常に困難なことが判明した。

したがって、本レイアウト解析調査で実施したように、ラベルの文字列の検出によって書誌情報項目を取得する場合には、公報の種別及び発行年代、並びに、検出対象の書誌情報項目に応じて、使用する OCR ソフトの種類を選定するなどの適切な設計を行うことが必要であると考えられる。

なお、OCR ソフトにより完全に自動でレイアウト解析を実施するためには、専用の処理システムの開発が必要なうえに、抽出されたテキストの品質保証という課題が残されるため、公報のイメージデータ以外の外部データとして書誌情報が存在する公報については、当該外部情報との紐づけ処理を行う検討をしたほうが品質・コストともに優位性があると考えられる。

図 4-8　外部書誌情報との紐付けの概念図



3. 処理能力に関する調査

以下の 3 種の既存 OCR ソフトについて、画像読み込み～レイアウト認識・文字認識～データ保存までの時間を測定した。調査にあたっては、400 枚の画像を処理し、その平均値を算出した。

使用した既存 OCR ソフト（レイアウト解析で使用したものと同じ）

B：PANASONIC ソリューションテクノロジー「読取革命 Ver.15」

C：NTT データ NJK　「e.Typist v15.0」

D：ABBYY　「FineReader15 Standard」

その他の条件）

・検証用データから 400 ページ分の画像を抽出し測定。平均値を算出。

・使用した PC のスペック　CPU : Intel Core i7-9700 3.00Ghz（8 コア）　RAM : 16GB

処理能力に関する調査結果は下記のとおりである。

・OCR ソフト B: 平均 5.7 sec / 画像ファイル

・OCR ソフト C: 平均 3.8 sec / 画像ファイル

・OCR ソフト D: 平均 1.6 sec / 画像ファイル

仮に B・C の 2 種の OCR ソフトを並列で使用し、OCR 前後の処理（画像データの取得、データベース登録のための後処理、等）と合わせて、約 10 sec／画像ファイル程度の処理時間が必要とされるものと想定すると、約 1,500 万文献（約 1 億画像）を処理するためには、50 台のサーバを約 8 カ月連続稼働させるようなシステムが必要と試算される。

実際にこのようなシステムを構築するためには、

・OCR ソフトのライセンス費（初期コスト・固定費）
・バッチ処理システム開発費（初期コスト・固定費）
・外部書誌情報データベースとの連携システムの開発費（初期コスト・固定費）
・OCR により得られるテキスト情報と外部情報を適切に処理して構造化されたテキストデータを作成するシステムの開発費（初期コスト・固定費）
・システム運用費（変動費・ランニングコスト）

等の要素を検討する必要がある。概算コストに関しては、第 7 章総括において述べる。

4. 小括

複数の既存 OCR ソフトについて、文字認識精度、レイアウト認識（文字領域の検出)、レイアウト解析、処理能力について、フォーマット調査の結果による 32 区分毎に評価を行った。

行矩形が正しく与えられた場合の文字認識精度はもっとも精度がよいツールで平均 91.4%であった。

レイアウト認識については、もっとも精度がよいツールで平均 80.3%であった。

レイアウト解析については、6 種類の書誌情報項目の「ラベル」部分が取得可能かどうかに関して検証を行い平均 44%程度であった。

そして、いずれの精度も、区分毎のバラつきが大きいことが判明した。

各区分の最適な手法に関しては、第 5 章及び第 6 章の結果を踏まえて第 7 章（総合分析）で述べる。

第 5 章 テキスト化精度向上策の検討・評価

第 4 章で評価した OCR ソフトから、さらにテキスト化精度を向上させるための手法について公報種別毎に検討を行い、各手法を適用した場合の評価を実施した。

第 5 章では、まず複数の既存 OCR ソフトを組み合わせ、多数決処理を実施した場合の文字認識の精度に関して、「1. 各ツールの組み合わせ等による精度向上」で調査結果を述べる。つぎに「2. その他精度向上策の検討」において、「(1) レイアウト認識と文字認識の組み合わせ検討結果」と「(2) 文脈補正の検討結果」について述べる。

使用した既存 OCR ソフト

A：Google、ネバダ大学　「Tesseract　OCR」

B：パナソニックソリューションテクノロジー「読取革命 Ver.15」

C：NTT データ NJK　「e.Typist v15.0」

D：ABBYY　「FineReader15 Standard」

1. 各ツールの組み合わせ等による精度向上

複数の既存 OCR ソフトで並行して文字認識を行い、多数決処理を実施することによる、文字認識精度の向上の可能性について検証する。

多数決処理の仕組みを以下に示す。

図 5-1　多数決処理による精度向上の原理

正解テキスト	OCR ソフト①		OCR ソフト②		OCR ソフト③		出力結果	
ゐ	ゐ	○	ぬ	×	ゐ	○	ゐ	○
さ	き	×	さ	○	さ	○	さ	○
が	が	○	が	○	か	×	が	○
め	め	○	ぬ	×	め	○	め	○
精度	**75%**		**50%**		**75%**		**100%**	

32 区分毎の、各既存 OCR ソフト及び多数決処理後の文字認識率を以下に示す。

表 5-1　既存 OCR ソフト及び多数決処理後の文字認識率

区分 No	検証用データ数	OCR ソフト			ABC の多数決処理後
		A	B	C	
A-①	13	72.4%	84.7%	93.2%	93.8%
A-②	13	67.3%	76.5%	88.1%	88.6%
A-③	1	79.2%	94.1%	90.6%	94.2%
A-④	1	85.2%	94.5%	91.8%	97.0%
A-a-①	13	80.9%	91.4%	95.6%	92.6%
A-a-②	1	76.9%	93.8%	90.4%	94.8%
A-a-③	1	85.1%	96.7%	87.1%	96.2%
A-b-①	13	64.1%	86.4%	86.3%	87.4%
B-①	14	40.2%	67.6%	70.3%	78.7%
B-②	3	50.9%	54.6%	83.7%	85.7%
B-③	1	48.6%	71.7%	40.8%	72.1%
B-④	13	27.3%	64.0%	22.2%	47.1%
B-⑤	26	86.1%	93.1%	97.0%	96.9%
B-⑥	1	87.8%	94.3%	94.8%	96.7%
C-①	14	80.4%	91.0%	94.7%	95.5%
C-②	13	85.3%	92.6%	95.1%	94.6%
C-③	1	82.4%	93.6%	93.9%	95.6%
C-④	1	87.5%	95.5%	92.2%	96.8%
H-a-①	10	77.8%	91.6%	96.3%	93.6%
H-a-②	4	76.0%	90.8%	95.5%	93.0%
H-a-③	1	75.1%	84.6%	98.0%	88.8%
H-b-①	1	52.3%	86.3%	94.1%	89.0%
D-①	1	47.7%	59.5%	83.8%	83.6%
D-②	3	49.6%	45.1%	79.3%	80.6%
D-③	1	19.6%	37.5%	49.1%	58.7%
D-④	14	58.9%	86.4%	73.5%	81.5%
D-⑤	27	88.7%	95.9%	97.6%	96.9%
D-⑥	1	88.1%	94.8%	94.2%	96.3%
E-①	13	59.5%	77.4%	87.4%	88.0%
E-②	1	66.8%	88.8%	90.1%	92.7%
E-③	1	76.8%	95.3%	86.9%	95.2%

区分 No	検証用データ数	OCR ソフト			ABC の多数決処理後
		A	B	C	
E-④	1	76.5%	92.6%	89.9%	93.0%
合計		81.6%	90.5%	91.4%	88.7%

95.3% 多数決の結果、精度が向上した区分

94.8% もっとも認識精度が高かったOCRツール

32 区分中、19 区分において多数決処理後の文字認識率が向上したが、平均すると、OCR ソフト B 及び C よりも文字認識率が低くなる結果となった。また、区分 B -①と D-③における多数決処理後の文字認識率は、その区分で最も文字認識精度が高い単一の OCR ソフトの文字認識率と比較して 10%弱向上しているが、その他に精度が大きく向上している区分はない。したがって、公報の文字認識において多数決処理を行っても、大幅な文字認識精度の向上は期待できないといえる。

既存 OCR ソフトの多数決処理を行う場合、各ソフトのライセンス料や多数決処理プログラム開発費、多数決処理システムの構築運用費を考慮すると、単独の OCR ソフトを使用する場合に比べてコストが 3 倍以上になることが想定される。

よって、公報の文字認識においては、大幅な認識精度の向上が期待できない多数決処理するのは費用対効果に見合わない可能性が高いことから、多数決処理を行わず、区分ごとに既存 OCR ソフトを使い分ける手法が最適と考えられる。

2. その他精度向上策の検討

(1) レイアウト認識と文字認識の組み合わせ検討結果

レイアウト認識の文字領域の認識精度が最も高い OCR ソフトと、文字認識精度が最も高い OCR ソフトの組み合わせによる、テキスト化精度の向上の可能性を検討した。

図 5-2　レイアウト認識と文字認識の好成績ツール組み合わせ

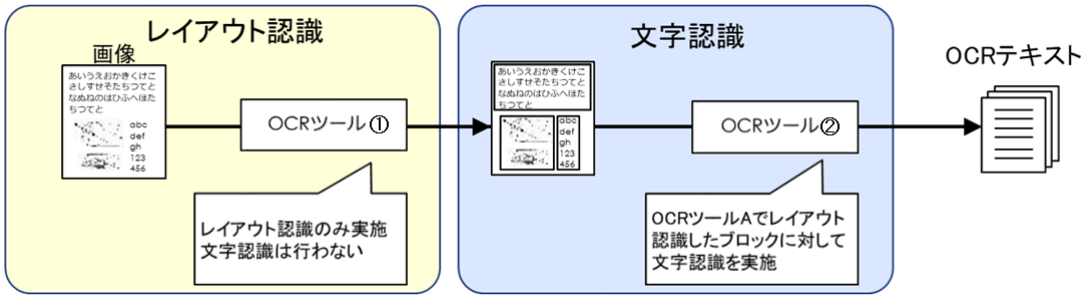


図 5-3　レイアウト認識と文字認識の好成績ツール組み合わせの原理

OCR ソフト	レイアウト認識	文字認識	認識精度
①のみ	①（**90%**）	①（80%）	90%×80%＝72%
②のみ	②（70%）	②（**90%**）	70%×90%＝63%
組み合わせ	①（**90%**）	②（**90%**）	90%×90%＝**81%**

以下に、レイアウト認識と文字認識の結果を示す。

表 5-2　OCR ソフトのレイアウト認識精度及び文字認識精度

	OCR ソフト			
	D	B	C	
レイアウト認識 （文字領域認識精度）	71.3%	41.2%	80.3%	①
	OCR ソフト			
	A	B	C	
文字認識精度	70.9%	85.0%	86.4%	②
		①×②	69.4%	

レイアウト認識の文字領域認識精度と、文字認識精度はいずれも OCR ソフト C が最も高かった。

したがって、基本的に複数ツールの組み合わせは行わず、OCR ソフト C 単体の OCR 処理が、最もテキスト化の精度が高くなるという結果となった。

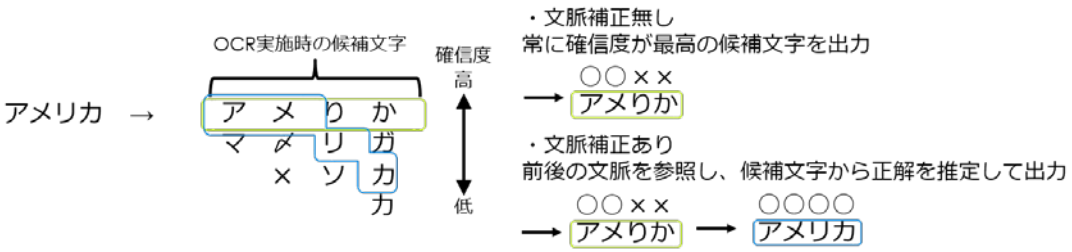
ただし、第 4 章の結果より、区分 B-③、および B-④については、OCR ソフト B の文字認識精度が OCR ソフト C よりも 30%以上高く、OCR ソフト C でレイアウト認識を行い、OCR ソフト B で文字認識を行うことによって、テキスト化の精度が高くなる可能性がある。

(2) 文脈補正の検討結果

複数の既存 OCR ソフトによる文字認識の結果を、単純な多数決処理ではなく、候補文字として扱い、別途文脈補正技術を導入することでテキスト化の精度の向上を図るシステム開発を想定した場合の基礎的な調査を実施した。

ここで、文脈補正とは、以下に示すように、確信度が最高の候補文字を出力するのではなく、前後の文脈を参照して候補文字から正解を推定して出力する手法である。

図 5-4　文脈補正による文字認識精度向上の原理



32 区分ごとの OCR ソフトの文字認識率と、候補内正解文字出現率を以下に示す。

表 5-3　既存 OCR ソフトの文字認識率及び候補内正解文字出現率

区分 No	検証用データ数	OCR ソフト			ABC 候補内正解文字出現率
		A	B	C	
A-①	13	72.4%	84.7%	93.2%	97.0%
A-②	13	67.3%	76.5%	88.1%	93.1%
A-③	1	79.2%	94.1%	90.6%	96.5%
A-④	1	85.2%	94.5%	91.8%	98.4%
A-a-①	13	80.9%	91.4%	95.6%	97.0%
A-a-②	1	76.9%	93.8%	90.4%	98.1%
A-a-③	1	85.1%	96.7%	87.1%	99.1%
A-b-①	13	64.1%	86.4%	86.3%	95.3%
B-①	14	40.2%	67.6%	70.3%	85.9%
B-②	3	50.9%	54.6%	83.7%	89.8%
B-③	1	48.6%	71.7%	40.8%	84.5%
B-④	13	27.3%	64.0%	22.2%	73.0%
B-⑤	26	86.1%	93.1%	97.0%	98.7%
B-⑥	1	87.8%	94.3%	94.8%	97.8%
C-①	14	80.4%	91.0%	94.7%	97.5%
C-②	13	85.3%	92.6%	95.1%	96.9%
C-③	1	82.4%	93.6%	93.9%	97.7%
C-④	1	87.5%	95.5%	92.2%	98.4%
H-a-①	10	77.8%	91.6%	96.3%	97.5%
H-a-②	4	76.0%	90.8%	95.5%	97.5%

区分 No	検証用データ数	OCR ソフト			ABC 候補内正解文字出現率
		A	B	C	
H-a-③	1	75.1%	84.6%	98.0%	98.3%
H-b-①	1	52.3%	86.3%	94.1%	97.6%
D-①	1	47.7%	59.5%	83.8%	87.1%
D-②	3	49.6%	45.1%	79.3%	83.4%
D-③	1	19.6%	37.5%	49.1%	67.3%
D-④	14	58.9%	86.4%	73.5%	93.4%
D-⑤	27	88.7%	95.9%	97.6%	98.7%
D-⑥	1	88.1%	94.8%	94.2%	98.0%
E-①	13	59.5%	77.4%	87.4%	93.2%
E-②	1	66.8%	88.8%	90.1%	95.5%
E-③	1	76.8%	95.3%	86.9%	97.5%
E-④	1	76.5%	92.6%	89.9%	98.2%
合計		81.6%	90.5%	91.4%	94.3%

95.3% 候補内の正解文字出現率が単一のOCR文字認識率より高い区分

94.8% もっとも文字認識率が高かったOCRツール

候補内正解文字出現率は、OCR ソフト A～C のいずれか 1 種類以上の文字認識結果が正しかった場合を正解文字とし、以下の計算式で算出した。

式 5-1　候補内正解文字出現率

候補内正解文字出現率　＝　正解文字数/総文字数

区分 B-①、B-③、D-③の旧字、旧仮名遣いの要素が存在する 3 区分では、候補内正解文字出現率はその区分で最も認識精度が高い単一の OCR ソフトの文字認識率より 10%以上向上しており、これら 3 区分においては、適切な文脈補正技術を使用することによって、テキスト化の精度が向上する可能性がある。

しかし、上記以外の区分については、精度の大幅な向上は見られず、効果は限定的であると考えられる。

AI－OCR による文脈補正については 6 章で述べる。

3. 小括

テキスト化精度向上策の検討として、複数の既存 OCR ソフトを組み合わせた場合の検討を行った結果、組み合わせによって達成可能な文字認識精度の向上幅は大きくなく、32 区分毎に最適な既存 OCR ソフトを使い分けるのが最も効果的であることが判明した。

第 6 章 人工知能技術を活用した OCR の検証

文書画像のテキスト化については、いわゆる OCR ソフトを活用することが通常であり、市販 OCR ソフトを活用したテキスト化精度の検証や、精度向上に向けた検討については第 5 章で論じた通りである。

本章では、人工知能技術を活用したテキスト化について検証を行い、さらなる認識精度の向上の可能性を検討する。

1. AI-OCR 検証概要

人工知能技術を活用した AI-OCR の検証は、以下のとおりに行った。

具体的には、第 2 章で抽出、作成した、2,019 件の実験用データのうちの 1,797 件の教師データを用いて AI-OCR を構築し、前記実験用データのうちの 222 件の検証用データに対して、構築した AI-OCR による文字認識を行い、AI-OCR の文字認識精度を検証した。また、教師データを漸増させていった場合における文字認識精度の変化の傾向も検討した。

AI-OCR の検証方法は、以下のとおりである。

① 文字認識精度の検証（文脈補正を含む）
　—1 文字認識：1 文字単位で切出された画像に対する文字認識
　—行内文字認識：1 行単位で切出された画像に対する文字認識
② 効率的な学習データ作成方法の検討
　—教師データの漸次増加
　—教師データを徐々に増やしていくことに対する効果検証
　—能動学習
③ 精度向上に必要な教師データ量の推定
　—文字種毎の字形数と精度による推定
　—全種別・全年代、及び、年代・画質別に実施

(1) 1 文字認識と行内文字認識

1 文字認識及び行文字認識の 2 種類の検証を実施した。

1 文字認識とは、文字毎の矩形を人手又はレイアウト認識で予め設定し、そのうえで各文字単位の矩形に対して AI-OCR による文字認識を実施するという方法であり、行文字認識とは、行単位で切り出した矩形を人手又はレイアウト認識で予め設定し、そのうえで各行単位の矩形に対して、AI-OCR による文字認識を実施するという方法である。

一文字認識は、検証用データに含まれる文字（全 343,374 文字）について、教師データによる学習によって構築した AI-OCR を用いて一文字ごとに文字認識を行い、文脈補正を行った上で、文字認識で得られた文字の文字コードと、正解文字の文字コードとを比較することで評価した。行内文字認識は、検証用データに含まれる行（全 18,910 行）の文字列につ

いて、前述の AI-OCR を用いて行毎に文字認識を行い、文脈補正を行った上で、文字認識で得られた文字の文字コードと、正解文字の文字コードとを比較し、第 4 章で用いた、レーベンシュタイン距離を使用した計算式により評価した。

なお、実際に公報のテキスト化を行う場合には、処理速度等を考慮するに、1 文字認識より行内文字認識の方が現実的であると考えられる。

本検証において、1 文字認識と行内文字認識の両方を実施した理由は、両者の検証結果を比較することによって、文字認識を誤る原因が、文字認識自体にあるのか、行単位で文字認識を行ったことにあるのか、区別して把握するためである。

(2) 文字種の扱い

半角英数字と全角英数字は同じ文字として扱った。データエントリー業務の観点から統一して扱う方が効率性の面から優れていると考えられるためである。

一方、アルファベットや仮名の大文字と小文字は表記上区別して書かれたと想定されるため、区別して扱った。

(3) 適切な教師データ量と効率的な学習についての検討

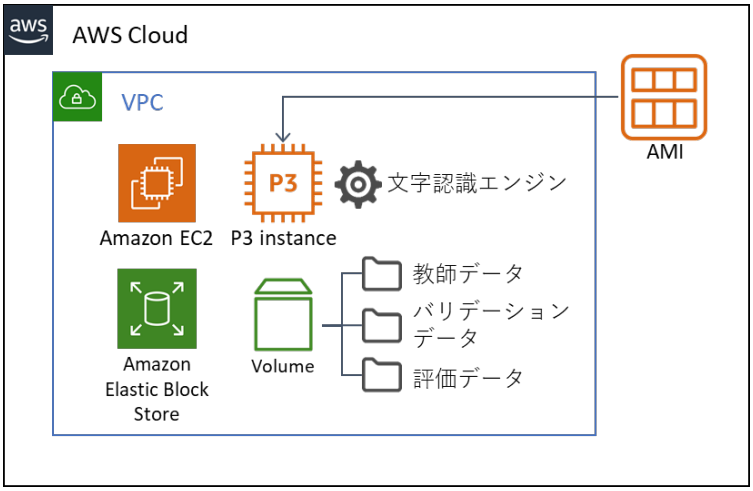
AI-OCR による文字認識の精度を高めるためには、教師データの量を増やしていくことが効果的と考えられる。しかし、教師データ量を無限に増やすことは物理的にも費用対効果の面でも不可能であるため、適切な教師データ量の検討と効率的な教師データの作成方法の検討が必要となる。そこで、教師データの適切な量と、特に能動学習を含めた効率的学習データの作成についても検討を行った。

能動学習とは、教師データの量を増やして文字認識精度の改善を図る際に、学習効率を上げるためにどのようなデータを優先的に学習していくかという点を AI に能動的に選択させる優先学習のことである。優先学習の方針については、その方針の検討をも AI にゆだねるディープラーニングの手法もあるが、文字認識の成否を分ける境目となる、重要性が高いデータと確信度の低いデータが相当程度一致するという仮説の下、文字認識精度が低い文字種から優先的に文字単位で学習していくという基本方針を設定し、それに基づいて具体的に学習するデータを AI に選択させる能動学習を行った。

本検討で用いたシステム構成は以下の図 6-1 のとおりである。文字認識ソフト以外は、合理的な費用で一般的に調達可能な構成とした。

文字認識ソフトについては、様々な画質に対応可能するために、柔軟にカスタマイズが可能な当社開発のソフトを用い、かすれ・つぶれに対応可能なモデルを新たに設計した。

図 6-1　システム構成



Amazon クラウド（AWS）

―EC2 p3dn.24xlarge インスタンス

CPU	48※vCPU Intel(R) Xeon(R) CPU (CPU メモリ 768 GB)
GPU	NVIDIA Tesla V100（VRAM 32GB）

（※学習実施当時の令和 2 年 3 月時点では、96vCPU のインスタンスしか存在しない）

OS/ミドルウェア

―Amazon Linux

・Deep Learning AMI (Amazon Linux) Version 26.0

―Python 3.6.5

・深層学習ライブラリ Chainer 6.1.0

―CUDA 10.1

文字認識ソフト

―凸版印刷開発の文字認識エンジン

ベース VM イメージ：　Deep Learning AMI (Amazon Linux) Version 26.0

OS:　Amazon Linux2 (64bit)

開発言語：　python 3.6.5

GPU 向けツールキット：　nVidia CUDA 10.1

フレームワーク：　Chainer 6.1.0

文字認識ソフトの処理速度に関して、学習速度（教師データについて学習収束にまでかかった時間）及び予測実行速度は、以下のとおりである。

予測実行速度は、現状以下の速度となっているが、今後さらなる高速化が可能と考えられる。具体的には、認識に用いるソフトのパラメータのチューニング及び処理プロセスの改善が可能と考えられ、例えばデータ転送方法を最適化することで相当程度の処理速度改善が可能と考えられる。

―学習速度（教師データ（2,569,995 字形）について学習収束までにかかった時間）
・1 文字認識：516.3時間
・行内文字認識：693.2 時間

―予測実行速度
・1 文字認識：300 文字以上 /秒
・行内文字認識：100 文字以上 /秒

本検証に用いた文字認識用のデータおよび文脈補正用のデータは以下のとおりである。文字認識用のデータについては、第 2 章で作成した、正解データを付与した実験用データ 2,019 件を分割して使用した。文字認識の対象としては、JIS 第 1 水準、第 2 水準をベースとした 6,739 文字種を対象とした。

具体的には、学習フェーズでは、実験用データのうちの教師データ 1,797 件について学習用データを設定し、さらにそれらを教師データと調整用のバリデーションデータに分割した。具体的には、教師データのうち 1,718 件から、AI-OCR が認識対象とする 6,739 文字種に含まれる 3,158 文字種（2,569,995 字形）を、AI-OCR 学習用教師データの作成に用いた。しかし、これだけでは AI-OCR による文字認識対象の文字種の学習データとして不十分であったため、6,739 文字種の AI-OCR 学習用教師データの画像を追加で用意し、学習に利用した。加えて、AI-OCR 学習用データをさらに拡充するため、データ拡張手法[7]も併用した。

また、教師データのうち、公報の種別及び発行年代において 2%程度の文献を抽出して若干の予備値を加えた、上記 1,718 件を除いた残りである 79 文献・1,850 文字種（118,089 字形）について、調整用のバリテーションデータとした。バリテーションデータとは、学習が適切に行われているかを確認するためのデータのことである。学習が適切に進んでいない場合はニューラルネットワークの層数や各層に含まれる特徴マップの数など（これらを総称してハイパーパラメータと呼ぶ）を修正するチューニング作業を行い、AI-OCR の認識アルゴリズムについて学習を行った。

また、検証フェーズでは、第 2 章で抽出した検証用データ 222 件・2, 301 文字種（343,374 字形）に対して、AI-OCR による文字認識を実施することで、その効果を検証した。また、

[7] データ拡張手法：様々な画像処理を加えて教師データのバリエーションを増やすことで、未知データに対する識別能力を向上させる手法。一般的には、ノイズ付加、コントラスト変更、ガンマ変換、平滑化、拡大縮小、反転、回転、平行移動、部分マスク、トリミング、変形、変色などが用いられる

認識検証においては、本検証用に生成した文脈補正用言語モデルを認識エンジンに読込んだうえで認識精度の検証を行った。

文脈補正を実施するための言語モデルの生成にあたっては、特許庁から貸与を受けた特許及び実用新案の各種公報（ 1971 年から 1993 年、1,302,176,534 文字）及び、JAPIO（日本特許情報機構）所有の公開特許公報（ 2009 年から 2018 年、66,295,543,852 文字）を用いた。

また、本検証で文字認識の対象とする公報のイメージデータには、縦書き、横書きの 2 種類の行が存在することから、縦書き、横書きは別々で学習を行った。

―文字認識用データ

・学習用データ[1,797 文献]

・教師データ[1,718 文献]　3,158 文字種、2,569,995 字形

・バリデーションデータ[79 文献]　1,850 文字種、118,089 字形

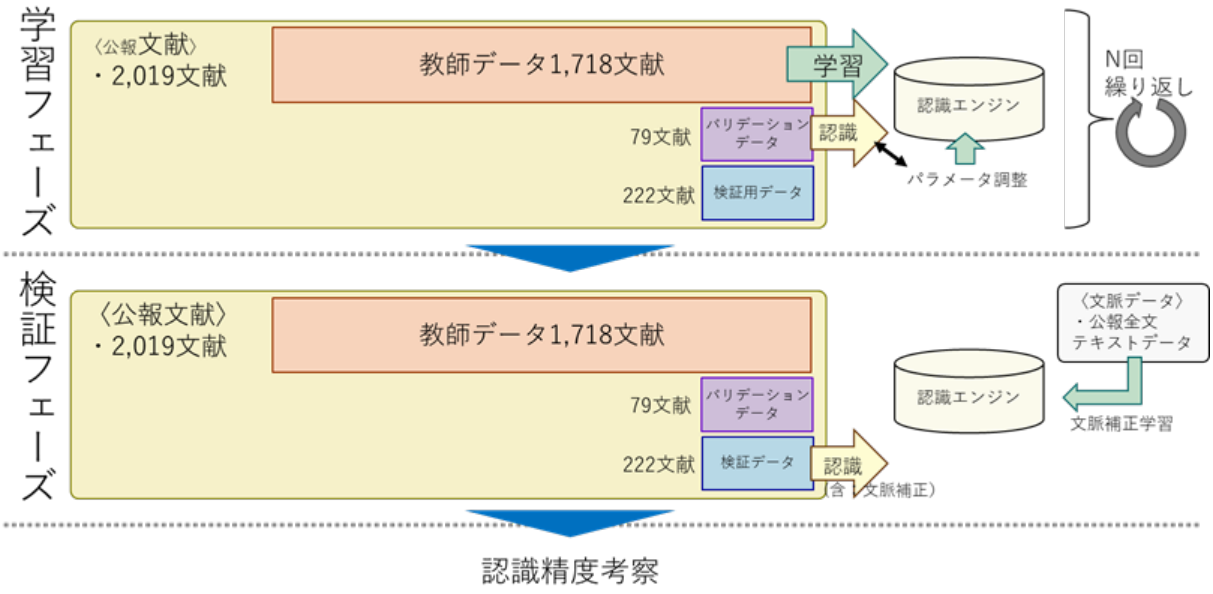
―検証用データ[222 文献]　2, 301 文字種、343,374 字形

―文脈補正用データ

・特許庁貸与データ　特許及び実用新案公報 1971 年～1993 年、1,302,176,534 文字

・JAPIO 所有データ　公開特許公報 2009 年～2018 年、66,295,543,852 文字

図 6-2　文字認識検証フロー（1 文字認識・行内文字認識）



2. AI-OCR 検証結果および考察

(1) 文字認識精度結果と効率的な学習データ生成

a) 文字認識精度結果

まず、1 文字認識の精度は、97.52% [334,871 正解/343,374 文字]であった。これは、後述する文脈補正プロセスを実施したうえで算定された結果である。このうち、認識対象に設定した 6,739 文字種に限定した場合の正解率は、98.35% [334,871 正解/340,472 文字]であった。

最も認識率が低い 78.02%であった 1948 年の公告実用新案と最も認識率が高い 98.91%であった 1970 年代の公告実用新案では、20.89%の認識率の差があった。これは 1948 年の公告実用新案の認識率が例外的に低い（同様のフォーマットである 1948 年 公告特許と比較しても明らかにつぶれとノイズが酷い事に起因）ためと考えられ、他の年代種別区分ではいずれも 90%を超える認識精度が記録されている。今後の更なる改善を含めて考えれば、例外的に低い認識率となる年代のみを除外すれば、人工知能 OCR によるテキスト化の取り組みは有用といえる。

なお、文字認識にあたっては、特許庁から貸与を受けた公開広報のテキストデータを利用して文脈補正を施したうえで処理を行った。また、文字認識と文脈補正の処理は、ソフトウェア的に統合されており個別算出にはソフトウェアの改変が必要となるため、本検証では文脈補正後の精度のみを利用した。

表 6-1　区分毎の認識精度（1 文字認識）

区分	精度(%)
A-①	98.89%
A-②	96.03%
A-③	97.47%
A-④	94.88%
A-a-①	98.10%
A-a-②	97.39%
A-a-③	97.96%
A-b-①	97.50%
B-①	95.94%
B-②	96.34%
B-③	91.43%
B-④	94.99%
B-⑤	98.59%
B-⑥	98.39%
C-①	97.65%
C-②	98.11%
C-③	98.28%
C-④	97.04%
H-a-①	98.70%
H-a-②	98.38%
H-a-③	96.93%
H-b-①	98.26%
D-①	94.64%
D-②	96.56%
D-③	**78.02%**
D-④	96.23%
D-⑤	**98.91%**
D-⑥	98.54%
E-①	97.31%
E-②	95.95%
E-③	96.38%
E-④	95.28%

次に、行内字認識精度は、93.61% [320,618 正解/342,504 文字]であった。同じく認識対象に設定した 6,739 文字種に限定した場合の正解率は 96.79% [320,618 正解/331,316 文字]であった。行内文字認識では、複雑な予測を行うことになるため、当然に認識精度は 1 文字認識よりも下回ることになる。

年代、文書種別ごとの精度のばらつきは、1 文字認識の場合よりも大きくなる傾向があった。最も認識率が低い 53.35%であった 1948 年の公告実用新案と，最も認識率が高い 97.76%であった 1970 年代の公開特許では、44.41%の認識率の差があった。1948 年の公告実用新案の認識率が非常に低い理由は当該文献がつぶれとノイズが酷いことに起因していると考えられるが、その他の区分でも、70%〜80%台の認識率にとどまる区分が複数存在しており、これらの区分は事前のレイアウト調整など、一定の前処理を行ったうえでなければ実用水準の認識精度は出ないと推測される。少なくとも 90%を超える文字認識精度がなければ、文字認識処理後に人手で補正する必要が生じ、多大なコストがかかることになる。

また、文字認識にあたっては、1 文字認識と同様に、特許庁から貸与を受けた公開広報のテキストデータを利用して文脈補正を施したうえで処理を行った。ここでも前述の理由により、文脈補正後の精度のみ利用した。

表 6-2　区分毎の認識精度（行内文字認識）

区分	精度(%)
A-①	**97.76%**
A-②	90.64%
A-③	94.07%
A-④	86.07%
A-a-①	95.28%
A-a-②	95.32%
A-a-③	96.00%
A-b-①	90.40%
B-①	92.89%
B-②	92.36%
B-③	80.26%
B-④	88.81%
B-⑤	95.31%
B-⑥	96.47%
C-①	94.61%
C-②	92.50%
C-③	95.57%
C-④	87.76%
H-a-①	95.78%
H-a-②	95.78%
H-a-③	94.92%
H-b-①	85.65%
D-①	90.89%
D-②	91.93%
D-③	**53.35%**
D-④	92.58%
D-⑤	96.74%
D-⑥	95.88%
E-①	93.48%
E-②	86.55%
E-③	89.39%
E-④	74.51%

総じて、1 文字認識においても行内文字認識においても、市販の既存 OCR ソフトに比して、AI-OCR は、高い認識精度での処理結果となった。しかしながら、特に行内文字認識においては、十分な実用に耐える精度とまでは言えないことから、今後、認識ミスの分析を含めたさらなる認識率の向上を試行錯誤し、比較的認識率が高い年代文献種別から優先順位をつけてテキスト化を図っていく必要がある。

b）効率的な学習データ生成

今後テキスト化を実用段階に進めていくにあたっては、効率的な学習データの生成が不可欠である。そのため、人工知能自身に、一定の方針に従ってどのような学習データを追加すべきか選択させる能動学習の手法が、効率的な学習データ生成に有用か否かについて、検討した。

b-1）能動学習による教師データの拡充の有用性検証

教師データの拡充は、文字認識精度の向上に一定の寄与を与えると考えられるが、文字認識精度が高くなるに伴い、教師データ量の増加による効果は低減する（文字認識精度が頭打ちになる）ことが予測されるため、費用対効果のバランスを適切に取ることが重要である。

そこで、効率的に教師データを拡充していくために、どのような教師データを拡充すべきかという点について AI 自身に選択させる能動学習の方法を採用することの有用性を検証した。

本検証では、ランダムに教師データを増加させる漸次増加学習と、特に認識精度が低いものを優先的に学習対象とする方針の下で AI に学習対象を選択させる能動学習の二つの方法を比較した。

b-2）漸次増加学習と能動学習の比較方法

本検証では、特許庁から貸与を受けた各種公報の貸与データから、発行年代及び画質で特徴的な発行年代及び発行種別を 4 種類抽出し、それぞれについて、比較実験を行った。

抽出種別①　発行年代：戦前（1930 年代）[300 文献（教師 290 件：評価 10 件）]
抽出種別②　発行年代：戦後（1950 年代）[300 文献（教師 290 件：評価 10 件）]
抽出種別③　画質：底（かすれ・つぶれの両方が多い）
[250 文献（教師 238 件：評価 12 件）]
抽出種別④　画質 高（かすれ・つぶれの両方が少ない）
[250 文献（教師 238 件：評価 12 件）]

教師データを増加させる方法は、教師データを 1/10 ずつ増やして学習を行い、精度を確認しながら、それを 10 ステップ繰り返すというものである。

漸次増加学習では、教師データをランダムに選択し、能動学習では、教師データを選択する前に、その時点の AI-OCR を用いて選択対象のデータを文字認識し、AI-OCR の出力する確信度の数値が低い文字を優先して選択するという方法を採用した。（初回の学習は両方ともランダムで選択）

そして、4 つの抽出種別に対して、ステップ毎に、漸次増加学習のモデルと能動学習のモデルの精度を比較する検証を行った。

その結果、以下グラフのとおり、能動学習は、漸次増加学習よりも低い精度にとどまるという結果となった。この傾向は、抽出した 4 種別においてすべて同じであった。すなわち、この方法による能動学習では成果が上がらないことが判明した。

なお、10 ステップ目では全教師データで学習するため、同等の精度になっている。

図 6-3　かすれ・つぶれの両方が多い種別 A-b 及び E

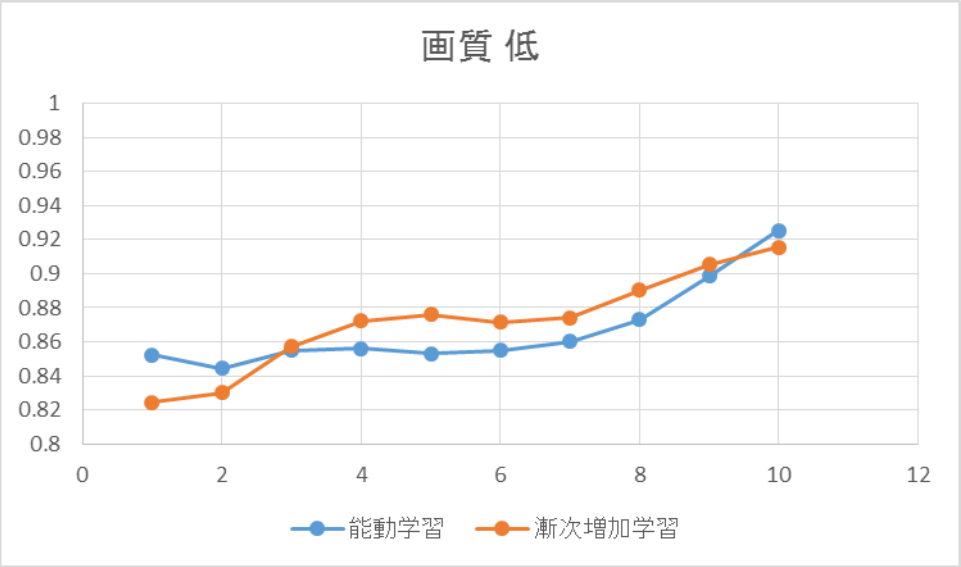


図 6-4　かすれ・つぶれの両方が少ない種別 B 及び C

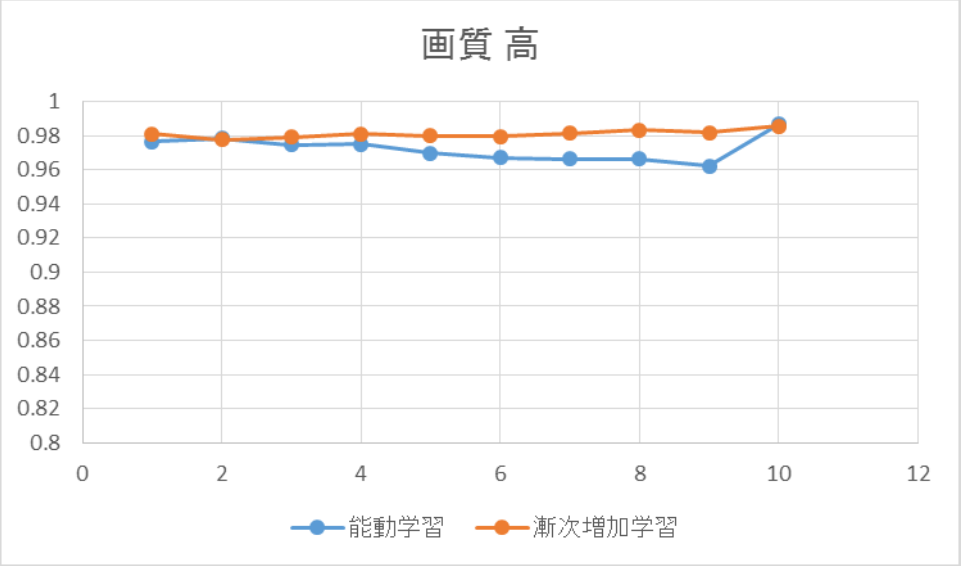


図 6-5　戦前（30 年代の種別 B 及び D）

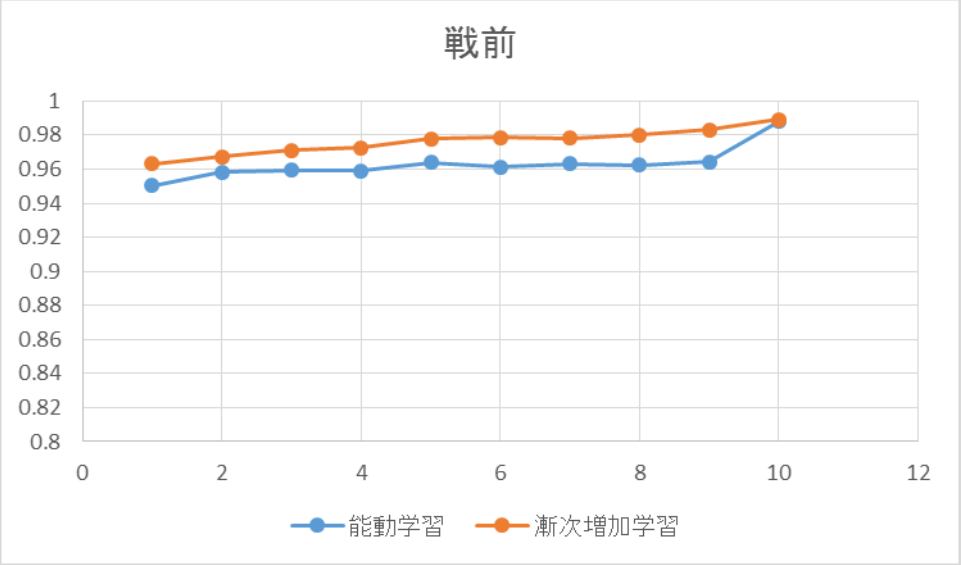
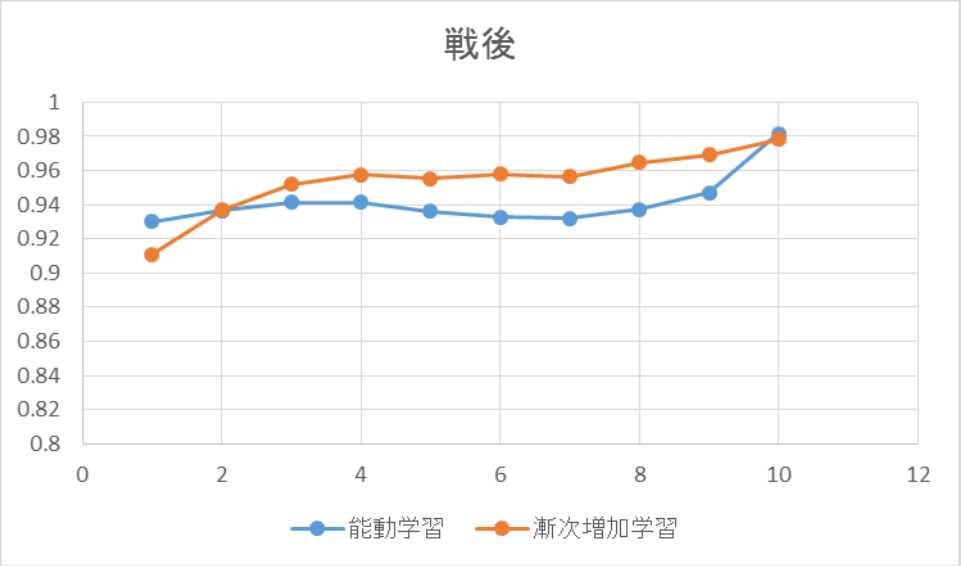


図 6-6　戦後（50 年代の種別 B 及び D）



b-3）能動学習において効果が出なかった原因

このように能動学習における効果が出なかった原因は、AI-OCR の確信度の低いものを優先的に学習データとして投入した点にあると考えられる。このようなデータを優先させたのは、認識の成否を分ける境目となる重要性が高いデータと確信度の低いデータが相当程度一致するという仮説によるものであったが、実際にはそのようになっていなかった。

すなわち、確信度が低いものは、実際には、画像がつぶれてしまっているなど、人間でも認識困難な文字である場合が多く、成否の境界ではなく明らかに認識できない文字であったため、教師データを増やしても認識精度が向上しなかったものと考えられる。

また、認識が困難な同一文字種ばかりが選択される傾向もあり、やはり学習の効果が低くとどまってしまった。

一方、ランダムによる漸次増加であっても、一定の効果は発揮されることから、能動学習にこだわらず、単純に教師データ数を増加させるという方法を選択することも現実的と考えられる。今後、能動学習の取組みを継続する場合には、単純に認識精度や確信度が低いものを優先するのではなく、認識候補には正解が含まれていたが最終的には正解を選べなかったデータを選択させるなど、より高度な優先順位付けを行ったうえで能動学習をさせていく必要があると考えられる。

能動学習をさせる際に、どのような方針で優先的に教師データを増やしていくかというアルゴリズム自体についても、今後さらに最適なものを検証し、チューニングしていく必要がある。

c）目標認識精度の達成に必要な教師データ量の推定

教師データの数を増加させるにつれて、文字認識精度は向上するが、その向上幅は文字認識率が高くなるほど低減していくため、費用対効果の観点からは、目標認識精度を設定して、それに向けてどの程度の教師データが必要となるかを予め推定することが必要である。

そこで、全検証用データに対して 1 文字認識を行い、99.00%の精度を目指した場合に、現状の何倍の教師データが必要かを推定した。具体的には、検証用データの文字種毎の教師データ数と誤認識率の関係から推定した。使用した教師データは、第 6 章 1.c)で教師データから選択した 1,718 件、約 257 万字形である。また、公報の発行年代又は画質別にも同様の推定を行った。具体的には、文字種毎の教師データ数と認識精度の関係を、誤認識率と教師データ数の関係に式 6-1 にて近似した。ここで、近似曲線はデータの性質と実際のデータ分布から冪乗近似を選択し、θ は勾配法[8]を用いて求めた。

式 6-1

精度：p
誤り率：$E = 1 - p$
認識クラス数：　N
最大誤り率：　$E_{max} = (N-1)/N$
教師データ数：$x \in \mathbb{N}^+$

$$E = E_{max}\frac{1}{(x+1)^{\theta}} \qquad (\theta > 0)$$

得られた近似式を用い、下式よって目標精度に対して現在の教師データ量の何倍が必要となるかを計算した。下式は、教師データ数 x が多く、対象は教師データが存在する（x がゼロより大きい）文字種のみとしていることから、上記の式において、$E \fallingdotseq E_{max}/x^{\theta}$ という近似を行って求めたものである。ここで、C は識別器の認識対象で教師データが存在する文字種の現状の誤認識率を何倍にするかを示す値である。例えば、認識率 98.0%の識別器を 99.0%の精度に近づけるには、誤認識率を半分にすれば良いため C=0.5 とすればよい。

式 6-2

誤り率の目標倍率：C
データ数の倍率：k

$$k = \frac{1}{\sqrt[\theta]{C}}$$

式 6-2 を用い、一定の文字認識精度近くに到達するために必要な教師データ量の推定を

[8] 勾配法：行列計算で解析解を求めることできず単回帰分析・重回帰分析で解を求める場合にそのフィッティングのパラメータを推定するために用いられる代表的な方法。確率的勾配降下法、最急降下法などが存在する。

行った。

図 6-7　誤認識率と文字種毎の教師データ数

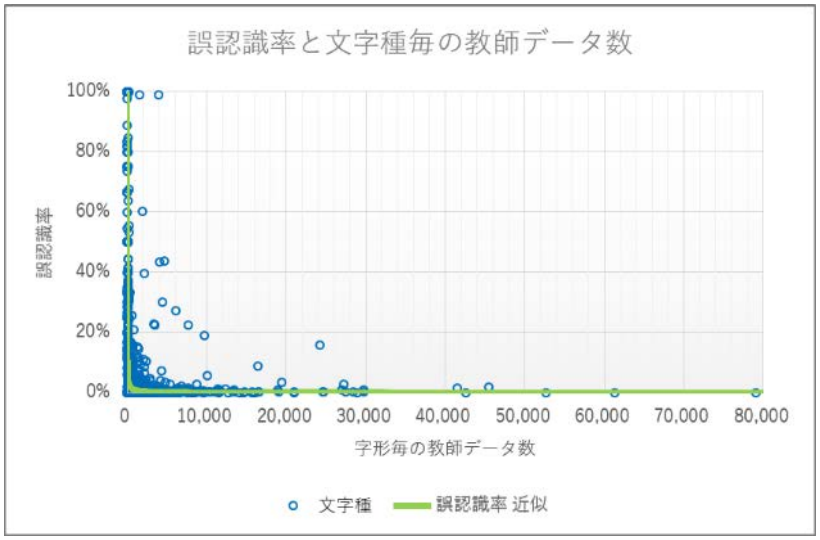
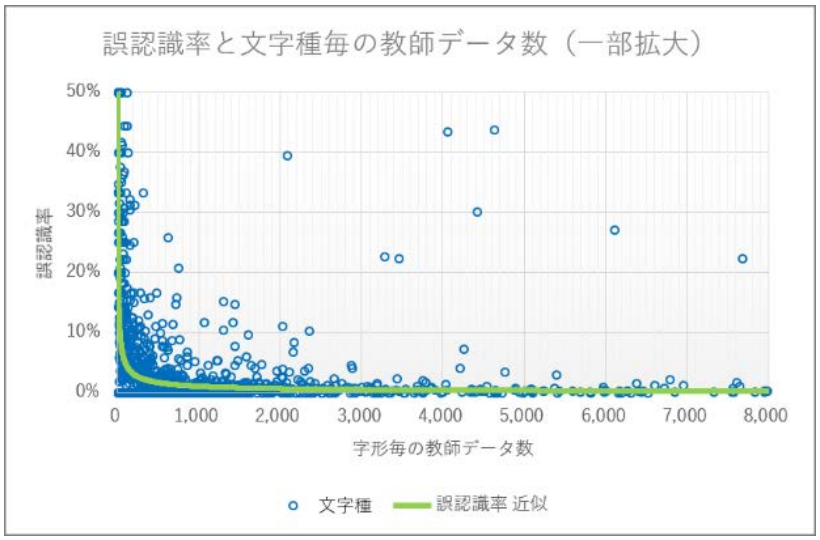


図 6-8　誤認識率と文字種毎の教師データ数（一部拡大）



上記のグラフは、縦軸が誤認識率（1-認識精度）であり、横軸が文字種ごとの教師データ数である。文字種毎の教師データが多いほど一般には誤認識率が低くなる。グラフ上にいくつかみられる例外値は、画像等の問題や、人間でも判別が難しい文字等、教師データが多くても認識できない文字種ということなる。

その結果、教師データ全体では、99.00%の認識精度を目標とする場合、現在の約 3.4 倍、874 万字形程度の教師データが必要であると推定された。1 文字認識よりも認識精度が下が

る行内文字認識においても 99.00%の認識精度を目標とすれば、現状の約 7 倍、1800 万字形程度の教師データが必要と推定される。

同様の推定を、公報の発行年代が戦前のもの（1930 年代の文献種別 B, 種別 D）と戦後すぐのもの（1950 年代の文献種別 B, 種別 D）のデータでも行ったところ、同じく 1 文字認識で 99.00%の精度を目標とする場合、戦前では現状の約 1.3 倍、戦後すぐでは現状の約 2.5 倍の教師データが必要と推定された。行内文字認識で 99.00%の精度を目標とする場合、戦前では現状の約 5 倍、戦後すぐでは現状の約 42 倍の教師データが必要と推定された。

画質の低いものと高いもののそれぞれについても同様の推定を行ったところ、画質の低いもの（かすれ・つぶれの両方が多い種別 A-b:80,90 年代 種別 E:70,80 年代）では、1 文字認識で 99.00%の精度を目標にした場合、現状の 3.2 倍の教師データが必要であり、画質の高いもの（かすれ・つぶれの両方が少ない種別 B:80,90 年代 種別 C:70,80 年代）では同じく現状の 1.1 倍の教師データが必要と推定された。行内文字認識で 99.00%の精度を目標とする場合、画質の低いものでは現状の約 10 倍、画質の高いものでは現状の約 5 倍の教師データが必要と推定された。

上記の時代・画質毎の推定結果は、本検証結果で得られた AI-OCR エンジンパラメータを基に、それぞれの分類に対して特化した再学習を行って精度を向上させるための教師データ量の目安を示している。

時代別推定

図 6-9　誤認識率と文字種毎の教師データ数：戦前

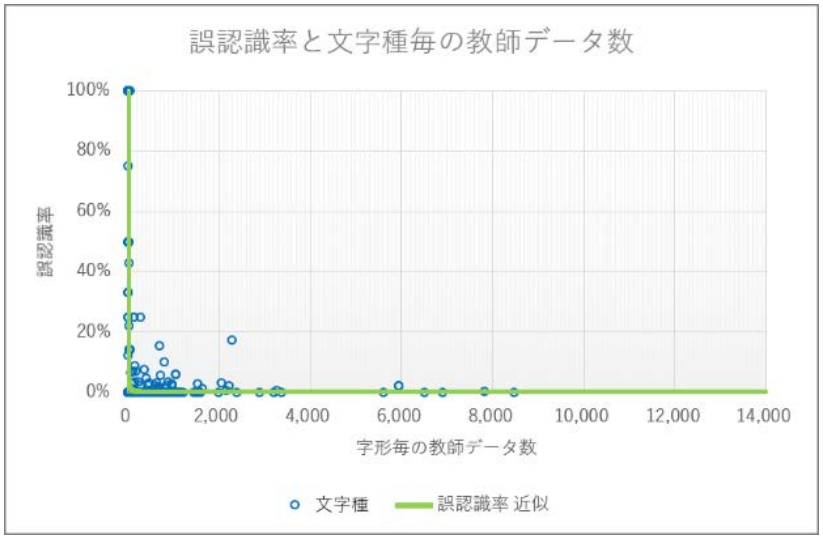


図 6-10　誤認識率と文字種毎の教師データ数：戦前（一部拡大）

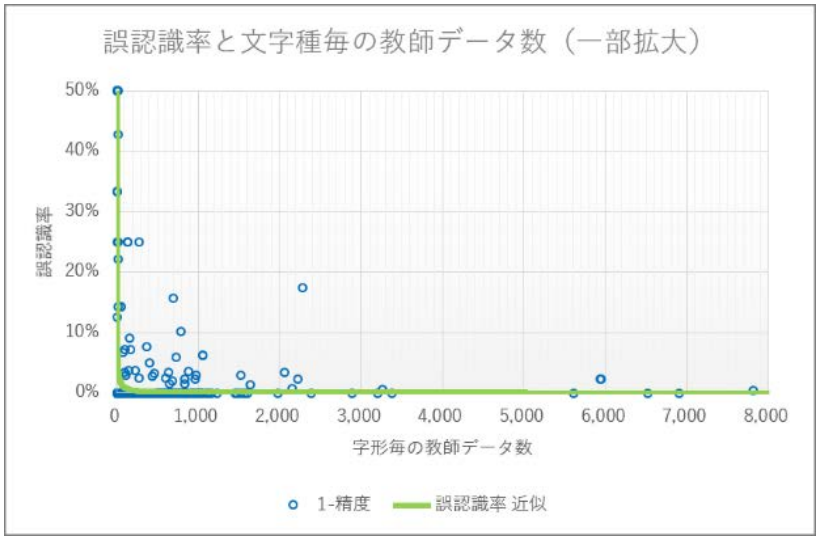


図 6-11　誤認識率と文字種毎の教師データ数：戦後すぐ

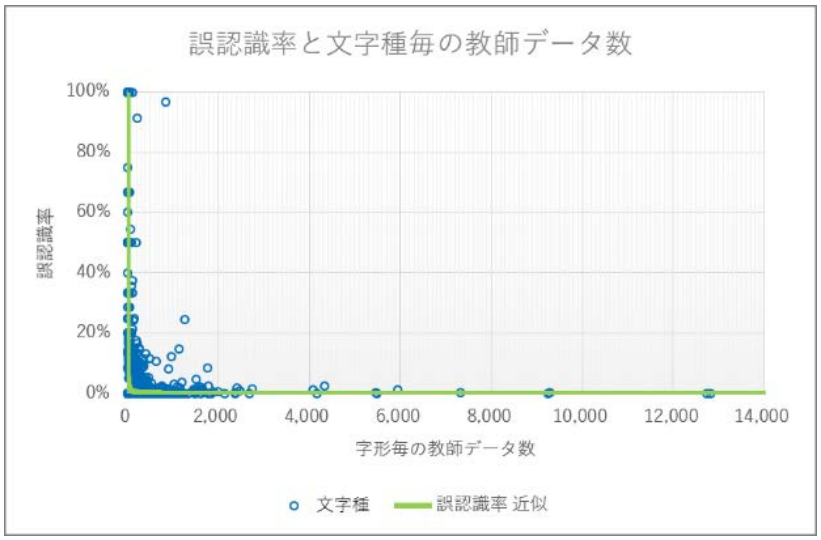
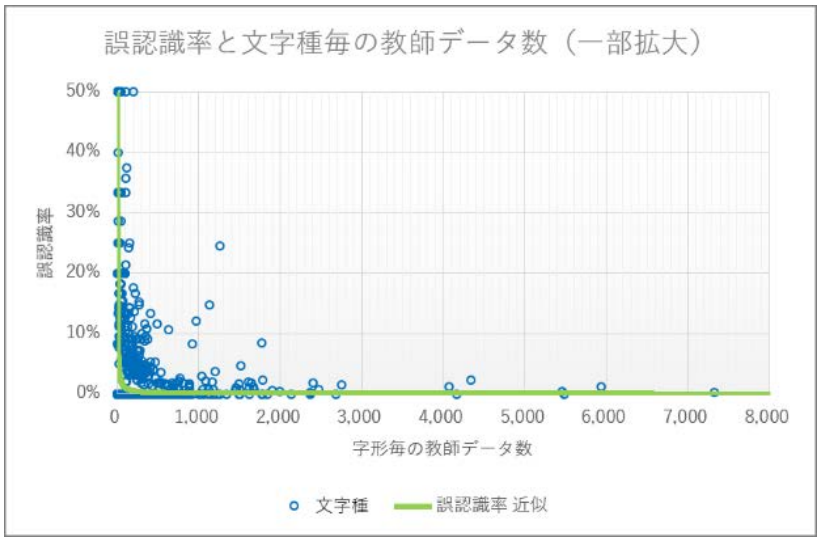


図 6-12　誤認識率と文字種毎の教師データ数：戦後すぐ（一部拡大）



画質別推定

図 6-13　誤認識率と文字種毎の教師データ数：画質低

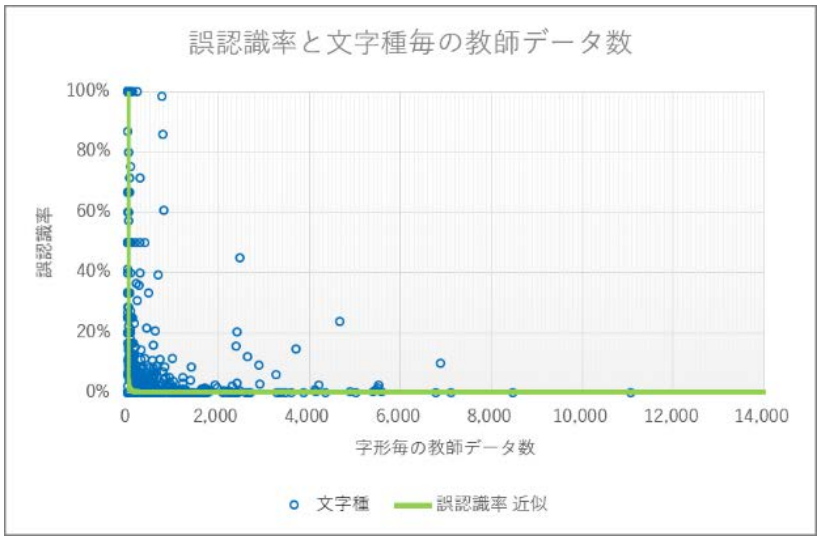


図 6-14　誤認識率と文字種毎の教師データ数：画質低（一部拡大）

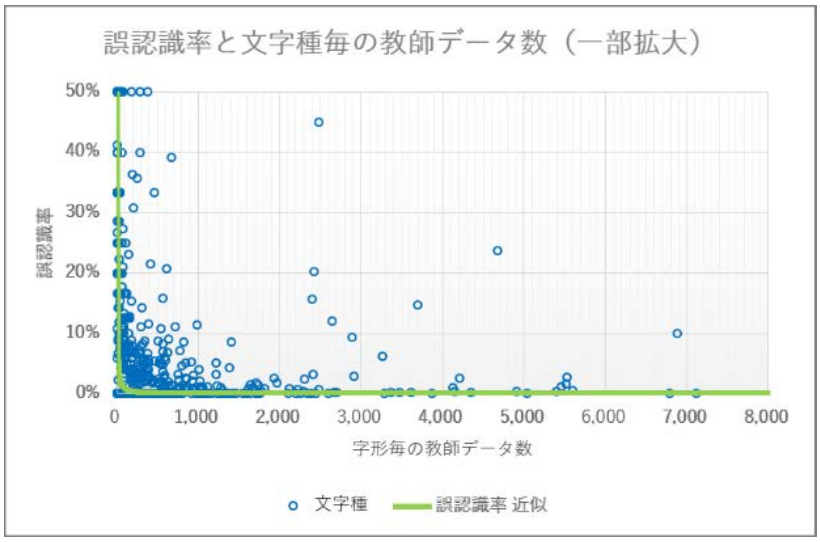


図 6-15　誤認識率と文字種毎の教師データ数：画質高

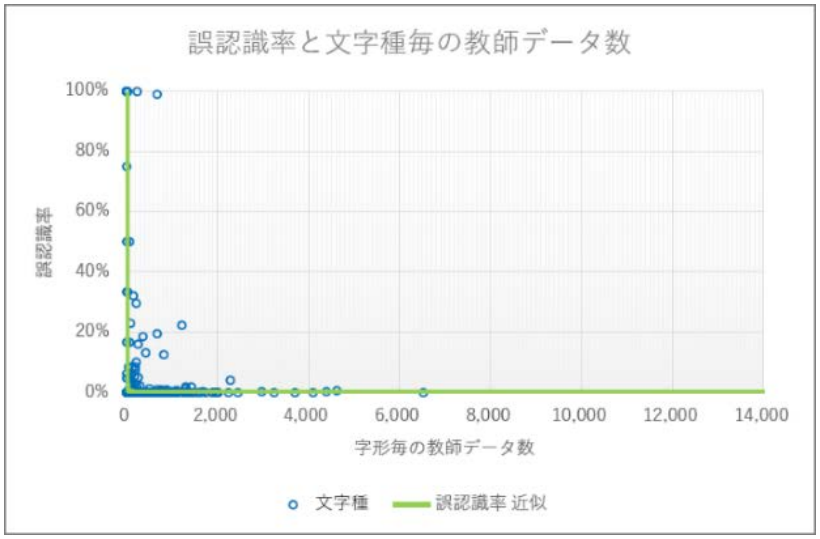
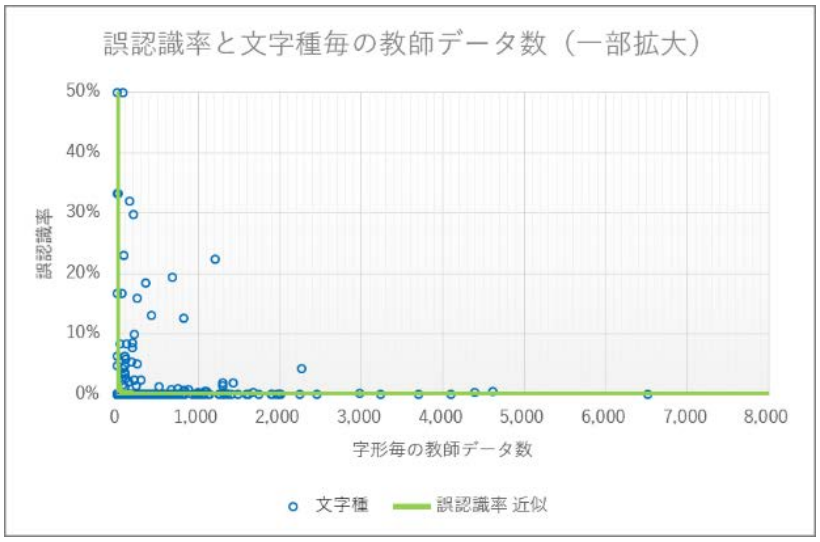


図 6-16　誤認識率と文字種毎の教師データ数：画質高（一部拡大）



今後の教師データの拡充にあたっては、限られた資源をどの教師データの増強に振り分けるかについて判断が必要となる。上記結果からすれば、1 文字認識ベースで 99.00%程度の認識率を目標とする場合には、必要となる教師データの数は、現実的に増加可能な範囲に含まれているものと考えられる。

(2) 認識精度向上策

単純な教師データ量の増加以外にも認識精度を向上させる方法は複数考えられる。そこで、ここではいくつかの認識精度向上策について考察する。

a) 認識精度向上のための 3 つの方法

今後、テキスト化処理の認識精度を向上していくための手段としては、①現在の認識対象となっている文字種に対する文字認識精度の向上、②認識対象の拡充、③文脈補正成功率の向上という 3 つの方法が考えられる。

そのうち、①現在の認識対象への文字認識精度向上については、後述 b) のとおり、認識ミスの原因に合わせた対策の実施が必要と考える。

認識対象の拡充としては、認識対象外で検証データに出現した 117 文字種 (5,802 字形、検証データ全体の 1.69%) について認識対象に含め、教師データを確保するという方法が考えられる。全体に占める割合は小さいため、この作業を実施することによる改善効果は必ずしも大きいとは言えないが、それでも一定の精度向上は見込まれ、比較的実施しやすい対応であるから、今後検討すべきである。

③文脈補正成功率の向上については、後述 c)のとおり、補正用データと対象データの年代・種別を合わせることや、文脈補正データの拡充が考えられる。

b) 認識ミスの分類と想定される対応策

今後の認識精度向上の取組実施に向けて、今回の認識実験において、どのような誤りが生じていたかについて分類すると、以下のような誤りの傾向があった。

まず、1 文字認識についての認識ミスを分類すると、以下のような 6 種類が代表的な認識ミスであった。

―濁点・半濁点の区別を間違えるもの

濁点・半濁点の有無に関しての誤認識をするケース。文字の潰れやノイズの影響も大きい。

―手書き文字であるため認識できないもの

想定していた対象文書は活字で記載された特許明細書であったが、一部に手書きで追記された箇所が存在した。ここでの検証は活字を用いて学習したモデルを採用しているため、手書きの箇所については基本的に上手く認識出来なかったものと推察される。

―傾き・文字回転のせいで認識できないもの

文字の傾きが想定を超えているケースである。極端な例では 90 度回転しているものも見受けられた。

―潰れ・ノイズ・かすれのせいで認識できないもの

広報文献の年代・種別によって文字の潰れやノイズ、かすれが酷い文献が存在した。中には人間でも判断できなさそうな程の状態になっている文字も存在している。

―類似形状の文字と誤認するもの

形状が似ている事によって誤認識するケースである。たとえば「5」と「s」のようなケースである．これ以外にも「【」と「〔」のようなものも散見された。中には、画像上では人間でさえ判別が困難な酷似形状のケースも含まれる。代表的な例として平仮名の「へ」を片仮名の「へ」として誤認識するような状況が挙げられる。

―ラベル誤りにより別の文字と認識してしまうもの

今回の検証では検証データの正解ラベルが誤っているケースが稀に存在した。

このうち、濁点・半濁点の区別については、文脈補正データの拡充によって相当程度改善するものと思われる。ただし、あるレベルを超えたノイズや潰れが存在しているものは、文脈補正によっても精度改善が難しいと考えられる。類似形状の誤認によるミスについても、同様に文脈補正の手法によって認識率向上が期待できる。

手書きで認識できないものや画像補正で改善しきれない潰れ・ノイズ・かすれが存在するものについては人力での処理に切り替えるほかないと考えられる。この場合、認識困難なデータをどのように抽出するかが今後の検討課題となる。これには例えば、識別器が出力する確信度の低いデータを認識困難文字として判断する等の方法が考えられる。

傾き・文字回転に起因して認識できないものについては、前処理の段階で傾き補正・回転補正をかけておくことが重要である。潰れ・ノイズ・かすれなどの画質の問題が併存しているものについては、傾き補正自体が有効に機能しない可能性が高いので、人力よる認識に切り替える対応を選択することになる。

潰れ・ノイズ・かすれ等の画質に関する問題については、その程度によって対応が異なる。問題が軽度にとどまる場合には、後述する人工知能を用いた画像補正や文脈補正によって改善できる場合があると思われるが、完全に文字がつぶれてしまっているといった重度の問題がある場合には、人手で認識を行うほかない。

総じて、画像に関する重度の問題が存在し、他の問題と併存しているような場合には、人手で認識を行う以外の対応がない。そのため、今後は、画像について潰れ・ノイズ・かすれ等の問題がどの程度存在するかによって、システムによる処理の可否を判断して分類する手法の開発が必要と思われる。システム開発等を伴わずにかかる分類を便宜的に行うとすれば、OCR 精度が特徴的に低く出る年代・種別のデータについて、画像に重大な問題があるものと推定し、テキスト化の優先順位を下げるといった対応が考えられる。

図 6-17　文字認識の誤りと分類

誤りモード	例	想定される対策
濁点・半濁点	×: ポ ○: ボ  ×: ペ ○: ベ  ×: て ○: で 	文脈補正データ拡充
手書き		リジェクト（人力）
傾き・ 文字回転	×: 凝 ○: 図  ×: 損 ○: 揖  ×: 目 ○: 四 	前処理 リジェクト(人力)
潰れ・ノイズ ・掠れ	×: 量 ○: 重  ×: ず ○: す  ×: Q ○: O  ×: 編 ○: 領 	前処理 文脈補正データ拡充 リジェクト(人力)
類似形状	×: 0 ○: 。  ×: へ ○: ヘ  ×: p ○: P 	文脈補正データ拡充
ラベル誤り	×:繼 ○:斷？  ×:徴 ○:微？  ×:眞 ○:真？ 	教師データ校正

次に、行内文字認識についても、同様に認識ミスを分類すると以下のようになる。

―行に傾きがあるために誤認識するもの

行（多くはページ全体）が想定以上に大きく傾いている事による誤認識ケース。文書のスキャニングの際に傾きが発生したものと推測される。

―化学式等で適切に認識領域を設定できないもの

行の中に化学式が含まれているケース。化学式に含まれる上付き・下付の数字や記号が他の文字よりも非常に小さくなってしまう事で誤認識をすると推測される。

―空白部分について空白と認識できないもの

離れた位置にある文字列を一つの文字列としていたり、文字間に空白を含む文字列を誤認識したりするケース。

―文字以外の映り込みによるもの

スタンプなど活字以外が行中に映り込み、それ自体を文字として誤認識たり、それと文字が重なってしてしまう事で他の文字と誤認識するケース。

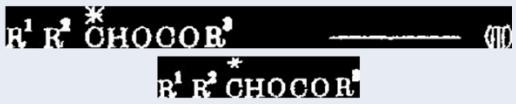
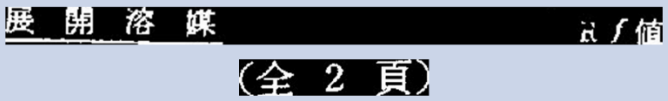
行の傾きについては一定の傾き補正をシステム上組み込んでいるが、傾きの程度が大きいものは認識に影響が出ている。傾き補正の前処理について更なる改善が必要と考えられる。

化学式のようなものは、専用のソフトウェア改良を行うか人力処理に切り替えざるを得ない。空白の処理も、現状のソフトでは十分な対応ができないため、ソフトの改良が必要と思われる。

文字以外の映り込みは、その程度によるが軽度のものであれば、ノイズの除去として前処理工程で改善できる可能性がある。重度の映り込みについては人力で修正せざるを得ないと考えられる。

文字認識位置の誤りは、認識矩形設定の際のミスであることから、学習用データを拡充することで矩形設定ためのデータ量が増えれば一定の改善があると考えられる。一行の中に異なるフォントや半角・全角が混在するような場合には、ソフトウェア自体の改良が必要となると考えられる。

図 6-18　行内文字認識の誤りと分類

誤りモード	例	想定される対策
行の傾き		前処理 ソフトウェア改良
化学式		リジェクト(人力) ソフトウェア改良
空白		前処理 ソフトウェア改良
文字以外の映り込み		前処理 リジェクト(人力)

c）文脈補正の実施

文脈補正については、特許庁貸与データと JAPIO 所有データの双方を用いて実施した。

その精度については、特許庁貸与のデータを活用したものの方が若干高いという結果を得た。具体的には、特許庁貸与データを活用した文脈補正を実施したあとに 1 文字認識を行った際の認識率は、上述の通り 97.58%であり、JAPIO 所有データを活用した文脈補正実施後の 1 文字認識の認識率 97.49%を上回った。

さらに文脈補正を行ったことで不正解であったものが正解に転じた数、逆に正解が不正解に転じてしまった数を計測すると以下のようになる。

表 6-3　文脈補正結果比較

JAPIOが優
貸与物②が優

	区分	補正成功割合※ JAPIO	貸与物	貸与物とJAPIOの差	悪化割合? JAPIO	貸与物	貸与物とJAPIOの差
	A−①	39.04%	45.19%	6.15%	0.11%	0.12%	0.01%
	A−②	31.11%	38.94%	7.83%	0.30%	0.34%	0.03%
	A−③	0.00%	25.00%	25.00%	0.00%	0.00%	0.00%
	A−④	16.33%	24.49%	8.16%	0.00%	0.00%	0.00%
	A−a−①	17.06%	25.10%	8.04%	0.11%	0.11%	0.00%
	A−a−②	25.71%	31.43%	5.71%	0.70%	0.61%	−0.09%
	A−a−③	17.39%	26.09%	8.70%	0.12%	0.00%	−0.12%
	A−b−①	47.52%	46.51%	−1.01%	0.29%	0.59%	0.30%
'30年代〜'50年代	B−①	7.63%	8.95%	1.32%	0.83%	1.22%	0.39%
'30年代〜'50年代	B−②	7.32%	7.32%	0.00%	0.94%	1.52%	0.59%
'30年代〜'50年代	B−③	6.98%	20.93%	13.95%	1.82%	2.43%	0.61%
'30年代〜'50年代	B−④	18.35%	23.85%	5.50%	0.96%	1.54%	0.58%
	B−⑤	29.95%	33.58%	3.64%	0.23%	0.23%	0.00%
	B−⑥	5.56%	27.78%	22.22%	0.00%	0.00%	0.00%
	C−①	9.77%	13.28%	3.52%	0.58%	0.54%	−0.04%
	C−②	12.81%	29.34%	16.53%	0.10%	0.09%	−0.01%
	C−③	0.00%	9.09%	9.09%	0.00%	0.00%	0.00%
	C−④	3.85%	26.92%	23.08%	0.42%	0.42%	0.00%
'30年代〜'50年代	D−①	11.11%	18.52%	7.41%	1.50%	1.50%	0.00%
'30年代〜'50年代	D−②	9.62%	11.54%	1.92%	0.47%	1.11%	0.63%
'30年代〜'50年代	D−③	11.90%	15.48%	3.57%	3.46%	3.81%	0.35%
'30年代〜'50年代	D−④	12.08%	18.60%	6.52%	0.70%	1.26%	0.56%
	D−⑤	13.18%	24.34%	11.16%	0.11%	0.14%	0.02%
	D−⑥	12.50%	31.25%	18.75%	0.00%	0.00%	0.00%
	E−①	30.79%	41.01%	10.22%	0.12%	0.14%	0.02%
	E−②	29.73%	35.14%	5.41%	0.00%	0.17%	0.17%
	E−③	8.33%	16.67%	8.33%	0.00%	0.00%	0.00%
	E−④	20.93%	34.88%	13.95%	0.18%	0.00%	−0.18%
	H−a−①	28.33%	39.25%	10.92%	0.11%	0.10%	−0.01%
	H−a−②	20.81%	32.21%	11.41%	0.07%	0.00%	−0.07%
	H−a−③	13.16%	28.95%	15.79%	0.00%	0.11%	0.11%
	H−b−①	20.00%	32.00%	12.00%	0.00%	0.27%	0.27%
	総計	26.32%	33.02%	6.70%	0.31%	0.42%	0.11%

※補正成功数/認識失敗数
補正によって不正解⇒正解に転じた割合
認識失敗文字をどれだけ救えたか

?補正悪化数/認識成功数
補正によって正解⇒不正解に転じた割合
認識成功結果にどれだけ影響を与えたか

これを見ると、1970 年代以降のデータは文脈補正の成功率が高いのに対して、それ以前のものは補正によって悪化してしまうケースが無視できない数に至っていることがわかる。これは、文脈補正に用いたデータが 1970 年代以降のものであったことが影響している可能性が推測される。そこで、今後文脈補正を進めていくにあたっては、補正に利用するデータについて、年代・種別が実施対象データに近いものを集めた方が良いと考えられる。また、補正対象によって言語モデルを切替えるといった対応も検証に値すると考える。

3. 人工知能技術を活用した画像補正検証概要

OCR ソフトで文字を読み取る際、そのままでは読取りが難しい画像に関しては、画像補正処理が求められる。第 2 章で調査したフォーマット等の調査結果により、現在想定している特許文献のテキスト化に当たっては、古い画像等を中心に画質が低いものが多数存在していることが想定される。

そのため、画質の改善策を具体的に講じることが、OCR ソフトの文字認識精度向上に不可欠であり、これを AI 等の活用によって効率的に実施することが求められるため、本事業において特に検証を行うものとする。

補正が必要な画像の例として、画質が悪いという状況をさらに分類すると、画像が不鮮明であるもの、文字が欠損しているもの、文字の一部がつぶれてしまっているもの、画像にノイズが混入しているものなどが考えられる。

画像が不鮮明なものとは、低解像度の二値化により、文字のアウトラインが不鮮明になっているもののことである。欠損画像とは、文字の一部が消えてしまったり、密度の高い箇所がつぶれてしまっているもののことである。ノイズ画像とは、原稿のテクスチャが、ノイズとして画像内に含まれてしまっているもののことである。

図 6-19　不鮮明画像の例

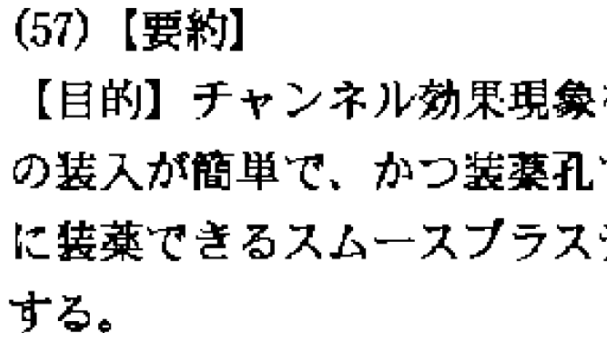

(57)【要約】
【目的】チャンネル効果現象
の装入が簡単で、かつ装薬孔
に装薬できるスムースプラス
する。

図 6-20　欠損画像の例

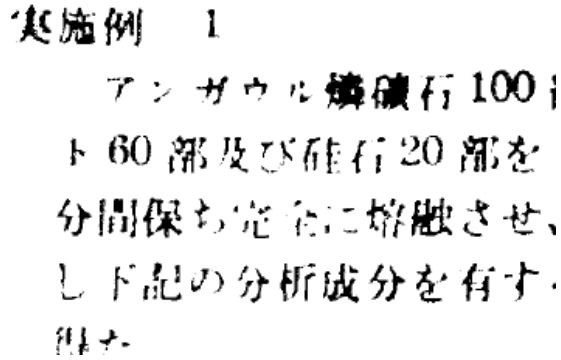
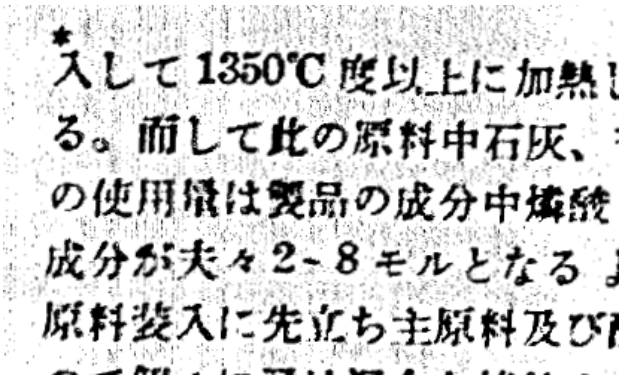

実施例　1
アンガウル燐鉱石 100
ト 60 部及び硅石 20 部を
分間保た定在に溶融させ、
し下記の分析成分を有す
得た

図 6-21　ノイズ画像の例



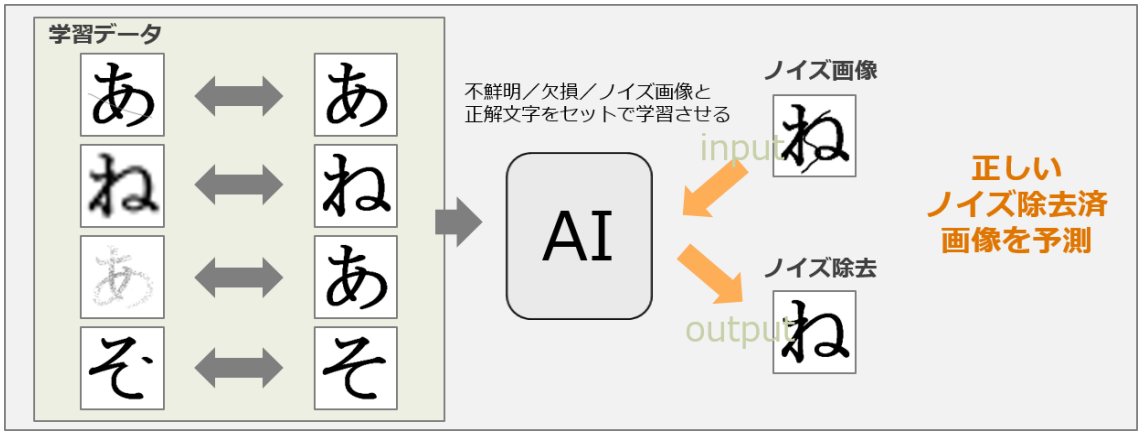
こうした要補正画像に対して、AI を活用した画像補正処理を行うことで、その後の OCR ソフトの文字認識精度の向上を検証した。

具体的には、人工知能に不鮮明・欠損・ノイズ画像と正解文字をセットで学習させることによって、要補正画像から正しい画像を予測させ、それに近づける処理を行った。

画像補正を実施するための AI の学習プロセスは、現時点において確立されたものではなく、期待する結果を出力できるように学習を収束させる方法について、試行錯誤が積み重ねられ、研究が進められている状況にある。

図 6-22　AI 活用補正

図 6-23　AI を活用した画像補正手法（ノイズ除去の場合）



AI 画像補正モデルの選定のあたり、AI を用いた画像補正の仕組みは、現在発展途上であるため、どのような機械学習モデルを用いれば最適な結果が得られるかは明確になってい

ない。代表的な機械学習モデルとしては、①CNN、②AutoEncoder、③GAN の 3 つが存在している。CNN（Convolutional Neural Network: 畳み込みニューラルネットワーク）[9]は、順伝播型のニューラルネットワークの一種であり、画像や動画認識に広く使われているモデルである。AutoEncoder[10]は、元の画像を復元可能なように次元を圧縮する、ニューラルネットワークを使用した次元圧縮のためのアルゴリズムである。GAN（Generative Adversarial Network: 敵対的生成ネットワーク）は生成ネットワーク（generator）と識別ネットワーク（discriminator）の 2 つのネットワークから構成され、2 つのネットワークが相反した目的のもとに学習するモデルである [11]。

---

[9] LeCun, Yann. “LeNet-5, convolutional neural networks”
http://yann.lecun.com/exdb/lenet/

[10] Geoffrey E. Hinton; R. R. Salakhutdinov (2006). “Reducing the Dimensionality of Data with Neural Networks”. Science 313 (5786): 504-507.
http://www.cs.toronto.edu/~hinton/science.pdf

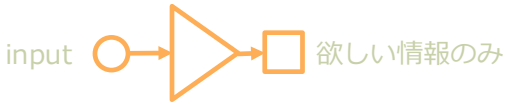
[11] Salimans, Tim; Goodfellow, Ian; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec; Chen, Xi (2016). “Improved Techniques for Training GANs”. https://arxiv.org/abs/1606.03498
Jianan Li; Xiaodan Liang; Yunchao Wei; Tingfa Xu; Jiashi Feng; Shuicheng Yan (2017). “Perceptual Generative Adversarial Networks for Small Object Detection”
https://arxiv.org/abs/1706.05274

＜代表的な機械学習モデル＞

① CNN（抽象化による特徴の明確化）

画像に応用すると、欲しい情報のみを取り出すことが可能。

図 6-24　CNN



※ネットワークの概念図

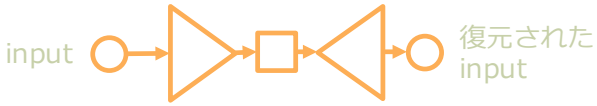
※〇がインプット、□がアウトプット、△がニューラルネットワークによる次元数の変化

② AutoEncoder（情報の圧縮と復元）

入力された情報から、元の情報を復元させることが可能

ノイズのある画像から、クリーニングされた状態を作り出す。

図 6-25　AutoEncoder



※ネットワークの概念図

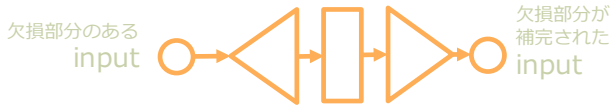
※〇がインプット、□がアウトプット、△がニューラルネットワークによる次元数の変化

③ GAN（正しい状態の予測生成）

正解データと対象データを読み込むことで、正解と同じようなものを生成することが可能

画像の欠損部分を予測し、本物の同等の状態に生成する。

図 6-26　GAN



※ネットワークの概念図
※○がインプット、□がアウトプット、△がニューラルネットワークによる次元数の変化

これらのうち、今回の検証では、GAN のモデルを用いた。今回は補正対象の画像が白黒の文書であり、カラー写真などと比べると補正の手掛かりとなる情報量が少ないため、写真等の画像の補正を得意とする CNN や AutEncorder はあまり向いていない。一方で、正解データが明確に存在していることから、対象画像を正解データに近づける作業をすることになり、この作業には GAN が適していると考えられたため、GAN を採用した。

GAN は、正式名称を「Generative　Adversal　Network」といい、生成器と選別器の二つの機構を持つ。生成器は選別器によって正解データと同じであると認識してもらえるような補正データ（すぐれた贋作）を生成することを目指し、選別器は正解データとの差分を発見して正解データと異なることを見極めることを目指す。この二つの GAN の中でも、対立する機構の切磋琢磨によって、補正後画像を正解画像に近づける学習が進んでいくという仕組みである。

さらに、今回の検証では、GAN の中から汎用的なモデルと高解像度化に適した二つのモデルを使用することにした。（以下、このモデルの 1 つをモデル A、もう一方をモデル B とする。汎用的なモデル ＝ モデル A、高解像度化に適したモデル ＝ モデル B）どちらのモデルも元画像（適当な画像）と正解（正しい、こうしたい）画像を 2 つの画像をペアで入力し、学習を行うことができる。

4. 画像補正の検証結果および考察

(1) 検証結果

人工知能を用いた画像補正の具体的な検証は、試行錯誤の段階であることから、仮説を立て、それを検証し、次の仮説につなげていくという仮説検証型アプローチによって進めた。

以下に、検討順に 1-1 から 7-3 までの合計 13 のアプローチについて仮説と検証結果を記載した。複数のアプローチで学習が収束せずうまくいかなかったものも存在したが、失敗したアプローチも次の仮説構築に繋がっていることから、それも含めて記載した。

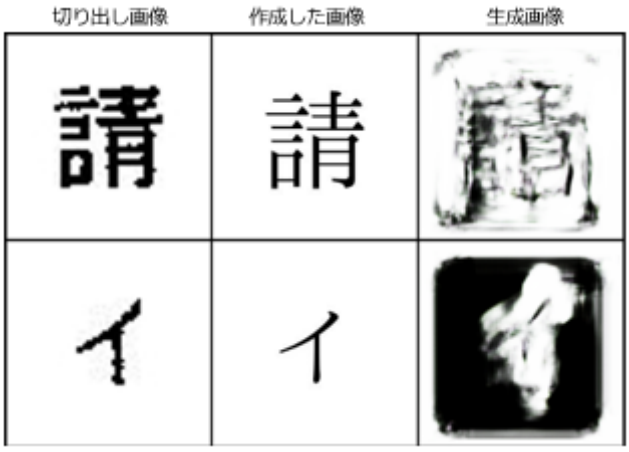
a）アプローチ 1-1（学習モデル：モデル B）

仮説：スキャン画像から切り出した 1 文字と明朝体フォントで作成した 1 文字画像をセッ

トで学習を実施することで、画像の補正と高解像度化ができるのではないか。

図 6-26　学習データ（アプローチ 1-1　モデル B）

図 6-27　学習結果（アプローチ 1-1　モデル B）



学習は収束せず失敗した。高解像度化もできず、スキャン画像から切り出した画像を使った学習ではうまく収束させられない可能性が高いことが判明した。

b）アプローチ 1-2（学習モデル：モデル A）

仮説：モデル A では、スキャン画像から切り出した 1 文字と明朝体フォントで作成した 1 文字画像をセットで学習を実施することで画像の補正と高解像度化ができるのではないか。

図 6-29　学習データ（アプローチ 1-2　モデル A）

図 6-30　学習結果（アプローチ 1-2　モデル A）

モデル A を用いても、やはり学習は収束せず失敗した。生成画像がフォント画像からかけ離れており、失敗に終わった。やはり学習データとしてスキャン画像から切り出した画像を使うと学習を収束させるのは相当難しいことが判明した。

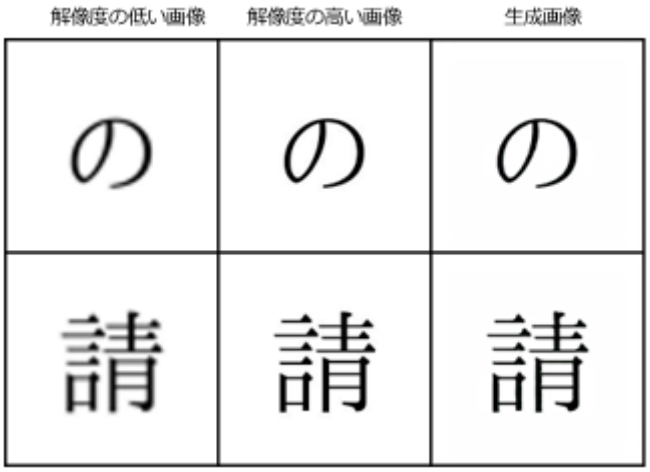
c）アプローチ 2（学習モデル：モデル B）

仮説：明朝体フォントで作成した解像度の低い画像と作成した解像度の高い画像をセットで学習を実施することで、高解像度化ができる

アプローチ 1 および 1-2 で学習が失敗したことから、スキャン画像ではなく、明朝体フォントから低解像度画像を作成し、それを用いて、そもそも文字の高解像度化ができるのかという実験を行った。

図 6-31　学習データ（アプローチ 2　モデル B）

図 6-32　学習結果（アプローチ 2　モデル B）



システム的に作成した低解像度画像という条件下ではあるが、モデル B による文字画像の高解像度化ができることが確認できた。ただし、同一フォント、同一サイズという特殊条件によるものなので、フォントのサイズや種類が違っても高解像度化ができるかについてはさらに検証が必要である（⇒アプローチ)。

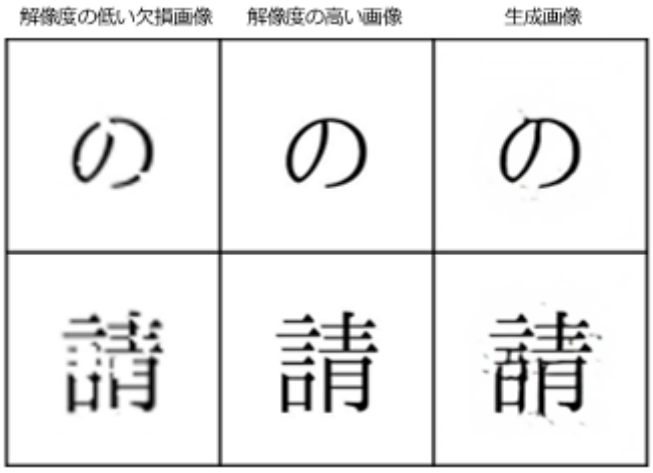
また、高解像度化のプロセスにおいて、欠損の補完やノイズ除去等も行えるかどうかについても検証が必要である。(⇒アプローチ)

d) アプローチ 3-1(学習モデル：モデル B)

仮説：欠損を加えた明朝体フォントで作成した解像度の低い画像と作成した解像度の高い画像をセットで学習を実施することで、欠損の補完と高解像度化が同時にできるのではないか。

図 6-33　学習データ（アプローチ 3-1　モデル B）

図 6-34　学習結果（アプローチ 3-1　モデル B）



アプローチ 2 と同様に高解像度化はできている。さらに、一部欠損の補完ができているが、一部不正確に補完してしまっている部分も存在している。今後適切な改善を図ることができれば実用に堪えるものとなると考えられる。

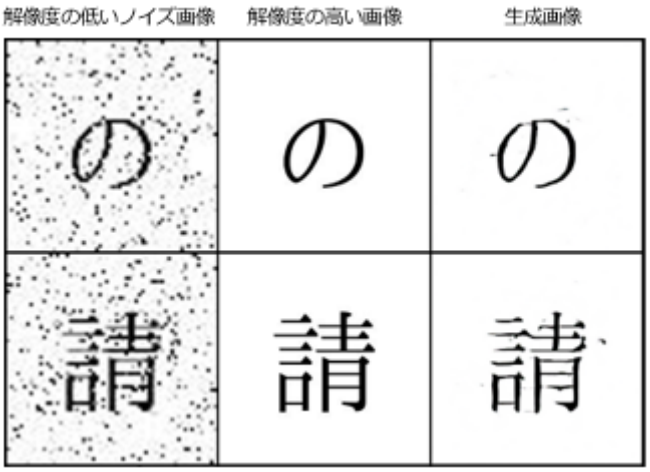
フォントのサイズ・種類違いについては適用可能か検証が必要である。

e）アプローチ 3-2（学習モデル：モデル B）

仮説：ノイズを加えた明朝体フォントで作成した解像度の低い画像と作成した解像度の高い画像をセットで学習を実施することで、ノイズ除去と高解像度化が同時にできるのではないか。

図 6-35　学習データ（アプローチ 3-2　モデル B）

図 6-36　学習結果（アプローチ 3-2　モデル B）



一部を除き、ノイズ除去ができており、高解像度化もできている。モデル B が高解像度化専用ではなく、ノイズ除去にも有用であることが示された。

フォントのサイズ・種類違いについては適用可能か検証が必要である。

f) アプローチ 4-1（学習モデル：モデル A）

仮説：モデル A においても、ノイズを加えた明朝体フォントで作成した解像度の低い画像と作成した解像度の高い画像をセットで学習を実施することで、ノイズ除去と高解像度化が同時にできるのではないか。

図 6-37　学習データ（アプローチ 4-1　モデル B）

図 6-38　学習結果（アプローチ 4-1　モデル B）

画像からのノイズ除去はできている。ノイズ除去の精度は、モデル B よりも高い可能性がある。その一方で、高解像度化はできておらず、モデル B がノイズ除去と高解像度化の両方を同時にできていたのとは異なる。

フォントのサイズ・種類違いについては適用可能か検証が必要となる。

g）アプローチ 4-2（学習モデル：モデル A）

仮説：ノイズを加えたゴシック体で作成した解像度の低い画像と作成した解像度の高い画像をセットで学習を実施することで、明朝体と同様にノイズを除去できるのではないか。

図 6-39　学習データ（アプローチ 4-2　モデル A）

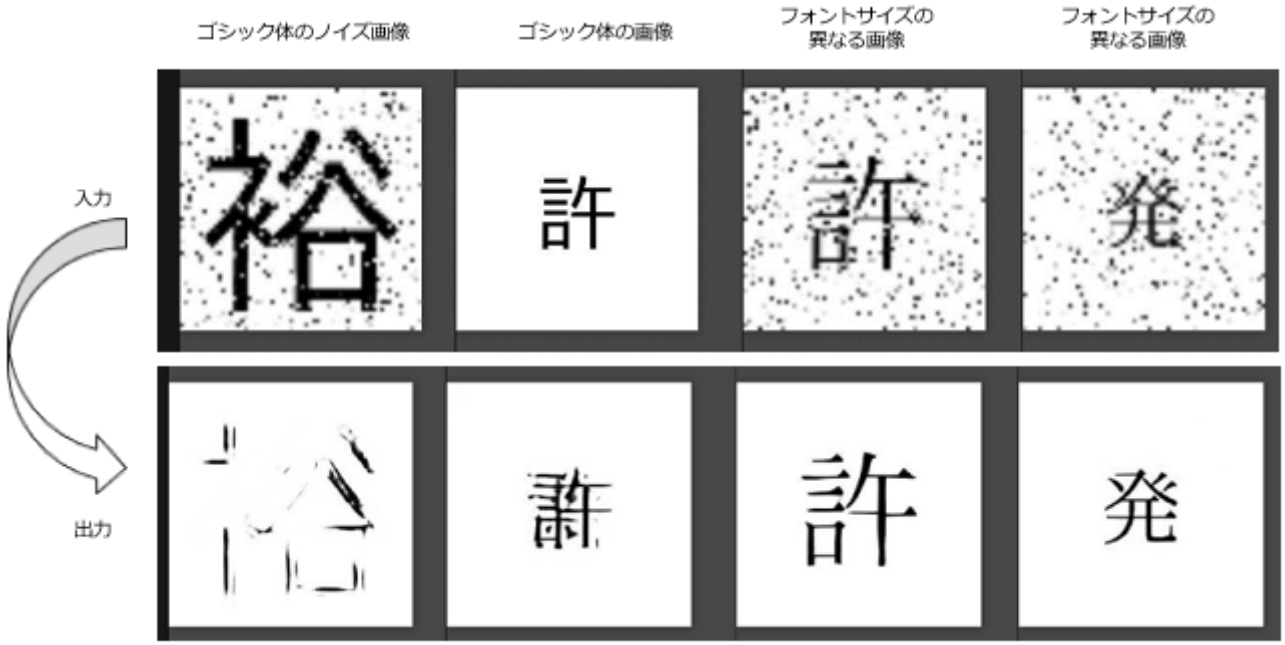
図 6-40　学習結果（アプローチ 4-2　モデル A）

ゴシック体を学習した場合でも、画像からのノイズ除去ができている。これにより、モデル A は、明朝体に限らず、適切なフォントを個別に学習すれば、ノイズ除去を実施できるものと考えられる。

h）アプローチ 5-1（学習モデル：モデル B）

仮説：ノイズを加えた明朝体フォントで作成した解像度の低い画像と作成した解像度の高い画像をセットで学習したモデルをゴシック体やフォントサイズの異なる画像にも適用できるのではないか

図 6-41　検証結果（アプローチ 5-1　モデル B）



同じ明朝体で、フォントサイズが異なる場合であっても、ノイズの除去を行うことはできる。一方、ゴシック体のノイズ画像は文字の要素までノイズと誤認して除去してしまう形になり、ゴシック体には対応できなかった。

学習データはフォントごとに準備し、フォント別に学習させる必要があると考えられる。

i）アプローチ 5-2（学習モデル：モデル B）

仮説：ノイズを加えたゴシックフォントで作成した解像度の低い画像と解像度の高い画像をセットで学習したモデルを明朝体やフォントサイズの異なる画像に適用できるのではないか。

図 6-42　検証結果（アプローチ 5-2　モデル B）

明朝体で学習したときと同じく、同じフォント体には適用可能だが他のフォントには適用が難しい。また、ゴシック体を学習させた場合には、学習時のフォントサイズへは対応できるが、サイズが異なると対応できていない。

j）アプローチ 6（学習モデル：モデル B）

仮説：複数フォント（明朝体、ゴシック体）で作成した解像度の低い画像と作成した解像度の高い画像をセットで学習を実施することで、どちらのフォントについても高解像度化ができる。

図 6-43　学習データ（アプローチ 6　モデル B）

図 6-44　学習結果（アプローチ 6　モデル B）

複数フォントを同時に学習した場合でも、学習結果の混乱等は生じず、双方のフォントについて高解像度化が実現できた。これにより一つの AI モデルに、複数のフォント画像を学習させることで、フォントの違いにかかわらず画像の高解像度化につなげられることが確認できた。

k）アプローチ 7-1（学習モデル：モデル B）

仮説：複数文字を含む学習データを利用することで、高解像度化の精度が上がるのではないか。

図 6-45　学習データ（アプローチ 7-1　モデル B）

この学習データは、元のデータに表れている文字をランダムで並べたものである。

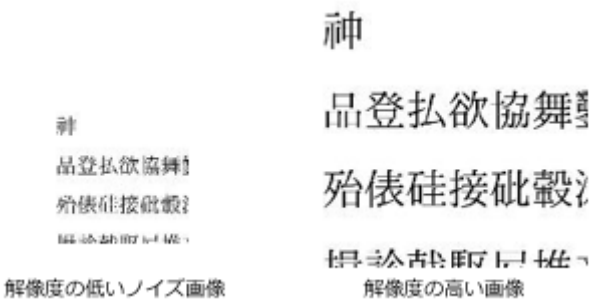
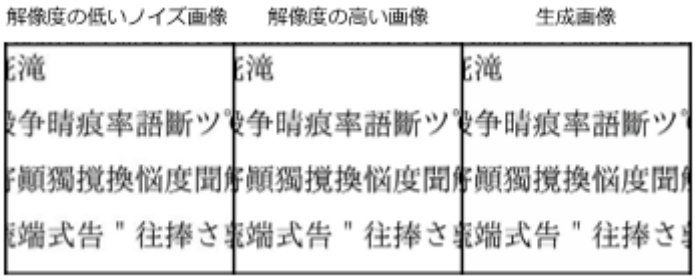


図 6-46　学習結果（アプローチ 7-1　モデル B）



テストデータでは高解像度化はできている。また、複数文字を含む学習データを利用することで、学習データを効果的に増加させることができ、高解像度化の精度は向上すると考えられる。

l) アプローチ 7-2 （学習モデル：モデル B)

仮説：複数文字を含む画像のデータを利用することで、欠損補完の精度も上がるのではないか。

図 6-47　学習データ（アプローチ 7-2　モデル B）

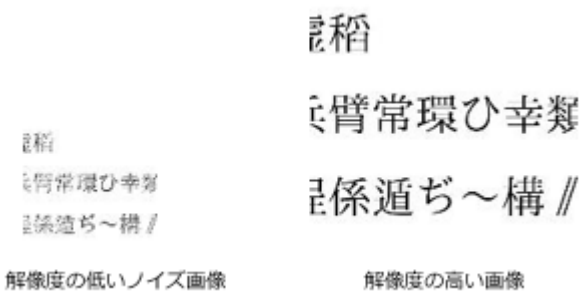
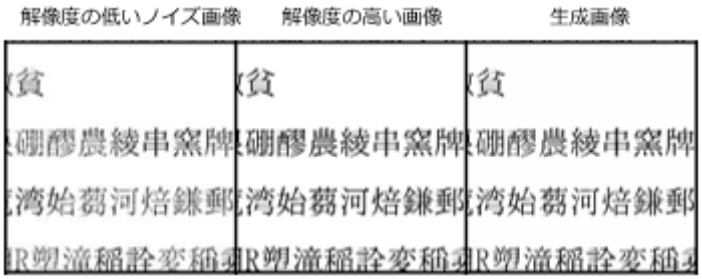


図 6-48　学習結果（モデル B）



テストデータでは欠損補完ができている部分もあった。しかし、欠損補完については不正確な部分もあり、完全とは言えなかった。

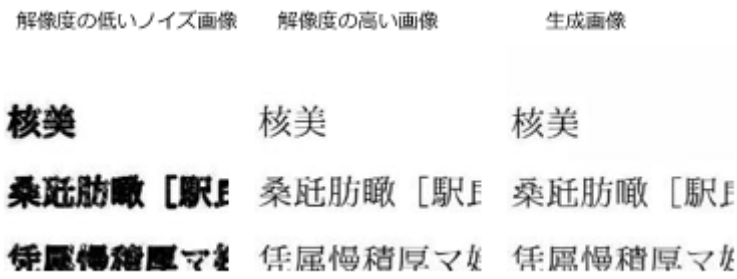
ただし、少なくとも、複数文字が並んでいる場合でも処理はできており、複数文字を含む画像を学習データとして活用することは、効率性の観点から有用であることがわかった。

m）アプローチ 7-3 （学習モデル：モデル A、モデル A_001 p3.8）

仮説：複数文字を含む画像のデータを利用することで、文字の鮮明化（滲みの対応）精度も上がるのではないか。

図 6-49　学習データ（アプローチ 7-3　モデル B）

図 6-50　学習結果（アプローチ 7-3　モデル B）



テストデータでは一部鮮明化ができているものの、補正前の画像から文字が潰れている部分は鮮明化ができていない。軽い滲み程度であれば効果があがると思われるが、文字がつぶれるレベルの滲みは補正できない。

(2) 仮説アプローチで検証された結果

ここまでの仮説アプローチの結果、以下のような示唆が得られた。

① モデル B でもモデル A でも、スキャン画像から切り出した文字を用いて学習した場合、AI がスキャン画像とフォント画像との同一性を十分に認識できないと思われ、結果として学習は収束せず失敗した。

② モデル B において、フォント画像を加工して低解像度の画像を作り、それと高解像度の画像を一緒に学習させた場合、低解像度の画像の高解像度化が実現できる。

③ モデル B において、欠損もある低解像度の画像を作り、それと高解像度の画像を一緒に学習させた場合、高解像度化と同時に一部の欠損補完が実現できるが、その効果は不十

分である。

④ モデル B において、ノイズもある低解像度の画像を作り、それと高解像度の画像を一緒に学習させた場合、高解像度化と同時に一定のノイズ除去が実現できる。

⑤ モデル A において、ノイズもある低解像度の画像を作り、それと高解像度の画像を一緒に学習させた場合、高解像度化はできないが、モデル B と同等以上のノイズ除去が実現できる。これは、明朝体を学習させた場合でもゴシック体を学習させた場合でも同様である。

⑥ 学習はフォントごとに実施しなければならず、異なるフォントの画像についてノイズ除去等の補正を行うことはできない。また、ゴシック体を学習させた場合には、フォントサイズが異なる画像にも十分に対応できない。

⑦ 学習データとして複数のフォントを同時に学習させても、特に AI の混乱は生じず、学習したフォント全体について高解像度化等を実現できる。

⑧ モデル B において、学習データとして複数文字を含む画像データを利用することで、より効率的に学習データ数を増加させることができ、高解像度化の精度は向上できた。一方、欠損、滲み等については、十分な精度向上までは確認できなかった。

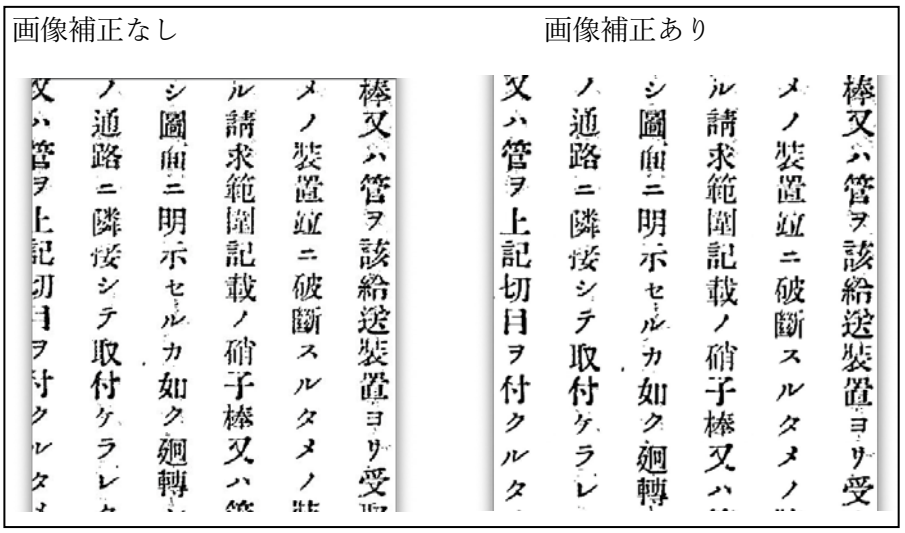
以上をまとめると、モデル B とモデル A のモデルを適切に活用して、フォントごとに AI の学習を行うことにより、OCR 認識精度の向上のために必要と考える画像補正のうち、画像の高解像度化と、ノイズの除去については、優位な精度向上につながる成果が生じるものと考えられる。また、より効率的な学習のためには、複数文字を含む画像データの利用が有力な手段であると言える。

ここまでの示唆を受けて、実データに対して画像補正を適用した結果を述べる。

a）高解像度化

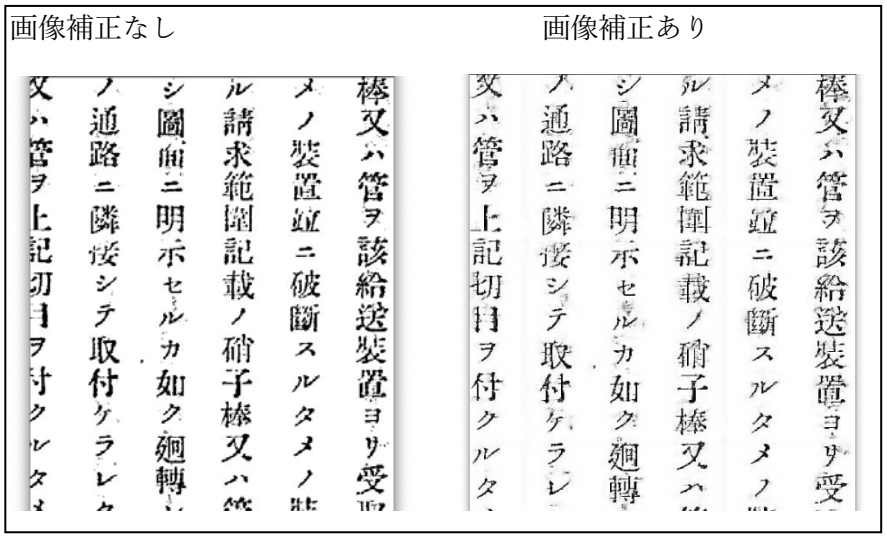
高解像度化は一定の補正の効果が定性的に確認できた。

図 6-51　高解像度化結果サンプル



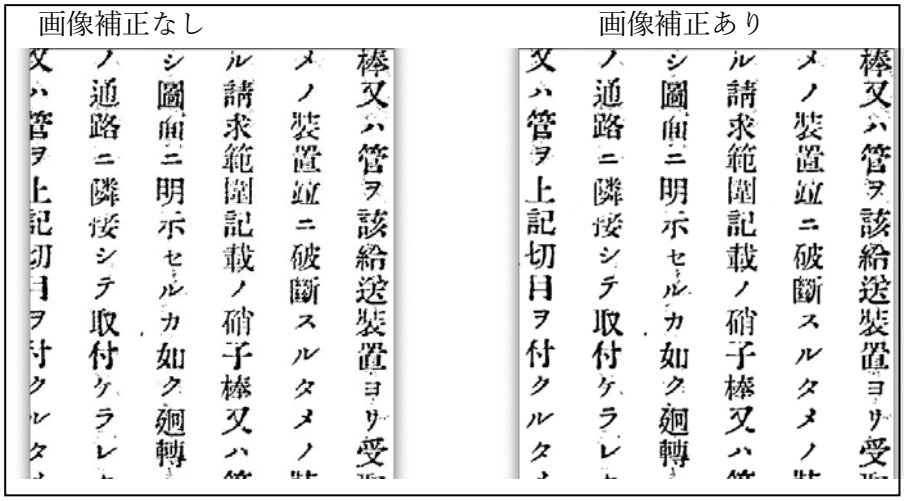
b）鮮明化（滲みへの対応）は十分な補正の成果が確認できなかった

図 6-52　鮮明化結果サンプル



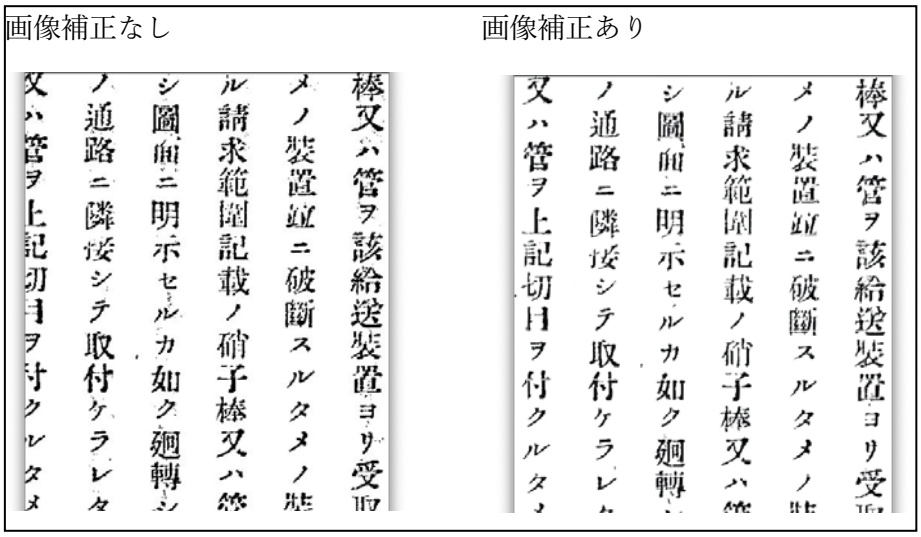
c）欠損補完は十分な補正の成果が確認できなかった

図 6-53　欠損補完結果サンプル



d）ノイズ除去は一定の効果が定性的に確認できた。

図 6-54　ノイズ除去結果サンプル



以上の結果は、(3)の仮説アプローチから得られたモデル上の示唆と同様である。やはり、高解像度化とノイズ除去の効果は認められるが、欠損補完や滲み対応は十分には行えなかったという結果である。

図 6-55　精度向上結果サンプル

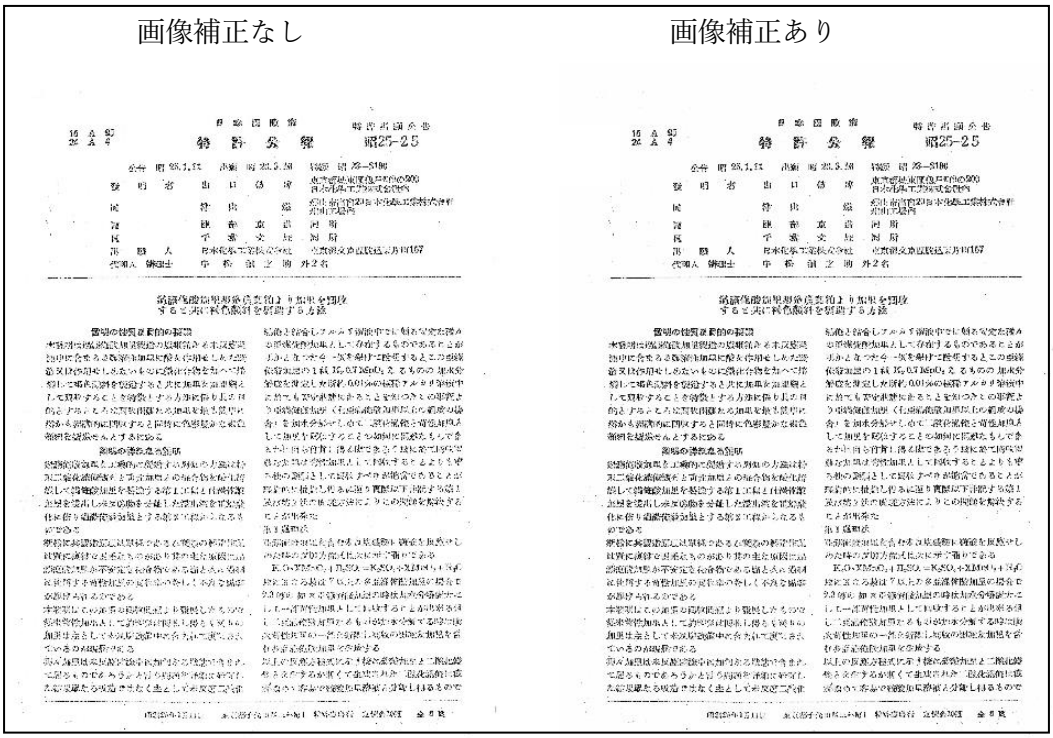
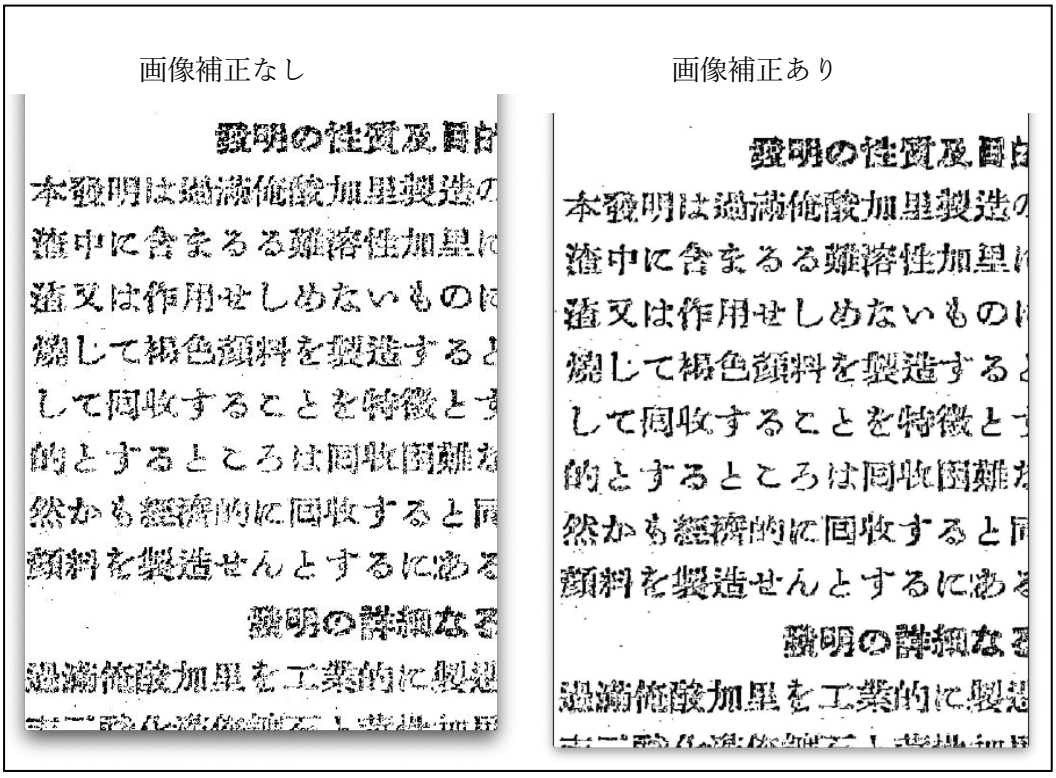


図 6-56　精度向上結果サンプル



定性的な評価に加え、画像補正を実施したデータに対して、市販 OCR ソフトをかけテキスト化を行った際の認識精度の変化を算定した。

これによると、画質が悪く、補正前の認識精度が低いデータについては精度向上が確認できた。その一方で、画質が良く、補正前の認識精度が高いものに対しては、改善は見られない、もしくは精度が低下した。

緑色：認識精度向上
黄色：認識精度低下

表 6-4　補正有無結果
AI 画像補正処理あり

年代	種別	TESSERACT	PANASONIC	NJK
30	全	46.65%	68.22%	75.34%
	B	46.65%	68.22%	75.34%
50	全	51.42%	78.71%	61.77%
	B	**43.52%**	**73.55%**	**49.44%**
	D	65.51%	87.91%	83.75%
70	全	81.94%	90.63%	95.68%
	A	72.48%	84.08%	93.12%
	B	85.81%	92.31%	96.83%
	C	80.35%	91.42%	95.23%
	D	**89.78%**	**96.60%**	**97.64%**
80	全	72.12%	83.90%	91.26%
	A	65.42%	73.51%	87.26%
	A-a	79.44%	89.44%	95.02%
	A-b	62.77%	84.05%	86.05%
	B	84.28%	92.44%	97.09%
	C	84.49%	91.56%	95.44%
	D	86.74%	94.12%	97.19%
	E	58.62%	76.34%	86.84%
	H	75.54%	88.91%	95.36%

AI 画像補正処理なし

年代	種別	TESSERACT	PANASONIC	NJK
30	全	40.21%	67.57%	70.30%
	B	40.21%	67.57%	70.30%
50	全	38.64%	72.02%	40.63%
	B	**27.29%**	**63.95%**	**22.22%**
	D	58.88%	86.39%	73.45%
70	全	82.14%	90.91%	95.68%
	A	72.44%	84.70%	93.21%
	B	86.41%	92.71%	96.99%
	C	80.40%	91.03%	94.67%
	D	**89.75%**	**96.56%**	97.56%
80	全	73.51%	85.93%	91.75%
	A	67.34%	76.45%	88.06%
	A-a	80.86%	91.41%	95.58%
	A-b	64.09%	86.41%	86.32%
	B	85.56%	93.70%	97.13%
	C	85.33%	92.60%	95.06%
	D	87.44%	95.10%	**97.66%**
	E	59.49%	77.45%	87.45%
	H	77.09%	91.33%	96.02%

(3) 認識精度向上策

今回の検証によって、画像の高解像度化とノイズ除去には、人工知能を用いた画像補正が有用であり、この画像補正を行うことで、現時点において画質が悪く認識精度が低いデータについては認識精度の向上が見込まれることが判明した。

人工知能を用いた画像補正の手法を今後の OCR ソフトの文字認識精度の向上に活かすためには、第一に、画像補正をかけるべき画像の識別を確実に行うことが重要であり、具体的には、特に現状の OCR ソフトの文字認識精度が低いデータを抽出し、これらに対して画像補正前後の認識精度の変化と差分をとったのが以下の 3 種のグラフである。

使用した既存 OCR ソフト（レイアウト解析で使用したものと同じ）

D：ABBYY　「FineReader15 Standard」

B：PANASONIC ソリューションテクノロジー「読取革命 Ver.15」
C：NTT データ NJK 「e.Typist v15.0」

横軸を画像補正前の文字認識精度、縦軸を画像補正後の文字認識精度として、検証用データ 222 件をプロットした。赤枠内は、画像補正により文字認識精度が向上した範囲である。

図 6-57 OCR ソフト D 結果

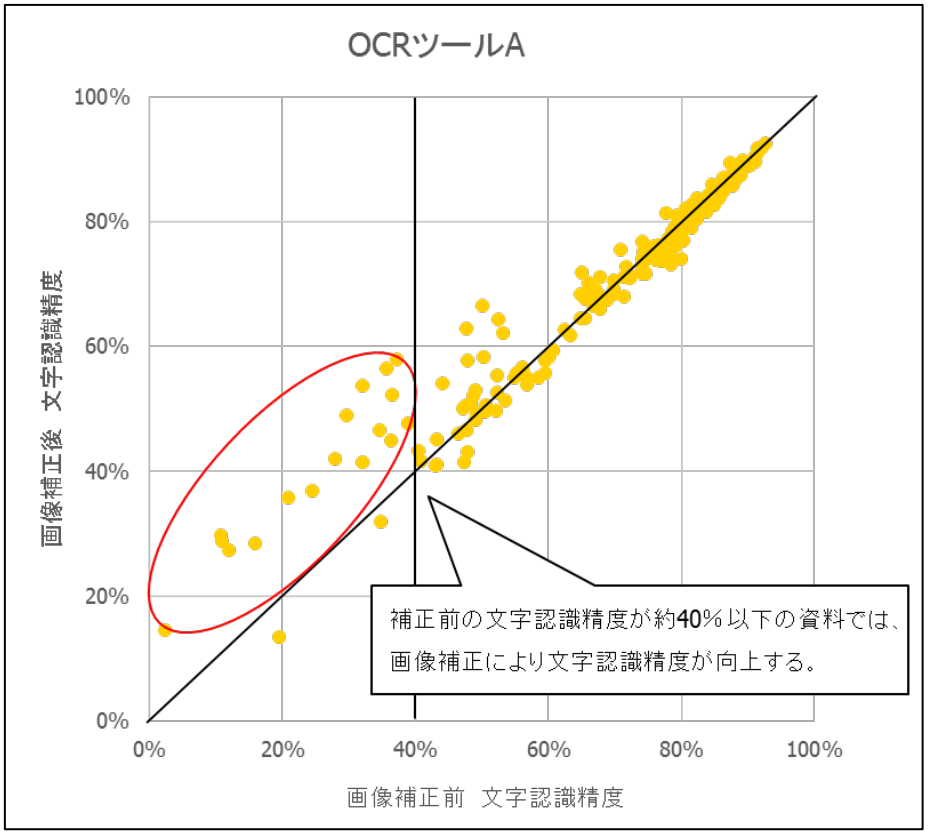


図 6-58　OCR ソフト B 結果

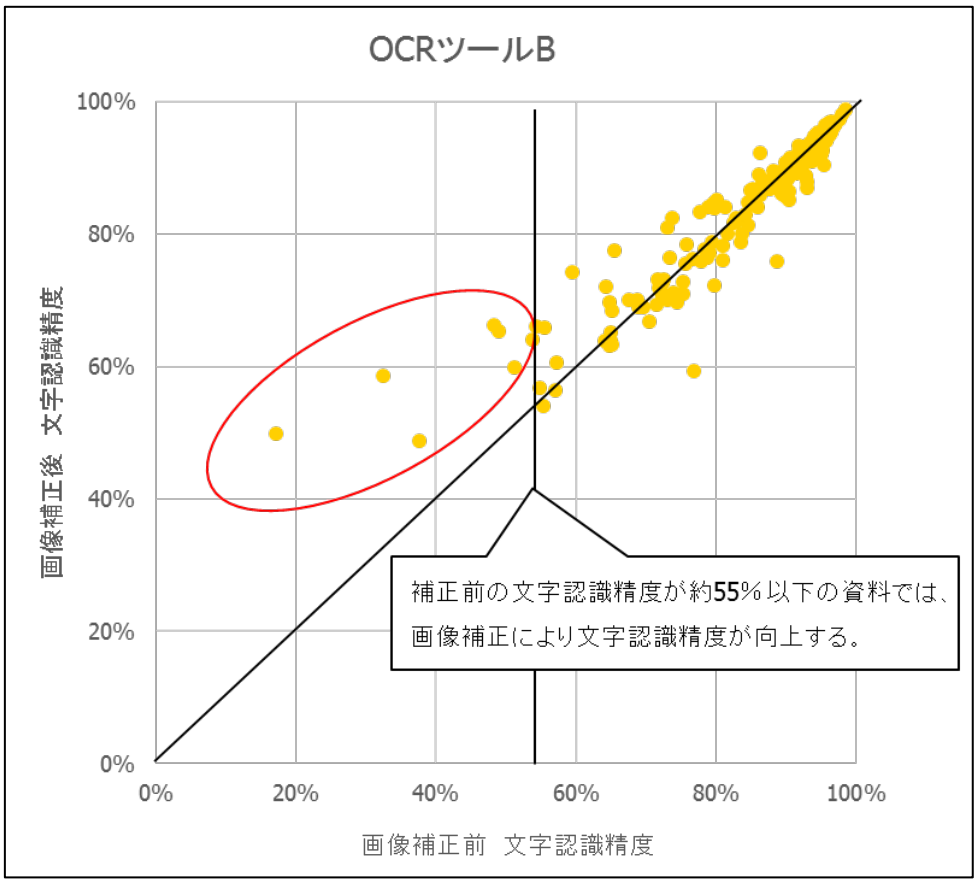
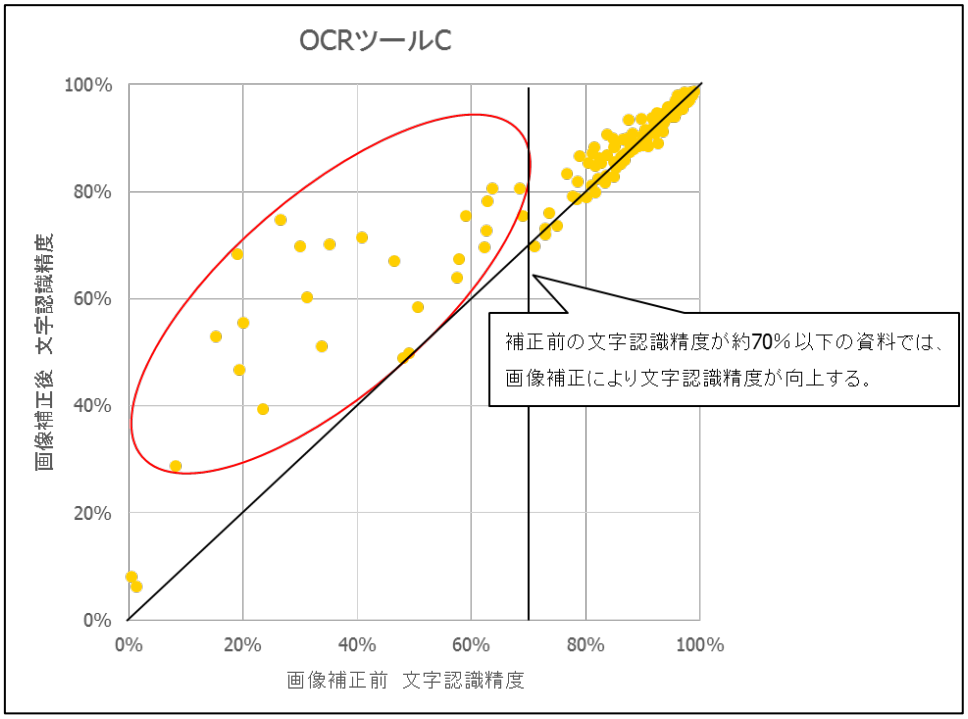


図 6-59　OCR ソフト C 結果



これを見ると、OCR ソフトによって差が発生しているが、現状の認識精度が、それぞれの、OCR ソフト D では 40%以下、OCR ソフト B では 55%以下、OCR ソフト C では 70%以下の低認識率の範囲に含まれている画像データについては、画像補正による認識精度向上が顕著に見て取れる。そのため、現時点の画像補正モデルのままであっても、こうしたデータについては、優先的に画像補正をかける効果が高いと言える。

また、モデル A、モデル B では、高解像度化とノイズ除去は効果的に実施できたものの、欠損補完や鮮明化は十分に行えなかったため、今後、この二つ以外の画像補正モデルの検討や、独自のモデル構築についても検討する必要がある。

加えて、現状の 2 モデルを利用する場合であっても、実データに近い欠損、不鮮明を再現した学習データを用意し、データ量を増やし学習することで欠損補完、鮮明化の精度向上を図ることも考えられる。

特許関連文献は、年代、種別によって、フォーマットが異なっていることが今回の調査で判明したため、画像補正の学習データについても、年代別・フォーマット別に、その画像イメージに近いデータで学習させれば、精度向上の可能性は高いと考えられる。

5. 小括

AI-OCR が、既存 OCR ソフトよりも高い文字認識精度でテキスト化を実現できる可能性が高いことが判明した。今後さらに、適切な文脈補正や教師データの増強、既存 OCR ソフトとの組み合わせ処理等を活用することで、さらなる認識精度の向上が期待できる。

また、人工知能を活用した画像補正についても、特に高解像度化とノイズ除去は優位な改善が生じる公算が高いことが判明した。特に、現時点で画質が悪く、認識率が低いデータについては、画像補正による認識精度向上の効果が大きいことが想定されるため、認識率が 70%以下のデータから優先的に画像補正を実施することが効果的である。

第 7 章　総合分析

第 3 章で調査した公報種別・フォーマット区分ごとに、既存 OCR ソフトおよび AI-OCR を単体または組み合わせることで、全件のテキスト化を実施した場合のテキスト化の精度、および必要な費用・リソースに関して調査する。

また、テキストデータ作成後に、特許庁の公報テキストデータフォーマットに沿うデータ形式に変換するための変換ツールや処理システムの開発等、既存 OCR ソフト以外に全件テキスト化実施時に必要となるツール等についても調査を行い、それらの費用・リソースに関して報告する。

1. 既存 OCR ソフトと AI-OCR の比較および組み合わせの検討

第 4 章および第 6 章の結果より、既存 OCR ソフトの上位 2 種（B、C）および AI-OCR の文字認識の結果を統合した結果を下表に示す。AI-OCR 単体の場合の精度に加えて、AI-OCR と既存 OCR ソフト上位 2 種との多数決処理を行った場合の精度に関しても集計結果を下表内に示す。

なお、以降の分析の見通しをよくするため、32 区分について開始日順で並び替えを行い、さらに文字認識精度への影響が大きい要素（縦組、旧字旧仮名遣い、画質）を表内に示した。

画質の項目に関しては、第 3 章の調査結果より文字のつぶれ・かすれが 50%以上の頻度で出現する区分について✔で評価している。

表 7-1　既存 OCR ソフト上位 2 種および AI-OCR の文字認識の結果

区分 No	開始日	縦組	旧字 旧仮名	画質が 悪い	OCR ソフト		AI-OCR	多数決 (B+C+AI)
					B	C		
D-①	1922/06/01～	○	○		59.5%	83.8%	90.9%	89.8%
B-①	1922/06/09～	○	○		67.6%	70.3%	92.9%	84.4%
B-②	1938/07/29～	○	○		54.6%	83.7%	92.4%	88.1%
D-②	1938/07/30～	○	○		45.1%	79.3%	91.9%	86.4%
B-③	1947/12/26～		○	✔	71.7%	40.8%	80.3%	79.6%
D-③	1947/12/27～		○	✔	37.5%	49.1%	53.4%	76.1%★
B-④	1949/12/23～		○	✔	64.0%	22.2%	88.8%	73.1%
D-④	1949/12/23～		○	✔	86.4%	73.5%	92.6%	92.1%
B-⑤	1963/12/28～				93.1%	97.0%	95.3%	98.1%★
D-⑤	1963/12/28～				95.9%	97.6%	96.7%	98.1%
A-①	1971/07/16～			✔	84.7%	93.2%	97.8%	96.0%
C-①	1971/09/13～				91.0%	94.7%	94.6%	96.6%★
E-①	1971/09/13～			✔	77.4%	87.4%	93.5%	92.2%
C-②	1977/06/30～				92.6%	95.1%	92.5%	95.7%

区分 No	開始日	縦組	旧字 旧仮名	画質が 悪い	OCR ソフト		AI-OCR	多数決 (B+C+AI)
					B	C		
A-②	1977/07/01～			✔	76.5%	88.1%	90.6%	91.0%
A-a-①	1979/07/26～				91.4%	95.6%	95.3%	95.4%
A-b-①	1979/08/09～			✔	86.4%	86.3%	90.4%	92.6%★
H-a-①	1979/09/06～			✔	91.6%	96.3%	95.8%	96.3%
H-b-①	1982/03/11～				86.3%	94.1%	85.7%	82.5%
E-②	1983/04/01～			✔	88.8%	90.1%	86.5%	93.9%★
H-a-②	1984/12/20～			✔	90.8%	95.5%	95.8%	96.6%
B-⑥	1984/12/27～				94.3%	94.8%	96.5%	97.4%
D-⑥	1984/12/27～				94.8%	94.2%	95.9%	96.9%
A-a-②	1985/01/10～				93.8%	90.4%	95.3%	93.7%
C-③	1985/02/13～				93.6%	93.9%	95.6%	97.2%★
A-③	1985/02/14～				94.1%	90.6%	94.1%	96.2%★
E-③	1985/02/14～				95.3%	86.9%	89.4%	97.0%★
C-④	1992/07/28～				95.5%	92.2%	87.8%	95.9%
E-④	1992/07/29～			✔	92.6%	89.9%	74.5%	94.6%★
A-④	1992/07/30～				94.5%	91.8%	86.1%	95.1%
H-a-③	1993/09/02～				84.6%	98.0%	94.9%	95.9%
A-a-③	1994/01/06～				96.7%	87.1%	96.0%	97.4%
				合計	90.5%	91.4%	93.6%	94.2%

89.7% OCRツール2種よりもAI-OCRの精度がよい区分

95.3% ★ 多数決の結果、1ポイント以上精度が向上した区分

上記集計結果より、以下のような傾向が確認される。

[既存 OCR ソフトと AI-OCR の比較]

- 今回の検証用データ全体での文字認識精度は、
  OCR ソフト B (90.5%) < OCR ソフト C (91.4%) < AI-OCR (93.6%)
  であり、32 区分中 18 区分で AI-OCR の文字認識精度が最も高かった。
- 既存 OCR ソフトは、20～60%程度と極端に文字認識精度が悪くなる区分が散在するのに対し、AI-OCR はほぼすべての区分において安定して高精度を実現している。
- 特に旧字旧仮名遣いの古い年代（1920～1940 年代）の資料、および画質が悪い区分の資料において、AI-OCR と既存 OCR ソフトの認識精度の差が大きい。

[既存 OCR ソフトと AI-OCR の多数決処理]

- 既存 OCR ソフト上位 2 種と AI-OCR の多数決処理により、認識結果が単一の OCR の場合よりも 1 ポイント以上改善された区分は 32 区分中 9 区分であった。認識精度向上の効果は限定的であると判断される。

2. 公報種別・発行年代別の最適な OCR ソフトの検討

前節の結果より「旧字旧仮名遣いの資料」「画質が悪い資料」「画質が良い資料」の 3 つの特徴が、最適な手法を決めるにあたって重要な指標になることが確認されたため、この 3 区分に従い既存 OCR ソフトおよび AI-OCR の精度をまとめると以下の表が得られる。

表 7-2　3 区分ごとの既存 OCR ソフトおよび AI-OCR の精度

レイアウトの特徴	検証データ数	区分 No	OCR ソフト B	OCR ソフト C	AI-OCR
[I] 旧字旧仮名遣い	50	B-①② D-①②	平均：62.1% 45.1%～67.6%	平均：74.2% 70.3%～83.8%	平均：92.6% 90.9%～92.9%
		B-③, ④-(1～2) D-③, ④-(1)	平均：74.1% 37.5%～86.3%	平均：48.5% 22.2%～73.5%	平均：89.1% 80.3%～92.6%
[II] 画質が悪い資料 （新字新仮名遣い）	68	A-①② A-b-① B-④-(3) D-④-(2) E-①②④ H-a-①②	平均：84.6% 64.0%～92.7%	平均：92.1% 22.2%～96.3%	平均：94.2% 74.5%～97.8%
[III] 画質が良い資料 （新字新仮名遣い）	104	A-③④ A-a-①②③ B-⑤⑥ C-①②③④ D-⑤⑥ E-③ H-a-③ H-b-①	平均：92.4% 84.6%～96.7%	平均：95.0% 86.9%～98.0%	平均：94.0% 85.7%～96.7%

3 区分それぞれの特徴を以下に記す。

[I] 旧字旧仮名遣いの資料

AI-OCR の文字認識精度が最も高い。
既存 OCR ソフトの中では、より古い資料では OCR ソフト C、比較的新しい資料では OCR ソフト B の精度が良いが、平均精度・バラつきの大きさともに AI-OCR と比較して大幅に劣る。

[II] 画質が悪い資料（新字新仮名遣い）
AI-OCR の文字認識精度が最も高い
既存 OCR ソフトの中では OCR ソフト C の精度が良いが、バラつきの大きさで AI-OCR に劣る。

[III] 画質が良い資料（新字新仮名遣い）
既存 OCR ソフト C の文字認識精度が最も高い。

したがって、文字認識に関する最適ツールに関しては以下のように結論付けられる。

表 7-3　3 区分に対する最適ツール

レイアウトの特徴	区分 No	推定 総件数	最適ツール	（参考）低コスト 優先の場合
[I] 旧字旧仮名遣い	B-①② D-①②	46 万件	AI-OCR （目標精度： 93%程度 ※1）	OCR ソフト B と C の組み合わせ （想定精度： 22%～86% ※2）
	B-③, ④-(1～2) D-③, ④-(1)	5 万件		
[II] 画質が悪い資料 （新字新仮名遣い）	A-①② A-b-① B-④-(3) D-④-(2) E-①②④ H-a-①②	507 万件	AI-OCR （目標精度： 95%程度 ※1）	OCR ソフト C （想定精度： 22%～96% ※2）
[III] 画質が良い資料 （新字新仮名遣い）	A-③④ A-a-①②③ B-⑤⑥ C-①②③④ D-⑤⑥ E-③ H-a-③ H-b-①	1,072 万件	OCR ソフト C （想定精度： 87%～98% ※1）	左同

※1) AI-OCR の精度に関しては、今後のチューニングによりさらに数ポイントの改善が可能と判断されるため「目標精度」として記載した。

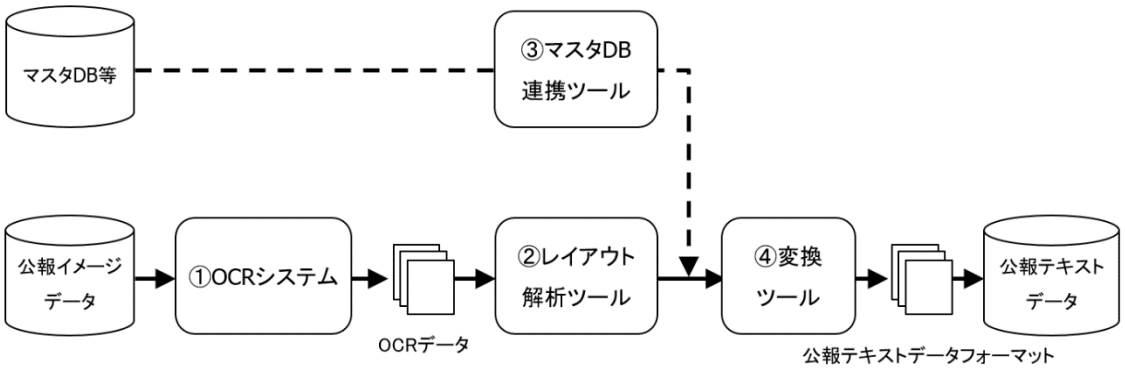
※2）画質の悪い資料に関しては、AI 画像補正により数ポイントの精度改善が可能である。

3. 全文テキスト化の必要費用とリソース

(1) システムの全体構成図

公報イメージデータを公報テキストデータへ変換する際に必要とされるシステムの全体構成図（想定）を以下に示す。

図 7-1　テキスト化システム全体構成図（想定）



① OCR システム：

イメージデータを OCR データ（座標情報付のテキストデータ）へ変換するバッチ処理システム。画像処理等の前処理、複数 OCR ソフトの連携処理等を含む。

② レイアウト解析ツール：

OCR データから文書構造情報を解析し書誌情報等を抽出する。

③ マスタ DB 連携ツール：

イメージデータ以外のマスタ情報が利用可能な場合、外部マスタから取り出された書誌情報を適切な形式へ変換し、イメージデータへの紐づけを行う。

④ 変換ツール：

OCR データおよび②または③から得られた書誌情報から、公報テキストデータフォーマットに沿うデータ形式に変換する。

(2) 想定コスト試算

a) OCR システム

システム条件（共通）

- 約 1,500 万件（約 1 億画像ファイル）を既存 OCR ソフトまたは AI-OCR を用いて、人手を介することなく、全自動でテキスト化する。
- IMG ファイルの解凍、画像処理、等の前処理を実施する。

・ クラウドデータセンターを利用し、セキュリティ要件は通常サービス程度を想定する。
・ 事業期間は、既存 OCR ソフト 1 種の場合は 10 ヶ月間、AI-OCR の場合は、18 ヶ月間を想定。
・ テキスト精度に関する品質保証はなし。
・ 目視によるテキストの修正（校正）は行わない。
・ 0.1%程度の抜き取り検査を実施し、システムが正常に稼働していることを確認する。

上記システム条件による想定コストを以下のように試算する。

表 7-4　OCR システムに関する想定コスト

手法	区分	内容	想定コスト	合計
既存 OCR ソフト（1 種）	ソフトウェア	OCR ソフトライセンス バッチ処理システム開発	約 45 百万円 ※1	約 100 百万円 （約 6.7 円/件相当）
	サーバ	サーバ構築・利用料 運用・監視	約 55 百万円 ※2	
AI-OCR	ソフトウェア	教師データ準備 AI-OCR エンジン生成 バッチ処理システム開発	約 100 百万円 ※3	約 600 百万円 （約 40 円/件相当）
	サーバ	GPU サーバ利用料 運用・監視	約 500 百万円 ※4	

※1　システム開発費： 40 人月と想定
　　内訳）調査・要件定義・設計：30 人月、開発・テスト：10 人月
※2　25 台×10 カ月間確保と想定
　　第 4 章 3.処理能力調査より 2 種 OCR の場合、50 台×8 カ月連続稼働と試算。
　　1 種の場合、サーバ台数は 50%の 25 台。連続稼働前後の期間のサーバリソース確保も必要であるため、10 カ月間確保と想定。
※3　教師データ準備：500～1000 万文字と想定
　　対象となる特許文献から 5,000～10,000 ページをサンプル抽出し、AI-OCR 生成のための教師データ形式（文字の矩形情報および文字コードの集まり）へ、機械的または目視による作業により変換する作業を実施。
※4　GPU サーバ 2～4 台と想定

b）レイアウト解析システム

第 4 章の調査を基に、前節で示した 3 区分に関してレイアウト解析システムを開発した

場合の想定コストを以下に示す。

システム条件

- OCR 認識結果情報（矩形情報付のテキストの集まり）から、書誌情報等を自動処理により抽出する。
- 書誌情報等の抽出精度に関する品質保証はなし。
- 目視によるテキストの修正（校正）は行わない。
- 0.1%程度の抜き取り検査を実施し、システムが正常に稼働していることを確認する。

表 7-5　レイアウト解析システムに関する想定コスト

レイアウトの特徴	区分 No	ラベル抽出精度	レイアウト解析システム開発想定コスト
[I] 旧字旧仮名遣い	B-①② D-①②	3%	約 20 百万円 （目標精度：約 50%）
	B-③, ④-（1～2） D-③, ④-（1）	14%	
[II] 画質が悪い資料 （新字新仮名遣い）	A-①② A-b-① B-④-（3） D-④-（2） E-①②④ H-a-①②	58%	約 20 百万円 （目標精度：約 80%）
[III] 画質が良い資料 （新字新仮名遣い）	A-③④ A-a-①②③ B-⑤⑥ C-①②③④ D-⑤⑥ E-③ H-a-③ H-b-①	54%	約 20 百万円 （目標精度：約 90%）

※システム開発費：各約 20 人月と想定
内訳）調査・要件定義・設計：10 人月、開発・テスト：10 人月

※レイアウトパターンの詳細な分析を、調査・要件定義時に改めて実施し、テンプレート化。レイアウト解析システムは、どのテンプレートに該当するかの判定とレイアウト解析を同時に行うような動作設計を想定。

[I] 旧字旧仮名遣いの資料

今回の予備調査では、ラベル文字列の抽出がほぼできなかった。より詳細な分析と専用プ

ログラム開発で、精度向上の可能性はあるが、50%程度にとどまると想定される。

[II] 画質が悪い資料

今回の予備調査では、ラベル文字列の抽出精度は 58%であった。専用プログラムの開発により約 80%程度の精度での抽出が可能と想定される。

[III] 画質が良い資料

今回の予備調査では、ラベル文字列の抽出精度は 54%であった。専用プログラムの開発により約 80%程度の精度での抽出が可能と想定される。

c) マスタ DB 連携ツール開発

システム条件

・連携対象マスタ DB に関する基礎情報（スキーマ情報、件数、等）は、事前に得られるものとする。
・連携対象マスタ DB からデータを抽出する作業は工数に含まない。ファイルとして、コンピュータで取り扱うことができる形式（CSV, XML 等）で提供されるものとする。
・イメージデータから得られる情報（公報種別・年代・ページ数等）およびイメージデータからテキスト化しレイアウト解析して得られた書誌情報と、マスタ DB を自動処理により紐付けるものとする。
・目視による情報の修正は行わない。
・0.1%程度の抜き取り検査を実施し、システムが正常に稼働していることを確認する。

約 20 百万円（一式）

※約 20 人月と想定　内訳）調査・要件定義・設計：10 人月、開発・テスト：10 人月

d) 変換ツール開発

システム条件

・OCR により得られたテキスト、レイアウト解析またはマスタ DB 連携ツールにより得られた書誌情報を基に、公報テキストデータフォーマットに沿うデータ形式に自動変換する。
・目視による情報の修正は行わない。
・0.1%程度の抜き取り検査を実施し、システムが正常に稼働していることを確認する。

約 20 百万円（一式）　※約 20 人月と想定

※約 20 人月と想定　内訳）調査・要件定義・設計：10 人月、開発・テスト：10 人月

(3) 区分ごとの想定コスト

以下に、前節で述べた 3 区分ごとに①～④の各システムの想定コストを割り当てた一表を示す。

表 7-6　想定コスト（システム全体）

レイアウトの特徴	区分 No	想定件数 (万件)	①OCR システム			②レイアウト解析ツール	③マスタ DB 連携ツール	④変換ツール	合計
			単価 (円/件)	最適ツールを使用	(参考) 低コスト優先				
[I] 旧字旧仮名遣い	B-①② D-①②	46	40	20 百万円 目標精度： 93%程度	3 百万円 想定精度： 22%	約 20 百万円 目標精度： 約 50%	約 20 百万円	約 20 百万円	約 80 百万円
	B-③,④-(1～2) D-③,④-(1)	5							
[II] 画質が悪い資料 (新字新仮名遣い)	A-①② A-b-① B-④-(3) D-④-(2) E-①②④ H-a-①②	507	40	203 百万円 目標精度： 87%程度	34 百万円 想定精度： 22%	約 20 百万円 目標精度： 約 80%	約 20 百万円	約 20 百万円	約 263 百万円
[III] 画質が良い資料 (新字新仮名遣い)	A-③④ A-a-①②③ B-⑤⑥ C-①②③④ D-⑤⑥ E-③ H-a-③ H-b-①	1,072	6.7	72 百万円 目標精度： 95%程度	—	約 20 百万円 目標精度： 約 90%	約 20 百万円	約 20 百万円	約 132 百万円

合計： 約 475 百万円

4. 総括

今後の本格的なテキスト化実施に向けて、既存 OCR ソフトおよび AI-OCR の使い分けや多数決処理により、テキスト化の精度を最大化するための方策を検討した。

既存 OCR ソフトには、20〜60%程度と極端に文字認識精度が悪くなる区分が存在するのに対し、対象文献に対してチューニングされた AI-OCR では、ほぼすべての区分において安定して高精度を実現可能であることを示した。今回の調査事業で試験的に生成した AI-OCR は、今後教師データの拡充、AI モデル見直し等を実施することでさらに数ポイントの文字認識精度の改善が期待可能と考えられる。

AI-OCR の認識精度が既存 OCR ソフトと比較して顕著に改善した区分は、旧字旧仮名遣いの古い年代（1920〜1940 年代）の資料、および新しい年代（1950 年代以降）の画質が悪い資料に関しての 2 つである。

旧字旧仮名遣いの資料の認識精度が悪い原因は、漢字の旧字体の問題と当時使用されていた活字の字形が現代の OCR と適合しないことの 2 点であると考えられる。特に旧字体には既存 OCR ソフトの仕様上読み取ることができない JIS 第二水準外の漢字が多く含まれるため、既存 OCR ソフトで対応することができない。かすれ・つぶれ等画質が悪い資料の文字認識精度は、第四章の調査の結果、レイアウト・活字の字形が類似していても OCR 認識精度低下の主要の要因になることが確認された。

一方、年代が新しく画質が良い資料に関しては、既存 OCR ソフトと AI-OCR の文字認識精度に大きな差はなく、費用対効果を考慮すれば、既存 OCR ソフトで処理することが最適であると考えられる。

したがって、今後約 1,500 万件のテキスト化にあたり最適な OCR ソフトを適用するための区分としては、[I] 旧字旧仮名遣い、[II] 画質が悪い資料（新字新仮名遣い）[III] 画質が良い資料（新字新仮名遣い）の 3 区分が妥当であると判断される。

また、レイアウト認識（文字領域の推定および行矩形の抽出）に関しては、年代・画質による影響よりも、どの紙面要素（書誌情報部分、本文、図版等）を優先的に処理するべきかという要件による決まるという性質があり、また容易にチューニング可能な項目であるため課題としての優先度は低いと考える。

レイアウト解析（書誌情報項目の抽出）に関しては、OCR ソフトによる文字認識後に実施する必要があるため、OCR ソフトの文字認識精度の影響を強く受ける。したがって、レイアウトの細かいパターンよりも、まず OCR ソフトの文字認識精度の向上を優先し、上記 3 区分による OCR ソフトの適切な使い分けが必須であると考えられる。レイアウトの細かいパターンの違いに対応したレイアウト解析機能は既存 OCR ソフトには存在しないため、いずれにせよ個別に開発する必要がある。

レイアウト解析の精度に関しては、文字認識精度による限界、書誌項目名の表記揺れの問題等があるため、自動認識処理では実用精度に達しない可能性がある点にも留意する必要がある。

全件のテキスト化にあたり、上記以外に必要とされるシステムは、外部のマスタ DB 等との連携のためのツールおよび公開テキストフォーマットへの変換ツールの 2 点であると想定される。

上記想定の基づき、全件のテキスト化を行った場合の総コストは、475 百万円程度と推計される。

また、本事業の目的である機械翻訳における使用について考察すると、高精度な機械翻訳を可能とするためには、平均的な文字認識率の向上に加えて、以下の自動翻訳特有の課題が想定される。

1）行のつながりを正しく推定すること。
OCR のレイアウト認識の際、行矩形の認識に成功したとしても、行矩形間のつながり方の推定に失敗すると、正しい機械翻訳の結果は得られない。

2）文・文節の区切りを正しく識別すること。
OCR は単純な記号の識別に失敗する傾向があるが、句読点の識別に失敗すると、文・文節の区切りの判断ができず、正しい翻訳翻訳の結果は得られない。

上記事例のように、少数の誤認識箇所が翻訳精度に大きな影響を及ぼすようなケースを分析し、重点的にチューニングすることができれば、特許庁がテキストデータを保有していない又は高精度なテキストデータを保有していない特許及び実用新案の各種公報について、その高精度な機械翻訳文の提供をすることが可能であると考えられる。