

# 機械学習を用いた文書検査法の提案と評価

プロジェクトマネジメントコース 矢吹研究室 1442069 須山武弘

## 1. 背景

人間とロボットの対話や、音声認識でのデバイス操作、機械翻訳などが増えており、コンピュータが自然言語を理解し、処理、出力することが要求されている。この際、言語を理解させ、人間とコンピュータの意思疎通を可能とさせる手段が自然言語処理である。

自然言語とは、人間がコミュニケーションを取る際、自然に発している言語のことである。自然に発している言語が故に、プログラム言語や数学的表現と違い、発言者自身の文化や曖昧さが含まれている。この曖昧な表現や発言者の文化を含んだ言語を直接コンピュータに認識させる事はできず、認識をさせるには自然言語処理が必要である。

自然言語処理とは、自然言語に含まれる発言者自身の文化や、曖昧さを含め、言葉の意味をコンピュータに処理させる一連の技術である。コンピュータと人間の意思疎通の必要性向上により、高度な自然言語の理解をする必要があるため、言語処理技術の向上が研究されている [1]。

本研究で用いる IBM Watson は、認知した情報に基づいて理解、推論し、学習するコグニティブ・コンピューティングのプラットフォームである [2]。13 種類の API があり、自然言語処理、画像処理、音声認識、データ分析など多岐にわたる分野で利用できるプラットフォームである。API が公開されており、多くの企業で質問応答システムや、自然言語のデータ化などのシステム構築がなされている [3]。

## 2. 目的

機械学習を用いた文書検査ツールを実装するため、自然言語処理に長けたプラットフォームである IBM Watson を使用する。この文書検査ツールにおいて、RedPen などの文書検査ツールより正確な文書の検査結果が得られるかを検証する。

## 3. 手法

本研究は、IBM WatsonAPI を用い、文書検査ツールを実装する。実装した文書検査ツールで文書検査を行う。次に、出力結果の有用性について検証、考察する。さらに、改善が必要な場合には改善を行い、検証、考察する。

その後、他の文書検査ツールを用い、出力結果の違いを検証、考察する。

## 4. 想定される成果物

文書検査ツールを実装し、その有用性や、他の文書検査ツールの出力結果の違いなどを検証し、考察する。その結果を踏まえて機械学習における文書検査の有用性の評価や改善をまとめたものを成果物とする予定である。

## 5. 進捗状況

IBM WatsonAPI の 1 つである Natural Language Classifier (NLC) を用い、入力した文の分類を認識させた。現在、IBM Watson の他の API を用いて文書の間違いを検査できるようにするように実装している。

## 6. 今後の計画

研究室内の文書データを用い、正確に文書の検査ができるかを試し、論文を執筆する。

## 参考文献

- [1] MahoTakara. 自然言語処理 (NLP) ってなんだろう? <http://qiita.com/MahoTakara/items/b3d719ed1a3665730826> (2017.09.22 閲覧)。
- [2] 井上研一. 初めての Watson API 用例と実践プログラミング. 株式会社リックテレコム, 2016.
- [3] 日本 IBM. IBM Watson 活用事例. <https://www.ibm.com/watson/jp-ja/use-cases/> (2017.09.28 閲覧)。