人工知能技術を活用したデータエントリー業務の高度化・効率化について

About advance and efficiency improvement of data entry making use of artificial intelligence technology

特許庁　審査業務部出願課知的財産情報分析官
藁谷　智雄
1986 年入庁、特許庁総務課、秘書課等事務関連部署に従事の他、近畿経済産業局特許室、(独)工業所有権情報・研修館などを経て、2017 年 7 月より現職
waragai-tomoo@jpo.go.jp　TEL 03-3580-5882(直)

1　はじめに

特許庁では、平成 28 年度から、特許行政の高度化・効率化に資することを目的に、人工知能（AI）技術の特許行政事務への適用可能性の検討を行ってきた。平成 29 年 4 月には、検討の結果を踏まえ、将来的な人工知能技術の活用を視野に入れ「特許庁における人工知能（AI）技術の活用に向けたアクション・プラン」を取りまとめて公表した。

平成 29 年度は、このアクション・プランに沿って、「電話等の質問対応」、「紙出願の電子化」、「特許分類付与（テキストに基づく付与）」、「先行技術調査（検索式の用語の拡張、ヒット箇所のハイライト表示）」、「先行図形商標の調査」、「指定商品・役務調査」の 6 つで調査事業を実施した。

本稿では、平成 29 年度に実施した調査事業のうち「紙出願の電子化」についての実証実験内容について紹介する。

特許庁への電子出願で可能な手続が書面で行われた場合に必要となるデータエントリー業務は、一定の専門的技術と相当の設備を要すること等から、外部の機関である登録情報処理機関を活用して行われている。データエントリー業務では、光学文字認識（OCR）の利用と人による手入力及び目視確認等の作業によって、年間約 16 万件に及ぶ出願書類等のテキストデータ化を行っている。

特許庁審査官による実体審査を初めとした特許庁の各業務は、このデータエントリーされたデータ（電子原本）で行われるため、極めて正確なデータエントリーが求められる。また、一次審査通知までの平均期間、権利化までの期間が短縮される中、データエントリー業務においても適時性が求められている。

現在、IoT によるデータ量の増加とそれによるビッグデータの活用、PC の処理能力の飛躍的な向上等を背景に、人工知能関係技術の急速な発展が見られ、多くの産業分野への応用が期待されている。

そこで、出願書類等のデータエントリー業務において、人工知能技術の活用による更なる業務の効率化、文字認識精度の向上、作業時間の短縮等を目指して、実証実験を実施した。

2　実証実験の全体像

本実証実験では、書面で提出された書類のうち、特許願及び実用新案登録願、並びに意見書及び補正書を調査の対象とし、比較検討用の既存の複数のシステムを活用して実施した。

書面で提出される書類には、パソコンで作成された「活字文書」と「手書き文書」が存在するが（ごく一部に活字・手書き混在も存在）圧倒的に活字によるものが多い。

表１　提出書類の活字割合

書類種別	活字割合
特許願	94.8%
実用新案登録願	95.8%
補正書	99.1%
意見書	96.7%
全体	98.0%

特許出願書類等のデータエントリーには複数の工程があるが、今回の実験では、活字文書文字認識、手書き文書文字認識、レイアウト認識の３つについて検証を行った。

活字文書と手書き文書では、OCR プログラムや AI を用いた文字認識の技術的難易度は大きく異なるため、本実験では活字文書と手書き文書を分けて扱うこととした。このうち、活字文書については、予備調査段階で OCR の有用性は既にほぼ明らかであったため、主に OCR プログラムと人工知能の比較検討が必要であるという趣旨から、①既製 OCR プログラム２種類と②人工知能による文字認識を行い比較検討する実験系を構築した。

一方、手書き文書では、そもそも機械入力の対象とせず手入力を維持するという選択肢も十分に検討の範疇であったため、①手入力と②既製 OCR プログラム及び③人工知能の３種類による検証を行う実験系を構築した。

文書のデジタル化を行う場合には、スキャニングした画像からそのレイアウトを解析し、文字領域と図表領域を分離し、認識すべき文字列を切り出したうえで認識する文字の矩形を特定する工程と、切り出した文字をデジタルの活字として認識していく工程の二段階が存在する。

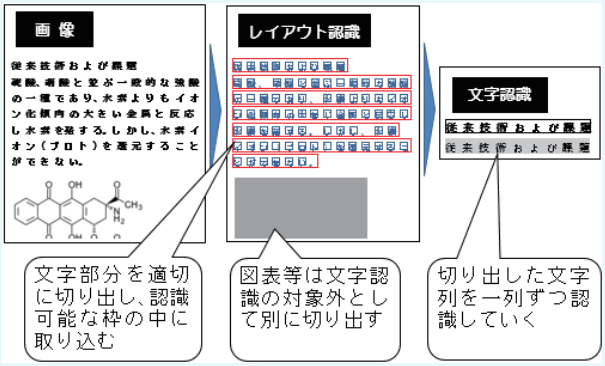

図２　データエントリーのイメージ

一般的に市販活字認識ソフトの文字認識の成功率は 99% 以上であると説明されることが多いが、これは、適切に文字列を切り取ったうえで認識矩形を特定することができた後の認識率であることがほとんどである。文字段階の認識精度がどれほど高くても、適切に文字列を切り取り、認識矩形を設定できなければ、正しく認識することはできない。

このように、文書のデジタル化を考える場合、レイアウト認識の段階と文字認識の段階のそれぞれについて、その精度を高めていく必要がある。したがって、本実験においては実験条件を揃えるために、レイアウト情報（文章領域＋行＋個々の文字の外接矩形）が正しく設定された座標データを所与のものとし、個々の文字の認識率のみのテストを各実験システムで実施した。

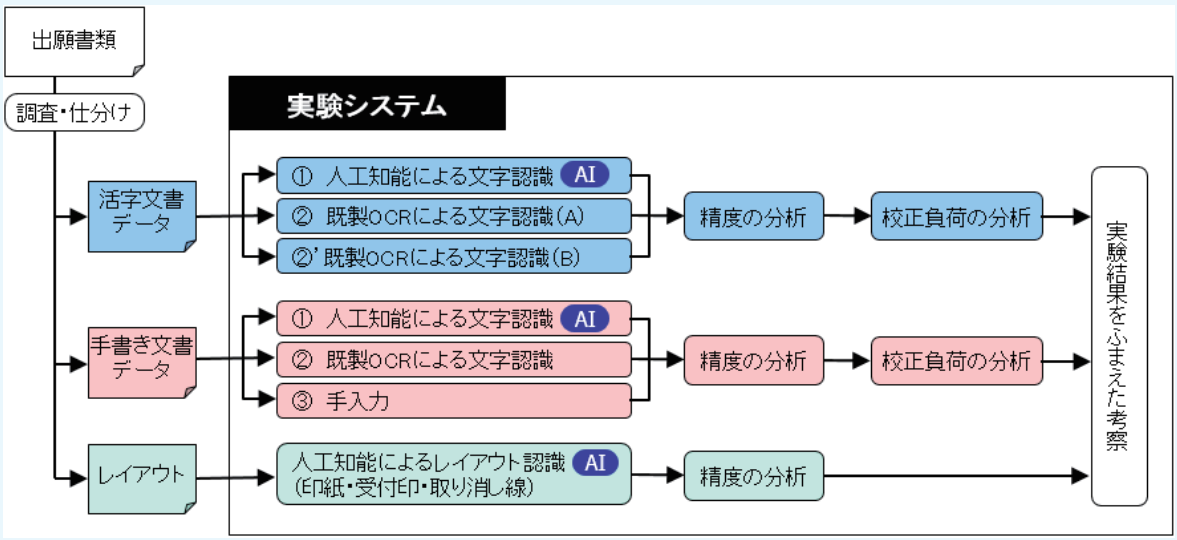

図１　実証実験の全体像

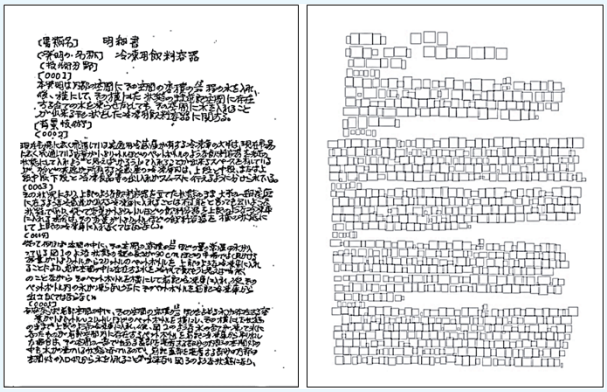

図３　手書き文書の文字枠設定イメージ

3　テスト用実験システム

3.1　活字文字認識テスト用実験システム

①　人工知能による文字認識システム（AI）
②　既製 OCR ソフト（A 社製）
②′　既製 OCR ソフト（B 社製）

データエントリーの処理内容は、人工知能にとってそれほど特殊な処理ではなく、今回の実証では一定の教師データの存在を前提にすることができることから、AI 実験系については、委託先の既存システムを活用した。

また、既製の OCR ソフトを活用した実験系は、信頼性の高い市販の二つの OCR ライブラリを活用することにした。この二つのライブラリは、その認識精度に定評があることに加え、アルゴリズムに差異がある。異なる OCR ソフトを選択することで、２つのエンジンの処理を掛け合わせた場合の精度向上を狙うために選定した。

3.2　手書き文字認識テスト用実験システム

①　人工知能による文字認識システム（AI）
②　既製 OCR ソフト
③　人力による文字入力（手入力）

AI 実験系については、活字と同じく委託先の既存システムと既製 OCR ソフトを使用した。また、比較のため、人力による入力テストを実施した。

これは、手書き文書においては、認識精度においても、認識処理にかかるコストにおいても、機械入力が手入力に対する優位性を持たない可能性があるという仮説を持っていたためである。

手書き文書のデジタル化に関し、AI や OCR を活用するとしても、認識精度が低い場合には、最初から手入力した方が、全体のコストが小さくなるということも十分に考えられる。特に、手書き文書の数が極めて少ないことから、手書き文書については従来通りの手入力を維持するという結論も十分に考えられる。

なお、人力による入力テストは、一般的な WindowsPC を使用した。

3.3　レイアウト認識テスト用実験システム

検証において検討すべきレイアウト上の要素は、レイアウト認識の精度向上を阻害していると想定された(1)印紙、(2)各種印影、(3)取り消し線について、これらの要素を人工知能技術の活用によって除去可能かどうかの検証を行った。

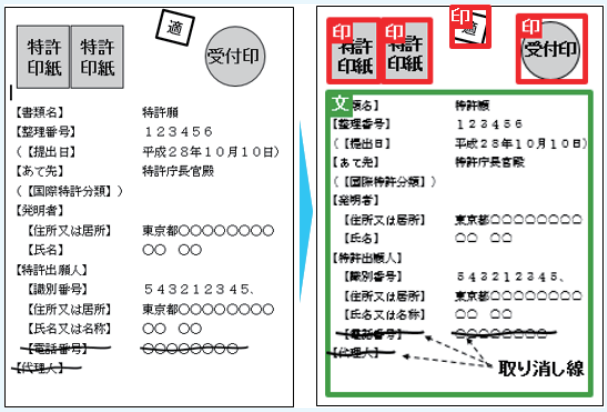

図４　レイアウト認識のイメージ

4　実験結果及び分析

4.1　活字文字の自動認識の結果

①　人工知能による活字文字認識の結果：98.7%
②　既製 OCR による活字文字認識の結果（A 社製）：97.6%
②′　既製 OCR による活字文字認識の結果（B 社製）：98.3%

4.1.1　結果の分析

人工知能による文字認識及び既製 OCR による文字認識ともに、97%～98% の認識精度を出すことができた。この認識水準は、「一般的な文書において、費用対効果の観点からは十分に実用にたえうる水準」であるが、特許出願書類等のデジタル化に際して求められる極めて高い品質水準を前提にする場合、「どうしても誤認識文

字はある程度残る」「校正の適切な実施が不可欠」であると考えられる。

誤認識文字については以下のような傾向がみられる。

（1）単純な形（線、点、丸、四角、カッコ類）の区別がつかない　例）()、〇、【】、□、–、…、等

（2）日本語文章中の欧文について認識精度が悪い

（3）カタカナの認識精度が悪い

人工知能システムにおいて、認識結果を蓄積したうえでそれを反映した教師データを組成し、教師データの高度化を行えば、更なる認識精度の向上が期待できる。

OCR システムにおいても、既存機能によるチューニングを充分に実施すればコンマ数ポイント程度の精度向上は容易に図れるものと推測する。

より効果が大きいと考えられるのは、特許出願書類特有の使用される文字種の傾向を処理ルールとして適切に登録することにより、処理ルールの高度化を図る方法である。

このように、今後、人工知能と OCR の両システムにおいて、更なる精度向上の余地があり、特に人工知能システムについては、教師データの高度化を継続的に図っていくことで、優位な精度向上が見込まれるが、本データエントリー業務における要求精度からすれば、校正は不要にはならない。

4.2　手書き文字の自動認識の結果

① 人工知能による手書き文字認識の結果：89.2%

② 既製 OCR による手書き文字認識の結果：57.0%

③ 手入力の結果（20 名が同一の書類を入力）

・入力効率：35.3 文字／分（24.53 文字／分～63.5 文字／分）

・入力精度：98.9%（98.31% ～100%）

表 2　1,000 文字あたりの誤字数

手入力	12 文字
AI	108 文字
既製 OCR	430 文字

4.2.1　結果の分析

（1）人工知能による手書き文字の認識精度は 89.2% であり、これは、手入力よりも精度が低いものの、既製 OCR（57.0%）よりは大幅に精度が高いという結果であった。

簡単な文脈補正によりさらに精度を数ポイント向上させることは容易と想定される。

（2）既製 OCR による手書き文字認識の精度は 57.0% に止まり、実用水準には満たなかった。既製 OCR ソフトの手書き文字認識は実用水準には至らないと想定していたため、本実験はその裏付けとなる結果であった。

（3）手入力については、1 回入力での精度は 98.8% 程度であった。今回、入力者本人による「見直し」をしないように指示したため、1.5 倍程度の時間をかけて見直しまで行わせれば 99.0% 程度に引き上げられると想定している。ただしその場合コストも 1.5 倍になる。

入力効率は平均 35 文字／分程度であり、一般にデータエントリー業務では 60～100 文字／分程度入力できるとされているが、それよりも低い結果になっている。その原因として、以下が推察される。

－乱筆の文書が多く見られる

－文字の密度が異常に高い文書が散見される

－使用されている語句が難しく、文脈による推論が困難

また、画像の解像度が高くなれば入力が効率的になると想定される。

（4）以上を総合すると、手書き文書の認識においても、特に人工知能を用いた文字認識には期待が持てる結果が出たといえる。

現時点では、90% 以下の認識精度と、実用水準とまではいえない認識率であるが、今後、手書き出願書類等から採取した字形を教師データとして人工知能に学習させることで十分精度の認識エンジンの開発は十分に可能と考えられる。

ただし、今回の実験は正しい文字枠を人力で作成しており、人工知能を活用した場合の総コストの低減効果は極めて限定的と言わざるを得ない。

実際の導入が可能なシステムにするためには、文字枠の自動切り出しの問題を先に解決する必要がある。しかし、手書き文書のレイアウトを適切に分析し文字枠を切り出すことができるシステムは、既存 OCR にも人工知能にも存在していないため、これを一から開発するためには、莫大なコストと時間が必要となる。

4.3 レイアウトの自動認識の結果

① 印紙 ：認識率 100%
② 受付印 ：認識率 100%
③ 取り消し線：認識率 69.1% 検出漏れ 30.9%

4.3.1 結果の分析

印紙及び受付印について 100% の精度で認識した。願書のレイアウト認識については、十分な自動化メリットが出しうると考えられる。

一方、取り消し線については、一定の検出漏れ（30.9%）や過検出が発生した。印紙や受付印に比べ、取り消し線は、本文中に予告なく存在し、また必ずしも正確な直線であらわされていないなど、検出上の難易度は高かった。

5 校正

校正作業は、それ自体は人力で行われるものであるが、特許出願書類等のデータエントリー業務のように、極めて高い認識精度が求められる業務では、必ず必要となるプロセスである。

どれほど精度の高い OCR や人工知能を用いたとしても、現状の技術水準を前提とすれば、自動化されたシステムのみで極めて高い水準の正確性を確保することはできないため、人力による校正プロセスが複数回にわたって必ず必要となる。

実験では、校正者 6 名により手書き文書の画像と入力したテキストデータをモニター内に並べて比較校正する「画面内の引き合わせ校正」を実施した。

5.1 校正の結果

（1）手入力によって作成されたテキストデータ（精度 98.9%）に対する校正
・校正効率：114.9 文字／分
・1 回目校正後の精度：99.3%（誤字見逃し率 59%）
（2）人工知能による文字認識により作成されたテキストデータ（精度 89.2%）に対する校正
・校正効率：32.5 文字／分
・1 回目校正後の精度：98.9%（誤字見逃し率 11%）

5.2 校正結果の分析

校正作業の作業効率は 99% 程度の高精度テキストデータからの校正の場合、115 文字／分程度であり、入力工程の約 3.3 倍の効率である。これは、校正には 1 回あたり入力の 30% のコストがかかることを意味している（人件費単価は同一と仮定）。

一方、89% 程度の低精度テキストデータからの校正では、32.5 文字／分と手入力 35 文字／分とほぼ同等であり、「初めから手で入力したほうが早い」という経験則を裏付ける結果となった。

また、1 回の校正ではすべての誤入力文字を発見・修正することはできず、高精度（99% 程度）からの校正の場合、約 59% の誤字を見逃してしまうことが確認された。

6 考察

6.1 活字文書

既製 OCR ソフト及び人工知能における認識精度が 98～99%と完成度は高いが、特許出願書類等のデータエントリーで求められる極めて高い精度を実現するためには、なお精度が十分ではなく、1% 程度の誤認識が残る状況である。

将来的な導入に当たっては、OCR の誤認識パターンを蓄積し、人工知能技術による文字認識エンジンへ再学習させるような仕組みにより、将来的に超高精度の文字認識技術が実現されるような取り組みが必要と考えられ、教師データの蓄積による精度向上が求められる。

6.2 手書き文書

人力による手入力と校正の組み合わせが当面は有効と考えられるが、人工知能技術による文字認識と手入力との差は急速に縮むことが期待される状況にある。一方、フリーレイアウトの手書き文書の文字枠を正確に切り出すことは今後とも困難であることが予想される。

むしろ、特許出願等の提出書面自体について、手書き用のフォーマットを作成公開し、当該フォーマットにはあらかじめ升目、枠線（OCR 読み込み時にノイズとな

らない工夫が必要）を入れておくことで、事実上レイアウトを固定するといった運用上の工夫によって人工知能の活用を短期的に可能とすることも考えられる。

6.3　校正

極めて高精度の品質保証のためには、目視による複数回の校正を繰り返すことが今後とも必須である。効率化のためには、校正回数を 1 回でも減らすことが重要であり、自動認識段階の精度を極限まで向上させる取り組みが必要と考えられる。

6.4　その他の施策案

・スキャニング画像の高精細化をはかり、文字の潰れを防ぎ、OCR 認識精度を向上させる
・自動認識に適さない書面を、人工知能技術により事前に仕分け、手戻りを防ぐ仕組みの実現
・願書特有の OCR 不要要素（日付印、取り消し線等）の人工知能技術による自動除去の実現

7　おわりに

現状ではどれほど人工知能を活用した文字認識技術が進んだといっても 100% 正確に認識することは困難であり、完全に自動化することはできない。特に手書き文書については、実用化にすら至らない状況である。ただし、今後の技術革新により人工知能を活用した OCR 文字認識と手入力との差は急速に縮むことが期待できよう。

今後急進する文字読み取り技術の動向を注視しつつ、今回の実証実験で得られた結果及び課題等を念頭に、引き続き更なる業務の高度化・効率化を目指してまいりたい。

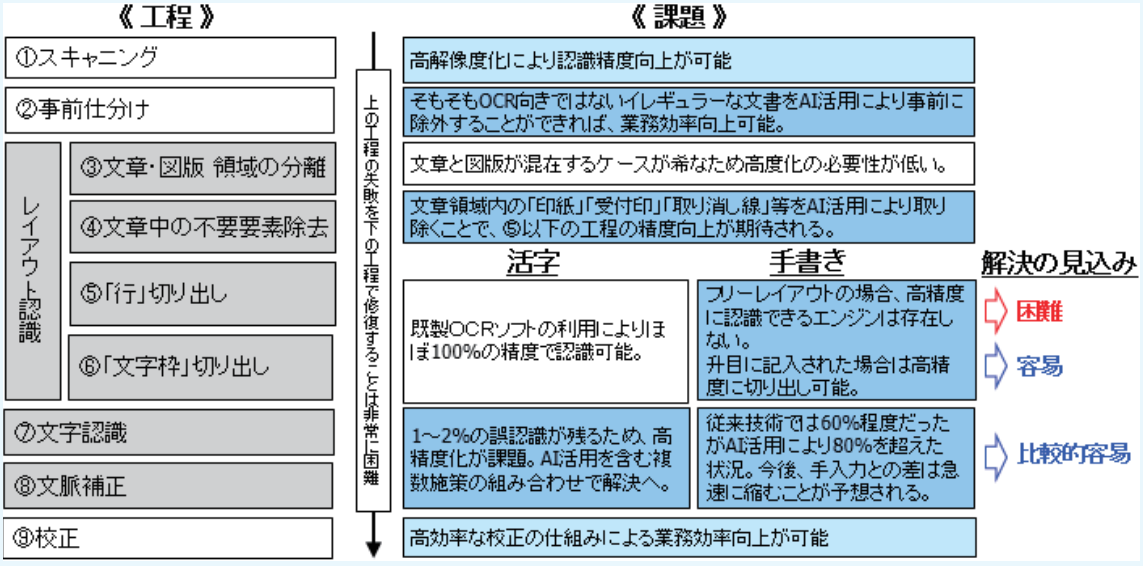

図 5　工程ごとの業務効率化・高度化のために解決すべき課題の考察