自然言語処理を利用した実証的な認知言語学の可能性

永田亮、高村大也

1. はじめに

本稿では、自然言語処理（以下、**言語処理**と省略）の変遷を紹介しつつ、言語処理と認知言語学を融合した実証的な研究の可能性を探る。過去から現在に渡って言語処理を概観すると共に、近年のブレークスルーを紹介する。このブレークスルーで培われた言語処理技術が言語分析のための強力なツールとなり得ることを実例を通じて紹介する。

2. 言語処理の変遷

言語処理の歴史は大きく次の三つの時代に分けることができる：ルールベース時代、データドリブン時代、エンドトゥーエンド時代[i]。

1990 年代後半までは、人手で作成したルールに基づいて言語処理を行うことが主流であった（ルールベース時代）。ルールは、言語学の知見や内省に基づいて作成する。例えば、与えられた文の品詞を推定する処理、品詞解析では、冠詞の後に名詞が出現しやすいなどの知見をプログラムとしてコーディングする。このアプローチは、言語学の知見を解析に利用できるという利点があるが、実用に耐えうる大規模なルール体系を作ることが難しい。例えば、上述のような品詞連接ルールを作成し、維持することには多大な労力を要する[ii]。

この問題を解決するために、2000 年頃から、データ駆動型の処理が主流になる（データドリブン時代）。この時代からは、機械学習[iii]と呼ばれる技術を利用して解析のためのルールを得る。代わりに、人間は与えられたデータに情報を付与するという作業を担当する（この情報付きデータを機械学習では**訓練データ**と呼ぶ）。例えば、品詞解析の場合、コーパスに品詞情報を付与する。訓練データを機械学習のプログラムに与えると、解析のための規則が自動的に作成される。このアプローチにより、様々な言語処理のタスクで性能が改善した。

その後、2013 年ごろより、エンドトゥーエンドとよばれる時代に入る。このアプローチでは、これまでの時代と異なり、途中の解析を伴わず所望の出力を得る。例えば、日英翻訳の場合、以前のアプローチでは、品詞解析、構文解析などの処理を入力文に順次適用し、その結果に基づいて翻訳処理を行う。一方で、新しいアプローチでは、日本語文を与えると途中の解析は行われず、対応する英文が出力される[iv]。このアプローチにより、今まで非常に難しかった言語処理の問題が次第に解決されつつある。その中心には機械学習の一種である**深層学習**の利用がある。

3. 深層学習と単語ベクトル

様々な分野で深層学習はめざましい成功を収めている。言語処理も例外でなく、自動翻訳、対話システム、文法誤り訂正など様々なタスクで大幅に性能が向上した。深層学習は大量のデータを必要とするが、インターネットの発達とともに、大量の言語データ（例えば、web コーパス）が利用可能になったことが追い風となったことは間違いない。

ではあるが、恐らく、最大の成功要因は単語の取り扱い方法の改善である。従来、単語はコンピュータ内で記号として扱われていた。そのため、ある単語が別の単語と異なるということは表せても、よく似ているとか、少し似ているというようなことは基本的に表せない。深層学習が使われ始めると、単語は数値列として扱われるようになる[v]。具体的には、深層学習を用いて、意味的、用法的に似た単語に、似た数値列を割り当てる[vi]。これにより、Japan は、apple より France に似ているということを数値として定量的に表すことが可能となる。なお、数値列をベクトルと呼ぶことがあり、単語に割り当てた数値列ということで**単語ベクトル**[vii]と呼ぶことがある。また、文や文書を数値列として表した文ベクトルや文書ベクトルも存在する。

4. 単語ベクトルを利用した言語分析

上述のとおり、単語ベクトルにより、単語の類似度を定量的に表すことができる。この類似度は、言語分析のための強力なツールとなり得る。文献 [高村大也、永田亮、川崎義史, 2017]では、単語ベクトルの類似度に基づいて、カタカナ語と英単語間の意味のずれを定量化した。日本語コーパス、英語コーパスそれぞれから、単語ベクトルを得て、対応する単語間（例：イメージと image）の類似度に基づいて意味のズレを定量化した。その結果、イメージ-image、プレゼント-present などが意味のズレの大きいペアとして特定された。また、歴史

コーパスから得られた単語ベクトルを通じて、単語の意味変化を捉える研究 [William L. Hamilton, 2016]がある。同様のアプローチを用いて、ラテン語からロマンス諸語への意味変化を分析した研究 [川崎義史, Maelys Salingre、Marzena Karpinska、高村大也、永田亮, 2022]もある。

上の研究では、単語タイプにつき一種類の単語ベクトルが割り当てられる。一方、単語トークンに対して単語ベクトルを割り当てることも可能である。すなわち、周辺単語を考慮して文中の単語トークンにベクトルを割り当てる[viii]。こうすることにより、文中の各トークン同士の類似度を定量化できる。文献 [永田亮、大谷直輝、高村大也、川崎義史, 2022]では、この方法に基づいて、better off の用法を自動的にグルーピングし、各用法の年代変化を可視化したところ、認知言語学で知られる仮説 [大谷直輝, 2020]に大枠一致することを示した。

5. 言語処理を用いた実証的な認知言語学の可能性

ごく一部であるが、言語処理を利用した言語分析の事例を、単語ベクトルに基づいたものを中心にして紹介した。これらの例に示されるように、単語ベクトル（や文ベクトル、文書ベクトル）は用例の柔軟なグルーピングや検索を可能にし、仮説の生成や検証をサポートするためのツールとなる。

ここで強調しておきたいことに、単語ベクトルをはじめとする各種ベクトルは、raw コーパスのみから得られるということである。すなわち、コーパス中の用例から単語ベクトルなどは得られる。このことは、認知言語学における用法基盤モデルの考え方に通ずるものがあり、認知言語学と言語処理の一つの接点となる可能性を秘めている。今後、単語ベクトルや文ベクトルが言語分析に、より一層利用されると期待される。

更には、単語ベクトルを基本とした各種生成技術も言語分析に有益となる可能性がある。文章から画像を生成する技術（**text-to-image generation**）を用いると文章が表す内容を画像として可視化できる。この技術を利用して、上り坂／下り坂のような **construal** の問題に新たな切り口が見出せるかもしれない。また、画像から言語を生成する技術（**image caption generation**）に使用される深層学習の内部状態と人が画像内容を言語化した際の脳活動との関係を調査する研究もあり、今後の発展が期待される。

6. おわりに

本稿では、言語処理の変遷を紹介しつつ、実証的な認知言語学の可能性を模索した。現代の言語処理では、単語ベクトルが重要な役割を果たすことを述べ、単語ベクトルに基づいた言語分析の研究を紹介した。単語ベクトルは用法基盤モデルと親和性が高く、認知言語学のための一つの強力なツールとなる可能性を秘めていることを述べた。

参照文献

川崎義史、Maelys Salingre、Marzena Karpinska、高村大也、永田亮 (2022). 分散表現を用いたロマンス語同源語動詞の意味変化の分析. *言語処理学会第 28 回年次大会発表論文集*, (p.p 1861-1866).

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (p.p. 1489-1501).

永田亮、大谷直輝、高村大也、川崎義史. (2022). 言語処理的アプローチによる better off 構文の定着過程の説明. *言語処理学会 第 28 回年次大会発表論文集*, (pp. 240-245).

高村大也. (2010). *言語処理のための機械学習入門*. コロナ社.

高村大也、永田亮、川崎義史. (2017). 外来語の意味変化に対する数理的分析. *言語処理学会第 23 回年次大会*, (pp. 907-910).

石井雄隆、石岡恒憲、金田拓、小島ますみ、小林雄一郎、近藤悠介、永田亮. (2020). *英語教育における自動採点*. ひつじ書房.

大谷直輝. (2022). better off 構文の定着過程に関する認知言語学的考察. *言語処理学会第 28 回年次大会発表論文集* , (pp. 235-239).

---

[i] 一つの分類に過ぎず、他の分類法もあり得る。また、三時代が明確に分かれるわけではない。

[ii] 例えば、品詞数が 100 の場合、二品詞の組み合わせは 10,000 にもなり、その全てについて連接しやすさを数値化することになる。

[iii] 機械学習については、文献 [高村大也, 2010]を入門書として挙げておく。

[iv] 品詞や構文などが完全になくなったわけでなく、処理の過程で明示的に現れなくなったということである。必要に応じて、品詞や構文の情報を出力することは可能である。

[v] 厳密には、以前の時代でも同じようなアプローチがとられることもあった。本格的に使用されるようになるのは近年からである。

[vi] 紙面の関係から、手法の詳細は割愛する。文献 [石井雄隆、石岡恒憲、金田拓、小島ますみ、小林雄一郎、近藤悠介、永田亮, 2020]5 章などを参照されたい。

[vii] このほか、単語分散表現という用語も使われる。

[viii] そのような手法に BERT や GPT-3 などがある。