エッセイ

自然言語処理「言葉の意味を表す技術」のブレークスルー

古宮嘉那子

こみや · かなこ。東京農工大学大学院工学研究院先端情報科学部門准教授。東京農工大学卒業。同大学大学院で修士号および博士号（工学）を取得。東京工業大学博士研究員、東京農工大学特任助教、茨城大学講師を経て、2021年より現職。専門は自然言語処理。著書に『機械学習教本』（森北出版）、『文書分類からはじめる自然言語処理入門』（科学情報出版）がある。

ChatGPT が登場してから 1 年余り。AI が質問に流ちょうな文章で答えてくれることが当たり前になりました。このようなことが可能になったのは、自然言語処理、特に、言葉の意味をコンピュータ上に表現する技術が急速に進歩してきたからにほかなりません。

自然言語処理とは、AI 研究の一分野で、日常私たちが使っている言語をコンピュータで扱う研究分野です。この分野において、ディープラーニングによる技術革新が起きたのは 2013 年から 2018 年のほんの 5、6 年のことでした。これは、言葉の意味を表す技術のブレークスルーだったと言えます。

●

自然言語処理では、主に言語学における分布意味論の「分布仮説」によって単語の意味を取り扱っています。分布仮説は、簡単に言えば「文脈の似た単語・句・文などは意味が似ている」ということ。ある単語を表す際、「one-hot ベクトル」と呼ばれる方法が用いられます。これは、語彙数分の要素を持つベクトル（数値の配列）を用意し、その単語に該当する要素だけを 1、その他の全ての要素を 0 とすることで単語を表したものです。ところが、one-hot ベクトルだけでは、単語の意味的な類似性を直接的には表現できません。例えば、「猫」と「犬」の意味的な類似性は「猫」と「ディープラーニング」の類似性より高そうですが、ベクトルの要素として単に「同じ単語であるかどうか」で考えると、共に「互いに異なっている」ということまでしか表せません。

そこで、分布仮説の出番です。意味を表したい単語の周辺の単語の頻度情報を「文脈」としてとらえるために、例えば周囲の one-hot ベクトルをいくつか連結したものを 1 つのベクトルとして単語を表すことにします。しかし、ベクトルの要素数がとても大きくなるわりに、周辺文脈に出現した単語以外の要素の値は 0 になるため、ほとんどの要素が 0 になってしまい、効率的に意味を表現できないという問題があります。

●

この問題に解を与えたのが、2013 年に Mikolov らの提唱した「Word2Vec」です。Word2Vec は、0 が多く疎であるベクトルの要素数（次元数）を圧縮する工夫を加えたものです。この際、ベクトルの要素数はもともとの語彙数よりはるかに小さくなります（語彙数は万単位ですが、Word2Vec のベクトルの次元数は通常たかだか 200 から 300 程度です）。

ここにディープラーニングの技術が用いられました。ディープラーニングは、次元数の大きい情報を圧縮することが非常に得意です。このように単語を低次元で 0 の要素が少ない密なベクトルで表すことを、「単語の分散表現」と呼びます。さらに、Word2Vec による分散表現には、言語学における「加法構成性」があるという特徴があります。これは、例えば「king − man + woman = queen」などの意味計算が可能だということです。

Word2Vec に代表される、語ごとの分散表現では、単語の見た目が同じであれば、文脈が異なっても同じベクトルとして表されます。これで困るのは、複数の意味を持つ単語です。例えば、「マウス」はコンピュータ・デバイスであると同時に動物の名前でもあります。でも、Word2Vec ではこれらを同時に 1 つのベクトルとして表すため、2 つの意味を区別できません。

●

2018 年に Peters らによって提唱された「ELMo」（Embeddings from Language Models）は、それを可能にしました。見た目が同じ単語でも、文脈が異なれば別の意味ベクトルとして表せるようになったのです。

さらに 2018 年に、ChatGPT の前身である初代 GPT（Generative Pre-trained Transformer）が発表されました。これで、Transformer という高速計算が可能なネットワーク構造を持つモデルで事前に大量の文書を用いて言語構造を学習させておき、タスクごとにモデルの Fine-tuning（微調整）を行って利用する「大規模言語モデル」（Large Language Model：LLM）の枠組みが確立しました。現在の自然言語処理では、この大規模言語モデルを利用する手法が主流です。

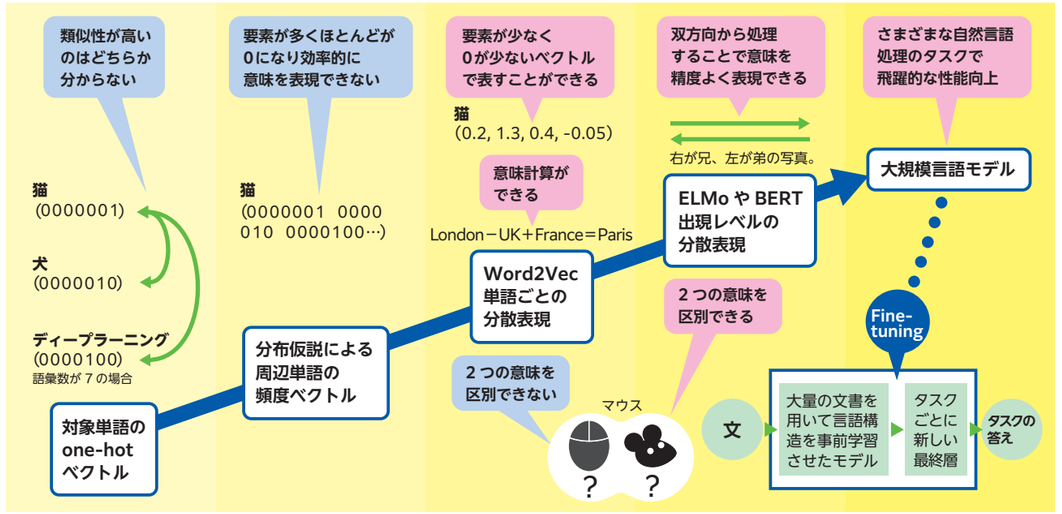
さらに、その年の秋に発表された「BERT」（Bidirectional Encoder Representation from Transformer）は、文脈を前方向からも後ろ方向からも処理する双方向の Transformer の利用と言語モデルの工夫で、さまざまな自然言語処理のタスクで飛躍的な性能向上を成し遂げました。また、BERT による分散表現は、日本語学や言語学の観点からも有益です。

●

筆者は学生時代から文中の単語を辞書の意味項目に分類する研究を行ってきましたが、上記のような一連の技術の発展に伴って精度が格段に上がってきています。例えば筆者の研究室では、日本語の BERT を使って「現代日本語書き言葉均衡コーパス（BCCWJ）」に出現する語に『分類語彙表－増補改訂版』の意味番号を振りました。文章のタイプによりますが 81.0 ～ 93.8% の正解率が得られています*。

以上のように、大規模言語モデルの発展により、自然言語処理技術は一定の円熟を見せ、応用的にも学術的にも非常に有用であることが明らかになり、研究者がモデルを自作する時代は終わりを迎えつつあります。筆者は、今こそ日本語学や言語学などの文系学問と、大規模なデータに立脚する自然言語処理とが、共に発展していくときだと考えています。これからの自然言語処理の発展に大いに期待しています。

自然言語処理による意味の表現形式の変遷



＊　浅田宗磨 · 古宮嘉那子 · 浅原正幸（2024）「『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号悉皆付与」『言語処理学会第 30 回年次大会発表論文集』