NTT DATA
株式会社NTTデータ数理システム

株式会社NTTデータ 様

電子カルテから新たな“価値”を抽出するテキスト処理システムを開発

ビッグデータの中でも、法整備により今後の活用が期待されている医療関連データ。そのデータを利活用するプロジェクトの1つである「千年カルテ」に携わっている株式会社NTTデータ様は、電子カルテに蓄積された医療情報を抽出するためのテキスト処理システムをNTTデータ数理システムと共同開発した。

製造ITイノベーション事業本部
第四製造事業部 課長代理 MBA
長谷川 義行 様

Interview

法律に基づいた匿名加工医療情報を提供

「千年カルテ」とは、どのような取り組みなのですか。

長谷川 医薬品開発などで医療系ビッグデータの活用が進められていますが、医療データは個人情報を多く含むため、活用にあたっては厳しい制限があります。こうした課題に対し次世代医療基盤法が2018年に施行され、国の認定を受けた事業者は医療機関等から提供された医療情報を、個人を識別できないようにしたうえで匿名加工医療情報として提供できるようになりました。千年カルテはその仕組みにもとづいたプロジェクトの1つで、一般社団法人ライフデータイニシアティブが推進する医療情報連携基盤です。当社はこのシステムの技術開発・運用を行っており、私はそのサービス企画と開発を担当しています。

患者情報抽出技術のために新しいテキスト処理システムを開発したそうですね。

長谷川 千年カルテでは、レセプトデータなどから得られた診療行為の実施情報や、電子カルテ情報をもとに診療行為がどのような結果をもたらしたかという臨床アウトカムをご提供してきました。しかし、電子カルテのテキスト情報は分析が困難で、ご提供できるデータが限られていたのです。今回、医療情報抽出のためのテキスト処理システムをNTTデータ数理システムと共同で開発したことで、より詳細な臨床アウトカムのご提供が可能になりました。これにより、医薬品開発やオーダーメイド医療などにおいてエビデンスを早期発見できるようになると期待されています。

電子カルテテキストを処理する言語モデルイメージ

	ルールベース（正規表現マッチ）	機械学習（系列ラベリング）
概要	● 人手でルールを書いて処理する	● 教師データを用意し、そこから正解を導くルールを導出する
特徴	● 処理がわかりやすい 新たなルールに弱い（メンテが必要）	● 総合的に判断するため、表記ゆれや新たなルールにある程度対応できる

患者情報を抽出する手法として、特徴の異なる「ルールベース（正規表現マッチ）」「機械学習（系列ラベリング）」の2種類の手法について実現性・有効性を評価する。

PROFILE

株式会社NTTデータ様

「将来にわたるビジネス革新を、技術の活用により、ともに実現するパートナーになる」という理念のもと、ビッグデータ・BI、AI、ブロックチェーンなど最新技術の活用・応用によるソリューションやシステム構築を展開。新しい社会とその実現に取り組んでいる。1988年設立、連結売上高2兆2,668億円、従業員数1万1,515名（単独）。

※データは2020年3月末現在

［裏面に続く］

なぜ、NTTデータ数理システムに依頼したのですか。

長谷川　医師が電子カルテに入力したテキストデータを活用するには自然言語処理を行い構造化する必要がありますが、その処理が非常に難しい。どのような治療や投薬を行った結果、患者の症状がどうなったか、医学用語、医薬品名とその量、検査項目、患者の状態、医師の所見などが羅列されていて、そこには文法やルールは存在しません。通常の日本語文章とは異なる特徴をもつため、一般的なツールではとうてい処理できないと判断しました。そこで、**Text Mining Studio**（以下、**TMS**）の開発などで自然言語処理技術に定評のあるNTTデータ数理システムに相談したのです。

どのように開発を進めたのでしょうか。

長谷川　まず複雑に羅列されたテキスト情報を形態素解析し、特定の診療や検査に関連するテキストをセットで抜き出す工程が必要となります。それを行うためにはどんな方法や技術があるか、NTTデータ数理システムの技術者とディスカッションするところから始めました。ルールベースによる処理に加え、機械学習ベースでの分析も必要であると判断しましたが、本件はまったく新しい取り組みのため、機械学習させるにも教師データがありません。そこでルールベースで処理した結果を教師データとして機械学習し、テキスト処理システムを作りました。

今回、テキスト情報として用いたのは、千年カルテプロジェクトに参画されている宮崎大学医学部様が実際に使用されている電子カルテです。難しかったのは、ある時期を境に電子カルテのデータ構造が変わっていたり、医師や診療科によって入力スタイルが異なったりしていたこと。それらの問題に対して一つひとつ調整や修正を重ね、処理の精度を高めていきました。最終的に、電子カルテ上のデータの9割は狙い通りの処理ができるまでになっています。開発はアジャイル的に行い、9カ月程度で完了しました。

このシステムはPythonによるフルスクラッチで構築し、千年カルテにおいて非構造化データを利活用するためのベース技術として利用しています。

期待通りの自然言語処理技術で臨床アウトカムを的確に抽出

NTTデータ数理システムのパフォーマンスはいかがでしたか。

長谷川　技術力がある人たちが揃っていますね。どの技術者も専門性が光っていて、部門・役職問わず、非常に高いレベルで対応してくれます。千年カルテのように、他の誰もがまだ作ったことのないようなシステムでは、各機能の構築に技術力はもちろん、経験に基づいた知見や応用力が必要になってくるので、彼らの力は欠かせません。今後も千年カルテの取り組みを進めるにあたってはさまざまな課題や問題が生じると思いますが、その際のパートナーとしてもファーストチョイスになるでしょう。またNTTデータ数理システムは、純国産のテキストマイニングツール**TMS**を開発して以来、地道に改良を重ね、技術やノウハウを豊富に蓄積し続けています。私はそこに期待して今回の案件を依頼したところ、見事に期待に応えてくれました。言葉の表記揺れや細かな部分のまとめ上げなど、すみずみにわたって**TMS**のノウハウが活かされていると感じています。さらに、今回の案件を通して我々のスタッフの分析力やプログラミングレベルが一段と向上しました。NTTデータ数理システムの技術者が我々に丁寧にレクチャーし、丁寧にサポートしてくれたおかげだと思っており、その面からも感謝しています。

これまでの評価と今後の展望をいただけますか。

長谷川　このテキスト処理システム開発で、膨大な電子カルテ情報の中から臨床アウトカムを的確に抽出できるようになりました。宮崎大学医学部の教授からは、電子カルテデータの新しい活用方法として高い評価をいただいています。

今回の案件では、肺がん患者の遺伝子検査結果という領域に絞って処理アルゴリズムを開発しましたが、これを活用することで他の疾病や研究テーマにも横展開が可能となります。実際、我々の部内で乳がんを対象としたテキスト処理システムを試作しています。お客様である医療機関や製薬会社にご提供できるデータの種類や量を増やし、千年カルテのユーザーを開拓していくと同時に、データをご提供いただける医療機関も増やしていきたい。それにより次世代の社会インフラとして、千年カルテの価値をいっそう高めていきたいと考えています。

テキスト処理システム開発に協力させていただいた
NTTデータ数理システム TMS開発リーダーの古賀久芳（右）

NTT DATA　株式会社NTTデータ数理システム

営業担当

〒160-0016 東京都新宿区信濃町35番地 信濃町煉瓦館1階　　TEL 03-3358-6681（お問い合わせ：平日10:00-16:00）

〈e-mail〉sales@ml.msi.co.jp　〈URL〉https://www.msi.co.jp/solution/analytics/index.html

045-202307