Profile
情報アーキテクチャ分野
教授 ルパージュ イヴ
博士（工学）（グルノーブル大学）

研究分野
・機械翻訳
・自然言語処理
・用例知識取得
・相違と類似・類推関係

Web https://www.waseda.jp/fsci/gips/other/2015/07/15/303/

IPS 教員インタビュー

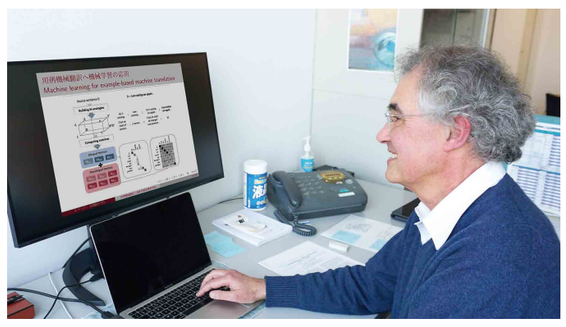
より人間的で表現力豊かな機械翻訳を実現させるために

スマートフォンやWEBブラウザに実装されたことで、「機械翻訳」が身近なものになってきた。しかし、身近になったからこそ、ニュアンスが微妙に異なる翻訳結果を目にしたり、翻訳できる語彙の限界などに気づいたりする機会も増えてきたのではないだろうか。話者が少なくデータ量が乏しいマイノリティ言語や、曖昧なニュアンスの言語表現は、機械翻訳が苦手とする分野だ。そうした中で大学院情報生産システム研究科のルパージュ・イヴ教授は、「類推関係」に基づく自然言語処理にターゲットを当て、機械翻訳ならではの弱点を補完しうる研究を進めている。

様々な分野の研究によって高度化する機械翻訳

近年になり、急速に普及・進歩してきた機械翻訳技術だが、そのルーツは意外に古い。開発構想が生まれたのは100年近く前に遡り、第二次大戦の頃にはデモ機が作成されている。その後、1970年代には、原言語（元の言語）と目的言語（翻訳すべき言語）との形態論・統語論のルールを設定し、これを機能させる「RBMT（ルールベース機械翻訳）」が開発された。ただ、文法ルールの入力は言語学者らによる手作業で、単語の設定にも大量の辞書が必要だったことから、80年代にはルール設定の作業を省き、文章例を軸にする「EBMT（用例ベース機械翻訳）」が開発され、その後2000年代に統計的機械翻訳（SMT）は開発されました。

このEBMTを開発・提案したのは、その後、京都大学の総長に就く長尾真京都大学名誉教授で、ルパージュ教授が来日するきっかけとなったのも、同氏の研究内容に強い関心を持ち、同ジャンルである「類推関係」に基づく機械翻訳の研究に注力するようになったからだという。「現在の機械翻訳は、LLM（大規模言語モデル）を活用するのが主流になっていますが、私はこれまで、翻訳アルゴリズムに関するプログラミングの研究や、確率モデルを作って自動的にデータ収集する『統計的機械翻訳』の研究を行ってきました。長尾教授がEBMTを開発されたことで、類推関係の研究に取り組むようになりましたが、これは、機械翻訳のベースとなる自然言語処理において、重要な位置を占めています」と、ルパージュ教授は語る。



類推関係に基づく自然言語処理とは、言葉や文を類似した言葉や文との関係性に基づいて理解したり、生成したりする技術。４つの事柄の間に成り立つ相違、いわば共通性では無く対応関係をシステムに学習させ、そこから新しい意味を持つ文章を推定したり生成したりする。「簡単な例ですが、『男性・女性』という関係性に基づけば、『夫』に対して『妻』、『王』に対しては『女王』あるいは『王妃』という語句が推定できます」、「それを私の娘に話したところ、人気のカップ麺を思い出したらしく、『じゃあ"赤・緑"だったら"狐・狸"だね』と笑っていました。他愛の無いジョークですが、類推のルールに何らかの条件を加味すれば、そういう語句の推定もあり得るわけです」。

類推関係の応用である「形態素解析」も、同教授が力を入れている研究分野だ。「『歩きます、歩きました、決めます、決めました』のように語句が変化して、同様に『walk、walked、decide』を与えられた時、ある文脈の下で『decided』という語形を生成でき、語彙の語系を正確に把握して翻訳精度を高めることが可能。語形の生成可能を確定させれば、類推関係に基づく文章の用例からスムーズな機械翻訳を行うこともできます」。

例えば、「このツアーの料金はいくらですか?」という日本語を、「How much does this tour cost?」と英訳した用例（過去事例）から類推し、同じHow muchを含む複数の構文で、「料金はいくらか」を「いくらの料金を払うか」と言い換える「How much do you charge for this tour?」が生成可能だ。単語の「cost」を「price」に置き換えた「What's the price of this tour?」という文章も生成できるだろう。ルパージュ研究室に修士学生として配属された学生たちも、類推関係について様々な角度から研究し、その内容を修士論文として積極的に提出している。

最近の修士論文題名
Titles of recent master's theses

- Exhaustive extraction of analogical clusters from word embedding spaces
- Transformer-based hierarchical attention models for solving analogies between longer and semantically richer sentences
- Compositional augmentation policy using different formulas for the notion of middle sentence for low resource machine translation
- An embedding-to-embedding method based on an auto-encoder architecture for solving sentence analogies
- A dual reinforcement method for data augmentation using middle sentences for machine translation
- Improving sentence embedding quality with sentence relationships from word analogies
- Extraction of analogies between sentences on the level of syntax using parse trees

大企業が手を出さない分野だからこそ研究のやり甲斐がある

ChatGPTに代表されるLLMの大半は、世界規模のビジネスを目指して構築・運用されている。だからこそ、莫大な資金を投じることができるわけで、逆に言うと、ビジネス上のメリットが見込みにくいマイノリティ言語に関しては、ほぼ手付かずの状態だという。「全世界には、6,000～7,000ヵ国（地域）の言語が存在します。ところが、現時点で世界最大級と言われているLLMでさえ、約200種類の言語にしか対応していません。日本語だけ見ても、琉球語やアイヌ語はカバーされていません。それら、LLMに"取り残された"言語に焦点を当てて類推関係を考察する研究は、大企業では真似できない分野と言えるでしょう」。

国内でChatGPTがリリースされた2023年秋頃、自然言語処理の研究者たちが集うフォーラムで、「ChatGPTで自然言語処理は終わるのか？」という標題のパネルディスカッションが行われ、参加者の多くは不安な表情になっていたそうだ。しかし、同教授の見解は違う。「ペシミスティックな考え方かもしれませんが、自分たちで機械翻訳システムを作ろうというのではなく、大企業が興味を示さない分野こそ、小規模な研究室にとって絶好の研究テーマと考えているのです」。その観点から、インド北東部・マニプル州で使われているマニプリ語の研究に注力。マニプリ語辞書を集めたり、現地出版社の協力を得て記事を集めたりしながら、同言語のコーパス(※)を完成させたという。

「ポーランド語、チェコ語、ロシア語など、同様のスラブ語族での言葉は動詞の活用だけではなく名詞、形容詞などの曲用も体系的な類似が見られる。これも類推関係を用いれば、語族ごとに"次に来るトーケン"を類推できる仕組みが作れるはずです」。データ量が少ないマイノリティ言語から、類推関係によって言語モデルを作成する手法や可能性について、同教授は「Continued pre-training on sentence analogies for translation with small data（小規模データでの翻訳における文の類似性を用いた継続的な事前学習）」、「Eliciting analogical reasoning from language models in retrieval-augmented translation under low-resource scenarios（低リソース環境における検索拡張翻訳において、言語モデルからアナロジー推論を引き出す」などの論文を発表。同時に、「Transformer-based hierarchical attention models for solving analogy puzzles between longer, lexically richer and semantically more diverse sentences（より長く語彙も豊富な類推関係パズルを解くための、Transformerベースの階層的モデル）」との論文では、従来のモデルよりも複雑な課題に対応できる新しいモデルについて提案している。

(※)大量の文章などを文字化し、コンピュータで処理できるよう電子化した言語資料

類推関係の、人間の学習分野への応用法も研究

同教授が手がける類推関係の研究は、語句や文章ばかりではなく図形の構造についても広がっている。対象として取り組んだのは、日本で使われている漢字と中国の漢字との類推関係。「日本人学生が、中国語の漢字をできるだけ効率的に覚えたいと言っていたのがきっかけとなり、図形としての類推関係を導き出す方法を研究しました」。

確かに日本と中国の漢字は、形状として似ているものが多い。この研究では、数値による属性のベクトルを用いた数学的な解析を用い、漢字の構成要素の中に含まれる構造を明示する手法を実践した。漢字の形状と、その発音との間の一致性や規則性を明確にすれば、言語そのものも覚えやすくなるだろう。「どんな言語にも普遍的な現象があり、個々の単語にも構造があります。まず、構造的な分析をコンピュータに行わせ、それを覚えてから文法や実用的な用例を人間が学習すれば、様々な国の言葉を習得しやすくなるでしょう。機械翻訳の話題からは少し外れますが、他国の言語を効率的に学習するためにも類推関係の研究は役立つのです」。

これは"裏話"だが、同研究室がマニプリ語のコーパスを完成させ、WEB上で公開した後、いつの間にかChatGPTがマニプリ語に対応できるようになったという。「『OpenAI』という社名とは裏腹に、実はClosedだよね…と、研究者たちが苦笑いするくらい、彼らのリソース収集元や翻訳のアルゴリズムはブラックボックス的です。ただ、自分たちの研究が盗まれたと憤るのではなく、LLMの急速な進歩に伴って生じる新たな課題は、常に私たちの研究テーマになり得るのですから、終わりのない研究領域だと考えています」。

好奇心旺盛な子どもが、お気に入りのおもちゃで遊び尽くして、さらに新しい遊び方を工夫する…。自身の研究ジャンルの魅力について、そのように語るルパージュ教授。「機械翻訳と聞いて、ICT系の研究ジャンルをイメージする学生も多いでしょうが、私たちの取り組みは言語学のウェイトの方が大きいです。義務教育で学ぶ英語以外の、他国の言語に興味を持っている学生には、是非うちの研究室に来て欲しいですね」。