青山学院大学理工学部情報テクノロジー学科
2024 (令和 6) 年度 卒業論文要旨

# OCR とラベリングによる書類整理自動化システムの構築と有用性の評価

嘉松 一汰　(15820094)
Dürst 研究室

## 1　はじめに

### 1.1　研究背景

日本企業の RPA (Robotics Process Automation) 導入率は全体で 38%, 中小企業では 25%となっており, 非常に少ないことがわかる図 1. また, 大企業と中小企業の間に 20%以上の差があり, 技術や規模による格差も見て取れる. これらの原因は主に 2 つある. 1 つ目は, RPA は紙媒体の処理等の, アナログの世界で行われる処理は非専門領域であるということである. 具体的には, 縦書き文字横書き文字が混在していたり, 旧字体や特殊文字等の組み合わせも考えられるため, 例外的な処理までを自動で行う必要があるためである. 2 つ目の原因は, 紙媒体の業務を行っているコミュニティの IT 知識の乏しさにある. 詳しくは次のセクションで説明する.


**図 1: 企業の RPA 導入率**

### 1.2　研究目的

本研究の目的は, RPA を使用して紙媒体の処理を自動で行うシステムを作成することである. 前述したように, システム障害への恐怖感や,IT 知識の乏しさ等によって, IT 知識に乏しいコミュニティが現代には数多く存在します. このような現状を改善するために, 一連の流れを RPA にすることで,IT 知識の有無に関わらず, システムが運用可能になることを目指す.

## 2　関連研究

本研究では, あくまでカテゴライズという目的でシステムを構築しているため, クラスタリング手法とは少し異なる点があるが, アルゴリズム等の観点で参考にしたため, 関連研究として以下に示す.

### 2.1　クラスタリングの既存手法

**非階層的クラスタリング**　目的のデータを事前に定義されたクラスタ数に分解することによって行われるクラスタリング方式のことである. 代表的な手法として, クラスタ内の分散を最小化するようにデータポイントをグループ化する k-means アルゴリズムが挙げられる.

**階層的クラスタリング**　データポイントを機構造の階層に分割して行うクラスタリング方式のことである. 代表的な手法には大きく分けて, 凝集型と分割型の 2 種類がある. 凝集型は, 木構造の下から上へクラスタを統合していく方法である. 対して分割型は, 上から下へクラスタを分割していく方法である.

## 3　提案手法

### 3.1　データベース設計

本研究で使用するテーブルには 3 種類あり, 図 2 にこれらの ER 図を示す

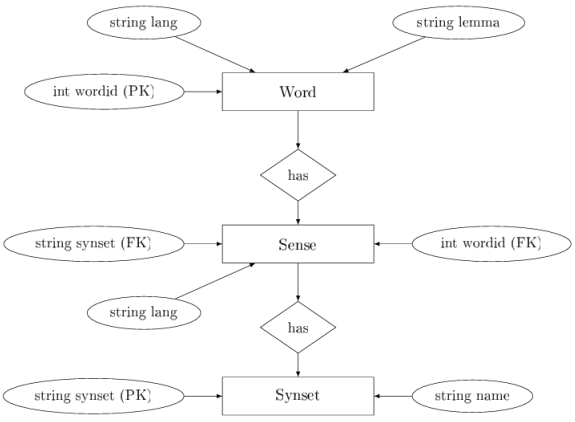


**図 2: データベース構成**

### 3.2　モデル設計

**Word モデル**　一対多の関係で, 複数の Sense モデルへのリレーションを持っており, Sense モデルを経由して, Synset モデルへのリレーションを辿ることで, 単語からその定義を取得する. 場合に応じて適切な言語のレコードを活用するため, 言語での絞り込みを行うスコープを持つ.

**Synset モデル**　一対多の関係で, 複数の Word モデルへのリレーションを持っており, Sense モデルを経由して, Word モデルへのリレーションを辿ることで, 特定の定義をもつ単語を取得する. Word モデルと同様に, 言語での絞り込みを行うスコープを持つ.

**Sense モデル**　上記の 2 つのモデルそれぞれと一対多のリレーションを持っており, 二つのモデルの中間テーブルの役割を担う. 単語から定義を取得する処理は,Synset テーブルを経由して行う.

### 3.3　提案する手法・アプローチ

カテゴライズしたいドキュメントファイルに対して Google Vision API を使用した OCR を行い, 内容の単語群を取得する. その後,WornNet データベースと連携し, 単語ごとの定義を取得する. 各々の定義を元に単語がどのカテゴリに属するかを判別し, 単語ごとにラベリングを行う. ラベリング結果から, 最終的なドキュメントのカテゴリを決定する.

## 4　実験・評価

本研究の実験で用意するカテゴリは, 経理, 人事, 事務の 3 種類であり, 各カテゴリごとに 5 種類ずつの計 15 種類のドキュメントに対してカテゴライズ処理を実行し, 精度の分析を行った. 結果として, 処理制度にはタイトルごとにばらつきが見られた. 特に, 旧字体を含む文書では処理が失敗する傾向が確認された. 一方で, 縦書きや横書きが混在する文書, 英語表記が一部含まれる文書, ゴシック体や太文字を用いた文書などにおいては, 問題なく処理が成功している. これらの結果から, フォーマットや字体の違いが OCR の精度に与える影響が明確になった. また, カテゴリごとの処理精度について, 経理カテゴリにおける精度が他のカテゴリと比べて著しく低いことが確認された. この主な要因としては, 経理カテゴリの文書に旧字体や特殊な専門用語が多く含まれている点が挙げられる. これにより, OCR 処理のカテゴリごとの安定性について一定の評価を下すことができる.

## 5　考察

### 5.1　結果の解釈

本システムの優位点については, ドキュメントの形式に関わらず同様の処理結果を得ることができる点である. 文書には, 縦書きや横書き及びそれらの複合など, 内容の形式はジャンルによって多岐にわたる. それら全てに対応出来なければ, 本システムの優位性は著しく低下してしまう. しかし, さまざまな形式の文書を実験では用意したが, 特に形式の差による精度の変化は見受けられなかった. そのため, 本システムにおいては, 文書をアップロードする前に形式を確認する等の校閲を行う必要がなく, その点ではストレスなくシステムを運用することができると考えられる. 本システムの有用性については, 第 **??**章で大まかに述べたが, 紙媒体中心の業務に不便を感じている人が現代に多くいるため, そのような状況を改善するという観点では, 要件を満たすことができているといえるだろう.

### 5.2　改善点

本システムの改善点は, 旧字体など, 現在使用されていない字体を用いた文書に関しては, 精度が落ちてしまう点である. 具体的には, OCR 処理では不自由なく単語ごとに分けることができているが,WordNet データベースにその単語が存在しないため, 単語のカテゴリが判別できないことが原因として挙げられる. 改善するためには, 現状使用している WordNet データベースの他に, 漢字字体規範史データセットのような旧字体を扱っているデータベースを使用し, 判別可能な語彙を拡張することが必要である. また, 現状のシステムでは, 日本語と英語を主な対象としているが, グローバル化が進む現代社会では, 多言語対応が求められる場面が増えている. これを実現するためには, 各言語に特化した語彙データベースを統合し, 多言語間でのカテゴリ解析を可能にするアルゴリズムの開発が必要である.

## 6　おわりに

### 6.1　今後の展望

今後の展望として, ユーザーインターフェース (UI) の最適化も重要である. 現状ではシステムのコア機能に重点を置いているが, 操作の直感性やユーザーエクスペリエンスの向上が求められる. 例えば, 解析結果を可視化するダッシュボードや, フィードバック機能を備えたインターフェースを実装することで, 利用者の利便性を高めることが可能である. また, 本システムをクラウドベースで提供することにより, 複数のユーザーが同時に利用できる環境を構築し, スケーラビリティを確保することができる. また,API を公開することで, 他のアプリケーションやサービスとの統合を可能にし, システムの拡張性をさらい高めることも有効である.

### 6.2　研究のまとめ

本システムは, 現在の業務フローにおける課題を解決し得る有効なツールであると結論付けられる. 現代では, 今後も IT 分野の著しい成長に遅れをとったコミュニティを中心に, さまざまな業務的ニーズが発生すると考えられる. 本研究では紙媒体中心業務をさらに細分化した, ドキュメントのカテゴライズ分野に着目してシステムの作成を行ったが, 本システムのようなソリューションが生み出されることで, IT 社会の課題が解決されることを願っている.