自然言語処理を活用した化学・材料開発

◆化学反応を自動で分類し可視化するAI

2021年1月、IBMは、既知の化学反応を自動で分類し、分類結果を可視化するAI（人工知能）を開発した。膨大な化学反応のデータをマッピングすることで、希望する化学反応に類似する反応を容易に探索でき、また研究者の先入観にとらわれない反応を発見できるなど、新薬や新材料の開発が効率化されるとしている。

自然言語処理技術を使って化学反応の分類を自動化したことが、新たに開発されたAIの特徴である。従来の分類では、反応フィンガープリントと呼ばれる、反応に含まれる分子数や、反応の中心となる部位、反応物と試薬の区別を事前にラベリングする必要があり、前処理に時間を要していた。これに対し新手法では、反応物と生成物の情報を、SMILESと呼ばれる化学構造を文字列化する手法で表し、かつ事前のラベリングがない状態で、文章の類似性判定の精度が高いBERT [1]をベースとした自然言語処理モデルで分類した。結果、分類精度は、既存のラベル付き分類モデルと比較して98.9%に達し、ラベリングがない反応データでも自動分類が可能になった。

また、得られた分類結果は、従来よりも反応の詳細な違いを表現した反応フィンガープリントとして使用できることができ、反応を適切かつ精緻にクラスタリングし、マッピングすることができた。

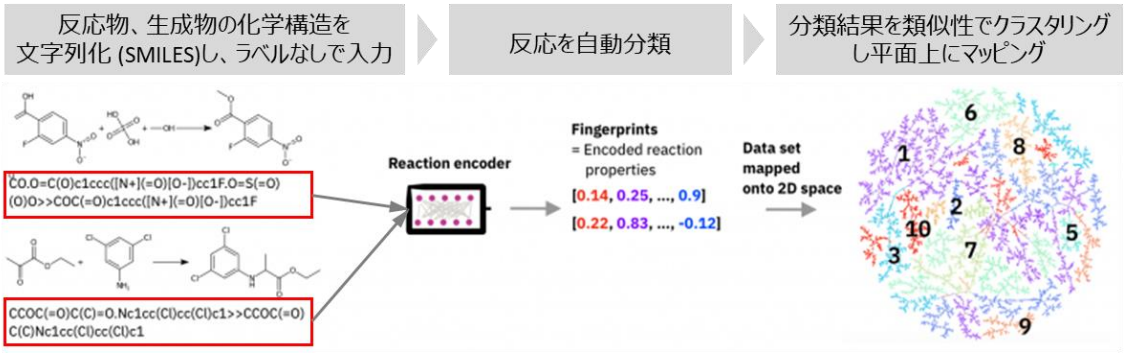

図　自然言語処理による化学反応の分類と可視化　出所)IBMホームページ情報をもとに一部ARC追記

[1] 2018年10月にGoogleが発表した自然言語処理モデル。文章の類似性判定や文法の妥当性判定、感情分析などの精度が高く、様々なタスクへの応用が進んでいる。

◆自然言語処理技術で材料物性や分子の動きを予測する

BERTに代表される、文脈などの情報の関係性を適切に解釈できる自然言語処理技術が発達したことで、これを科学技術の分野に適応する動きが広がっている。21年1月、清華大学の研究チームは、自然言語処理によって低分子材料の構造と物性の関係を予測する手法を開発した。研究チームは、音声認識や天気予報などの時系列情報に対して高い予測精度を示す、LSTM(Long short-term memory)を利用した。低分子材料のSMILESを入力値としたLSTMベースの予測モデルによって、有機材料の融点を予測でき、既存技術と同等以上の精度を示した。従来のルールベースや記述子ベースのモデルとは異なり、分子量、分岐の数、結合数の変化に対しても、精度よく予測できる。このモデルを応用することで、化学構造から各種物性値を予測できるとしている。

同じく1月、メリーランド大学の研究チームは、自然言語処理を使用して、分子動力学のシミュレーションと同等の予測ができると発表した。研究チームはLSTMベースの自然言語処理を用いて既存のシミュレーション結果を学習させ、別の分子動力学シミュレーションの結果を予測した。その結果、リゾチームタンパク質へのベンゼンの結合に重要なアミノ酸を予測することに成功した。

◆文章資産からナレッジグラフを作成し、開発に活用する

21年2月、ボッシュはナレッジグラフに関するワークショップを開催した。ナレッジグラフとは、情報を分類・要約し、可視化する技術であり、ボッシュは材料開発をサポートするシステムとして、ナレッジグラフを開発している。出版物や特許から自然言語処理を用いて抽出した4万点の資料を基に、ナレッジグラフが構築されており、材料の検索や、複雑な質問応答タスクに使用できる。

20年11月、長瀬産業は、ナレッジグラフを作成する材料探索プラットフォーム「TABRASA」を開発した。特許技術、論文、ビジネス文書などの複数の非構造化データから、自然言語処理によって情報を抽出し、構造化する。体系化した情報から新たな知見を見いだすことでき、固定観念にとらわれない物質の合成方法の発見や、開発期間の大幅な短縮が可能になるとしている。

情報の関係性をとらえる自然言語処理技術は、さまざまな課題を解決する手段になりうる。文章処理にとどまらず多方面での応用が期待される。【塚原祐介】