講演録

日経記事テキストに対する自然言語処理を用いた情報抽出とグラフ構造化

株式会社日本経済新聞社 日経イノベーション・ラボ　**安井 雄一郎** 氏

[Profile] 2016年12月まで、コンピュータに適した（ハードウェアを考慮した）並列アルゴリズムの設計と実装に関する研究に従事。2017年1月、日経BP データサイエンティスト。2019年4月より日本経済新聞社 日経イノベーション・ラボ 上級研究員。自然言語処理を用いた情報抽出、セマンティックウェブ技術、グラフデータベースを用いたデータ活用などに取り組む。

本日は、①日経が販売している記事データについて、②日経記事データを**Text Mining Studio**（以下、**TMS**）に適用した事例について、③私が取り組んでいる研究テーマである日経記事データの構造化についてお話しします。

メタ情報を付与した日経記事データ

日経では新聞記事データを日経コーパス*として販売しています。2020年9月からは、AWS Data Exchangeで分析用のデータが購入可能になりました。この記事データは機械学習やテキストマイニングなどの分析に使うことを想定し、豊富なメタ情報が付与されています。具体的には、テーマ、業界、地域、記事種別といった分類語や、人物名、企業名・企業コードといったキーワードが付与され、CSVデータで提供されます。販売するのは1981年10月以降の記事データ。記事データ以外にPOSデータも購入可能です。

＊ https://nkbb.nikkei.co.jp/nikkei-corpus/

日経記事データをTMSで分析

TMSは手間のかかるテキストマイニングが簡単に行えるツールです。分かち書き処理からネットワーク分析までさまざまな機能をGUI操作で実行します。私も使ってみて、典型的な分析を、評価軸を変えて試行錯誤しながら行うのにとても有用だと感じました。自由度が高くて使いやすいです。

このTMSを使い、日経記事データを分析してみたのでご紹介します。用いたデータは2020年1月～6月の日本経済新聞朝刊、一面に掲載された1万弱の記事です。

まず係り受けの関係が時系列でどう変化したかを見ました。上昇傾向にあった係り受けをプロットしたところ、「時間-かかる」が特に大きな領域を占めていました。そして同じ時期に「新型コロナ-対応」「検査-受ける」という係り受けが見られることから、「時間-かかる」は新型コロナに対する対応を指しているのではないかと推測できます。他にも「政府-まとめる」「経済対策-盛り込む」「影響力-強める」など、新型コロナに関係すると思われる係り受けがきれいに抽出されました（図1）。

共起ネットワーク分析も行いました。フィルター条件を設定しなかったにも関わらず、新型コロナへの集中が確認できました。国内での感染がまだ少なかった1～2月と、緊急事態宣言が議論された3～4月のそれぞれで見たところ、1～2月の段階でも小さいながら新型コロナのまとまりがすでに見られます。それが3～4月になると、とても大きなまとまりに変わったのが一目瞭然です（図2）。

このようにTMSを活用すると、あるトピックに対して何がどう変化したか効率的に分析できることを、今回の検証を通して確認できました。

図1　TMSで解析した係り受けの時系列変化

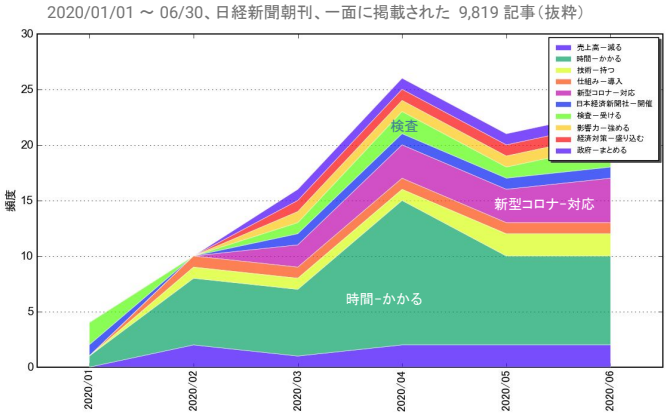


図2　TMSで解析した共起ネットワーク

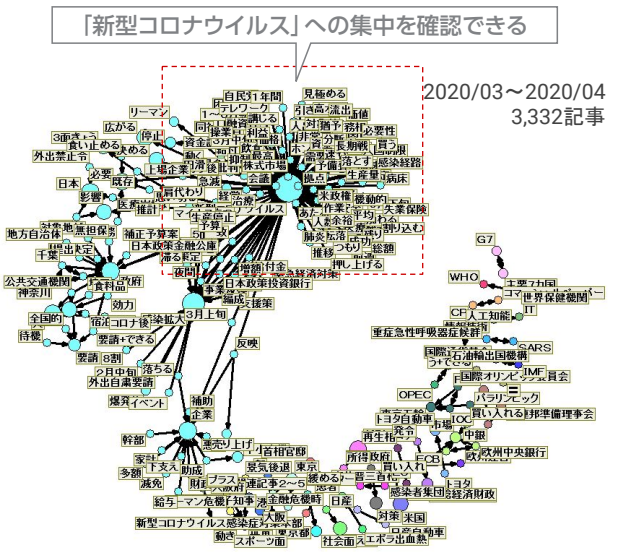
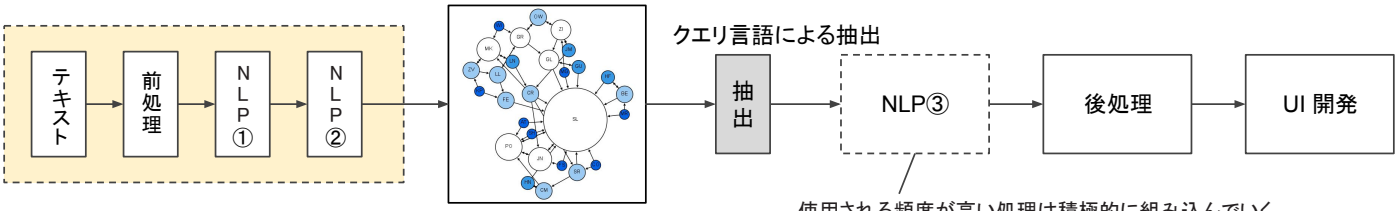


図3　処理工程の一部を構造化



*NLP：自然言語処理

日経記事データの構造化の取り組み

私は現在、自然言語処理（NLP）を用いた技術検証の効率化に関する研究に取り組んでいます。技術検証にはたくさんの煩雑な処理工程が必要で、定型化しづらいという難点があります。そこで処理工程を、典型的な処理と個別処理が必要な部分に分けて、典型的な処理の部分を構造化することを考えました（図3）。この構造化についてご紹介します。

●グラフ構造化の概要

記事は階層構造を持っています。メディア→面（ページ）→記事→パラグラフ（段落）→文章→単語という階層です。この階層構造に自然言語処理といった情報抽出を合わせ、ノードとエッジでテキストデータをグラフ構造化します。これにより、「特定の単語を含む記事」「特定の特徴を含む記事」「特定の単語と特徴を含むパラグラフ」といったように、構造を指定した検索が可能になります。ちなみに今回、文章から単語を抽出するのに固有表現抽出と係り受け解析を用いました。

こうしていくつかの処理を行ったところ、1記事あたり922ノード／1,351エッジのグラフ構造が抽出されました（図4）。半年分の記事データだと1,000万ノード／数千万エッジにのぼります。これらのグラフ構造を「Neo4j」というグラフデータベースに蓄積し、クエリ言語で抽出して利活用します。

●グラフ構造化を用いた実験

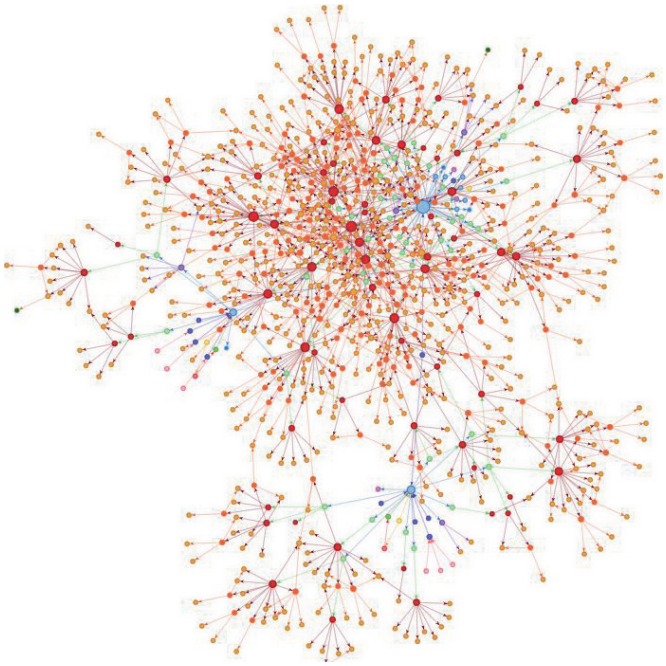
このグラフデータベースを用いた実験について紹介します。用いたデータは2020年1月～8月の日本経済新聞朝刊。粒度は月次集計です。

まず、設定をパラグラフでの共起と新型コロナとの共起とし、単語出現頻度の時系列変化を固有表現ごとに抽出。すると「地域」では「中国」が2～3月にピーク、「日本」「米国」はそれより後、3～5月にピークを迎えていました。「組織」では「政府」や「厚生労働省」の出現が多数でした。「役職」では「首相」が多いと予想していましたが、「社長」のほうが多く、ピークは5月でした。この頃、企業はさまざまな判断を迫られており、こうした記事が増えたのでしょう。

ほかにも「金融」「機械・エレクトロニクス」「資源・エネルギー」など業種ごとに単語の出現頻度を集計。それぞれに新型コロナとの関係や、それに伴う業界の動きが浮かび上がりました。

次に係り受けの時系列変化も解析しました。在宅勤務が普及していく様子を見るために、「在宅勤務」と「広がる」を係り受け構造に指定し、該当する記事を抽出しました。すると2月には1記事だけだったのが、3月5記事、4月15記事、5月20記事と増加。ところが6月は9記事に減りました。これは「広がる」状態が進んだことで他の単語に置き替わり、別の係り受け構造へ移行したと推測できます。

図4　1記事から抽出されたグラフ構造



922ノード、1351エッジのグラフ構造

日経メディア・情報サービスへの適用

このグラフデータベースはさまざまに利活用できると考えています。例えば、メディア（電子版）のバックエンドとして、あるいは情報サービスの向上に、また記者へのフィードバックにも適用できるでしょう。日経は新聞以外にも財務データや企業情報データを持っているので、それらを構造化し連携させれば、よりよい情報発信が可能になります。私自身の大きな目標は、社内外のデータ・情報からグラフ構造を抽出して蓄積・成長させていく、“日経ナレッジグラフ”ともいうべきプラットフォームをつくること。それを通してユーザーの皆様へよりよいサービスを提供していきたいです。

NTT DATA　株式会社 NTTデータ 数理システム
お問い合わせ：平日10:00-17:00（e-mail、FAXは随時受付）
Text Mining Studio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　FAX 03-3358-1727
〈e-mail〉vmstudio-info@ml.msi.co.jp　〈URL〉https://www.msi.co.jp/tmstudio/