生成AIを活用した用語抽出

伊澤 力
株式会社川村インターナショナル
rkizw@k-intl.co.jp

はじめに

背景

翻訳において、対訳用語集は人手翻訳における用語の統一を図るために不可欠な資料であるが、近年は機械翻訳においても用語を指定するために利用されており、その必要性は益々高まっていると言える。用語集を構築するにあたっては大量のテキストコーパスから対象となる用語を抽出し、適切な訳語を適用する作業が必要だが、そのほとんどは人間の目による確認や人手による作業が不可欠であり、実務における対訳用語集作成の課題となっている。

課題

用語抽出を自動化する方法は、これまで複数提案され実用化もされている。しかし、その多くは用語の使用における統計量をもとにしたものであり以下の点において未だその精度が実用レベルに達していないという課題が存在している。

- 特定分野のみの選択的な抽出
- 対訳での抽出

本研究では、化学分野の特許明細書の対訳コーパスをもとに、生成AIを活用することで上記課題を解決することができるかを検証し、さらに実用度を高めるためにモデルの学習を用いた精度向上の検証も実施した。

対訳コーパス

本研究では以下の対訳コーパスを用いた。

- 対象文書：　化学分野の特許明細書
- 原語ペア：　日本語、英語
- 文章数：　10,000文

文書の特徴として、化学分野の用語が多く含まれているものの特許明細書の性質上特許用語も頻出しており、従来の統計的手法による用語抽出では不要な用語候補（ノイズ）として抽出されることが予想された。

また、対訳コーパスは対応する日英の特許明細書を自動処理を用いて整合（アライン）した。そのため、**日本語と英語が正しく対応していないペアが一定数存在**する点も特徴的かつ実践的と言える。

生成AI

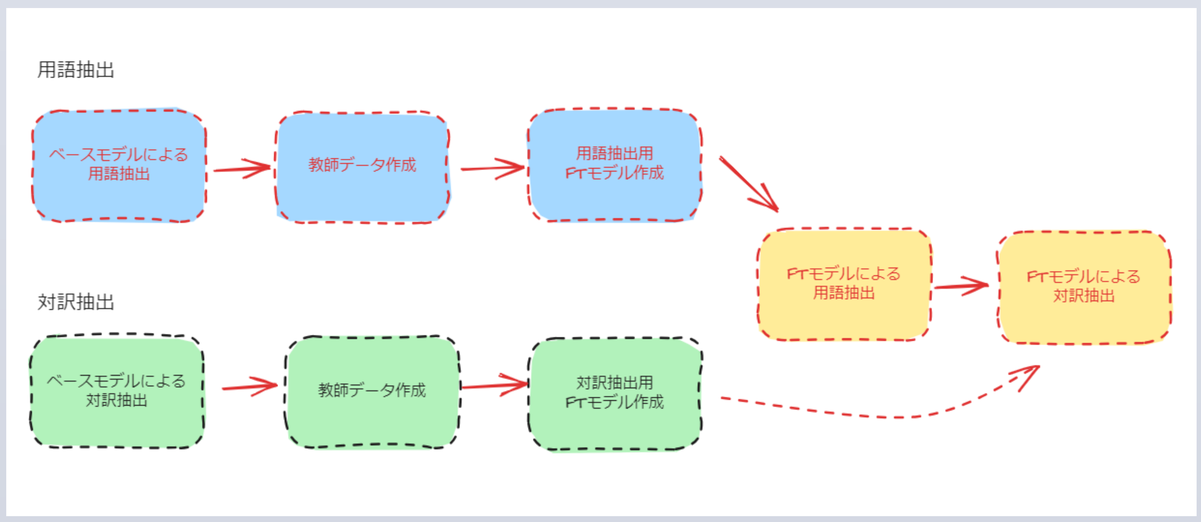
OpenAIのGPT gpt-3.5-turboを用いた。当時、GPT 4がリリースされていたが、費用に比して抽出結果に大きな差がないことから、速度と経済性の面からGPT gpt-3.5-turboがより実用的であると判断した。

プロンプトはSystem Promptで役割を定義し、User Promptには化学用語を抽出する旨を指定し、それらをすべて英語で記述した。

作業工程

本プロジェクトでは当初、特定分野の用語抽出と対訳抽出をひとつのタスクとして行う方法を模索したが、2つのタスクを一度に行ったためか精度に難があったため、各タスクを独立して行う方法（図1）を用いた。

図1. 用語抽出と対訳抽出を独立して実行するフロー



用語抽出

ベースモデルでの抽出

対訳コーパスのうち、原文となる日本語の文章を入力として化学用語を抽出する旨のプロンプトを用いたところ、選択的に化学用語を抽出することができたが、以下の問題が確認された。

- レスポンス形式の一貫性の欠如
- 用語候補の過抽出、漏れ

レスポンスの形式はプロンプトである程度制御することができるものの、用語とは関係のないモデルからの「返事」の内容が含まれていたり、複数の用語が抽出された場合の区切り文字が一定ではないといった結果が見られた。また、モデルが用語を抽出することを優先したためか、用語候補として分野外の用語を抽出するケースも多く見られ、結果としてノイズを含む多数の用語候補が抽出された。

作業工程（続き）

ファインチューニングモデルでの抽出

上記問題を解消するため、ベースとなる GPT gpt-3.5-turbo に対してファインチューニングを行った。教師データは、ベースモデルの出力からランダムに抜き出した100文を**専門知識を有する者の目で確認**し、修正を行った。また、出力後のデータの加工を効率的に行うために複数候補が抽出された場合の**区切り文字を統一**した。

ファインチューニングを施したモデルを用いてベースモデルでの出力と比較したところ、ベースモデルの出力における2点の問題がいずれも大幅に改善された。

対訳抽出

ベースモデルでの抽出

対訳抽出においても、用語抽出と同様にGPT gpt-3.5-turboを使用した。上記ファインチューニングモデルによる抽出結果に加え、対応する英文を入力としてモデルに与えた。

結果として、ベースモデルであっても概ね期待通りの結果を得ることはできたが、訳文中に対応する用語が存在しない場合に、誤った訳語を出力する傾向が見られ、この点について改善の余地があることが判った。

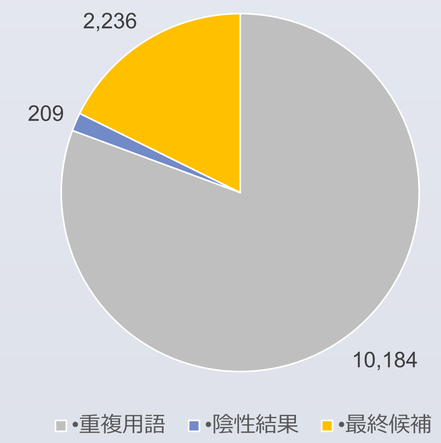
ファインチューニングモデルでの対訳抽出

用語抽出と同様に、ベースモデルの抽出結果からランダムに100文を抜き出し、期待する出力となるよう修正を行い、教師データとしてファインチューニングを実施した。このとき、ベースモデルにおける課題を解決するため、訳文中に対応する用語が存在しない場合に期待する出力結果を教師データに含め、プロンプトにも追加した。

結果

上記工程にて作業を実施した結果、12,629件の用語候補が対訳で抽出された。このうち、図2に示すとおり重複した用語10,184件および陰性結果209件を除いた**一意の用語は合計2,236件**となった。陰性結果は原文と対になる訳文側に、該当する用語が存在しないものを指す。陰性結果の発生原因は、翻訳の結果として訳文に該当する用語が訳出されなかった場合もあるが、主な原因は対訳コーパス作成における自動処理において、誤った訳文が紐づけられたことによるものであった。

図2. 対訳用語抽出結果



考察

本研究の結果から、生成AIを用いることで従来は難しかった**特定分野の用語を大量のテキストコーパスから選択的に抽出することが可能であることが確認できた**。ただし、学習を行わなかったベースモデルでは出力の安定性や精度の面で実用レベルに達していなかったため、本研究で行った**ファインチューニングを用いた出力制度の向上は必須**であると言える。

また、複雑なタスクを処理することが可能な生成AIにおいても、可能な場合はタスクをより**単純なタスクに分割して実行することが有効**であることが確認できた点も本研究の成果と言えるであろう。

もちろん、最終候補として抽出された2,236件の用語についても最終的な有用性については専門知識を持った人間、つまり実務において**最終利用者となる者による確認を経ることが理想**ではあるが、少なくとも大量の文書を前にして用語集を作成必要のある者にとって、生成AIを用いた抽出は救いの手となることは間違いないであろう。

本研究は、あくまで生成AIの活用によって従来の用語抽出の課題を解決できるかという点について、実務での活用を前提として検証を行った。そのため、従来の方法による作業との詳細な比較および品質面での評価については今後の課題とする。

今後の展望

本研究では前述の理由からgpt-3.5-turboを活用したが、本稿執筆時点でOpenAI社からGPT-4oなどのより新しいモデルがリリースされており、今後はこれらの新しいモデルにおいてより精度が向上するかを検証したい。

また、本研究ではファインチューニングにより精度が向上することを確認したが、ファインチューニングを複数回実施した場合に、精度がどのように変化するかといった継続的な改善についても、今後の検証を続けていきたい。

謝辞

本研究を実施するにあたり、言語商会代表の山本和英先生から研究の進め方等について多大なるご支援をいただきました。また、元インターンの高木創志氏には生成AIを活用した作業に協力いただきました。