人工知能（AI）による固有表現抽出

自然言語処理技術の一つである固有表現抽出は、金融機関における文書チェックなど非構造化データを扱う業務の効率化や自動化を高める可能性がある。しかし、ソリューションを導入する際には処理精度に加えて業務視点からの検討が欠かせない。

固有表現抽出の概要と利用が期待される業務

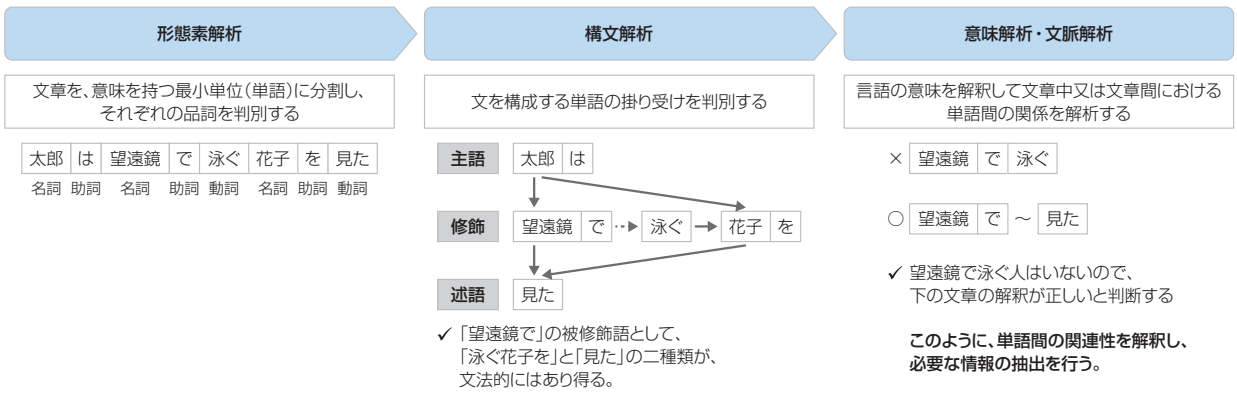
人工知能（AI）の自然言語処理技術の一つとして、「固有表現抽出」というものがある。固有表現抽出とは、文中から人名、地名等の固有名詞や、日付、金額等を、あらかじめ定義された項目に分類することができる技術である。固有表現抽出の一般的なプロセスは、最初に長い文章を単語に分解し、単語間の係り受けを判別して、意味を理解することから始まる（図表1参照）。この技術を利用すると、例えば「東証株価指数（TOPIX）が月間で4.91%上昇し、2019年1月末に1567.49ポイントとなりました」という文章（非構造化データ）から、TOPIXの値である「1,567.49ポイント」という情報（構造化データ）を自動的に抽出することが可能となる（図表2参照）[1)]。

この技術は、金融機関の報告書、レポート等の文書チェック業務などに活用されることが期待される。金融機関における投資家向けレポートや、監督当局向け報告書のチェック業務は、記載事項の正確性や、法令上求められている事項の有無を確認する重要な業務であるが、文書量の増加に伴い業務量も増加しており、チェックの品質を保ちつつも効率化が求められている。また、記載事項の正確性（例えば文章中に記載されたTOPIXの値）を確認するにも、現状は担当者の手作業に頼る部分が大きい。数十頁にわたるレポートの中から記載場所や書き方も異なっているTOPIXの値を、あらかじめ決められたロジックで抽出するのは意外と難しいからである。

従来型のシステム化の手法では、記載場所の特定や数値の取得方法を、網羅的にルール化することは困難である。チェックミスを防止するために、人が何重にもチェックプロセスを重ねるといった対応にならざるを得ず、自動化への壁が高い業務でもあった。

こうした業務に固有表現抽出を実現するソリューション（以下、ソリューション）を用いると、チェックしなければならない項目や、法令上求められている事項を文章の中から自動的に抽出することが可能となる。近年、このようなソリューションが提供されるようになり、文

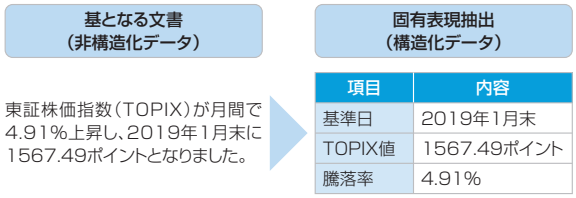
図表1　固有表現抽出のプロセス



NOTE

1) テキスト文章のように、「どこに何があるか」が明確に定められていないデータを「非構造化データ」と呼んでいる。一方、CSVファイルやExcelファイルのように、「列」と「行」の概念があり「どこに何があるか」が明確化しているデータを「構造化データ」と呼んでいる。「非構造化データ」から「構造化データ」に変換することで、後続のデータ処理が容易に行えるようになる。
2) これまで述べた留意点を踏まえ、野村総合研究所でも、固有表現抽出の精度向上に加えて、人間が利用しやすい機能を備え、継続的に利用し続けることを意識したソリューションの開発を行っている。

図表2　構造化された抽出結果



書チェックなどこれまで自動化が難しかった業務に対して、ソリューション適用の機運が高まりつつある。

固有表現抽出を業務に導入する際の留意点

このように固有表現抽出は業務の自動化を高める可能性があるものの、ソリューションの導入にはいくつか留意すべき点がある。

まず1点目は、導入を検討しているソリューションに人が介在しやすい機能が備わっているかどうかが重要なポイントとなる。

AIは、機械学習や深層学習（ディープラーニング）の技術進歩に伴い急速に精度が向上したものの、100%の精度を達成するのは困難である。AIを業務に活かすためには、AIの処理結果の確認や、誤った処理を行った際の修正を人が円滑に行えるよう、確認画面や修正機能など周辺機能の充実がポイントとなる。いかにAIが人を介さずに業務を行うか考えるより、業務の特性や必要とされる精度に応じて人とAIの業務範囲を整理し、どうしたら人が業務を円滑に行えるか、現実的な視点で考える必要がある。

2点目は、業務要件の変化に対して、ソリューションが対応しやすい機能を有しているかである。

次のようなケースがある。ある金融機関では文書から必要な情報を抽出する際にソリューションを利用している。当初は十分な精度で情報を抽出できていたが、業務を継続していく中で精度が低下する問題に悩んでいた。既存文書の書き方の変化や、新規文書の追加により、未学習の内容が増えて、当初の学習モデルのままでは正しく情報を抽出できないことが原因であった。

継続的な学習にその都度大がかりなチューニングが必要になると、学習頻度の低下が情報の抽出精度の低下要因となり、運用コスト面でも課題を抱えてしまう。

この点、データを抽出するために必要な特徴量（データにどのような特徴があるかを数値化したもの）をデータ自体からAIが自動的に学習できるソリューションであれば、学習頻度を保ち情報の抽出精度を維持し、継続的に利用し続けることができる。このような観点から、ソリューションを導入する際には、業務要件の変化に必要な、継続的な学習機能の利便性について留意する必要がある。

固有表現抽出を用いたソリューションを導入する際はいくつか留意点があるものの、AIを活用することで、これまで人手に頼っていた非構造化データを、効率的・自動的に取り扱うことが可能となる利点は大きい。本稿を、AIを上手く、長く使いこなせるソリューションを選定するための参考としてもらえれば幸いである[2)]。

Writer's Profile

須賀 直弘 Naohiro Suga
金融デジタル企画一部
上級コンサルタント
専門は業務改革、自動化支援
focus@nri.co.jp