日中韓辭典研究所
The CJK Dictionary Institute

LLM 翻訳用の用語集自動生成

春遍雀來（Jack Halpern）
株式会社日中韓辭典研究所
2024 年 8 月 29 日

1. LLM 翻訳

近年、GPT-4 等の**大規模言語モデル**（LLM）は日常的に使われるようになり、機械翻訳（MT）にも活用されている。これを **LLM 機械翻訳**もしくは **LLM 翻訳**と呼ぶ。大量の多言語のデータで訓練された LLM は、ニューラル機械翻訳（NMT）に匹敵する高精度な翻訳を行える。しかし、一般的に用いられる学習コーパスには、固有名詞や Point of Interest（POI）のような特定の分野のデータが十分に含まれていないか、ほとんど含まれていないことが多く、その結果、翻訳の質が低下してしまう。

2. 機械翻訳の精度向上

一部の研究者達は、特定分野の機械翻訳の精度を向上させるために、少数の例を用いた微調整（教師あり）やプロンプト指向の微調整（教師なし）などの技術を開発している。別のアプローチとして、推論時に翻訳の質を向上させるために、文脈内学習（ICL）やインストラクションチューニングといったプロンプトベースの技術を利用する方法もある。特に有望な技術の一つとして、外部データソースを追加の文脈として利用する、**検索拡張生成**（Retrieval Augmented Generation, RAG）に基づくプロンプト拡張が挙げられる。

これらの技術には、様々な利点（分野への適応性や文脈への適合性の向上等）と欠点（訓練や推論にかかるコスト等）があるが、それについては別の論文で報告することにする。これらの技術により翻訳の質は全体的に向上するが、固有名詞や POI の多言語データといった特定分野のデータ不足という問題は解決されない。

3. LRAG と用語集生成

当研究所は新しい斬新なアプローチを取ることにした。RAG の手法を語彙レベルで適用して大規模多言語データベースと組み合わせることで、特定分野のデータ不足という核心の問題に対処するツールを開発した。このツールは、翻訳対象の**原文**から抽出した用語に対して、**その分野専用の用語集**を自動的に生成する。この「特定の分野と原文に特化した」工程を、**訳語取得のための検索拡張生成**（Lexical Retrieval Augmented Generation, **LRAG**）と名付けた。**LRAG 用語集生成ツール**という補助ツールが、固有名詞や専門用語の大規模多言語データベース（**LRAG データベース**）から API アクセスによってデータを取得し、**LRAG 用語集**を自動的に作成する。この用語集は、ユーザー辞書を追加してカスタマイズすることもできる。

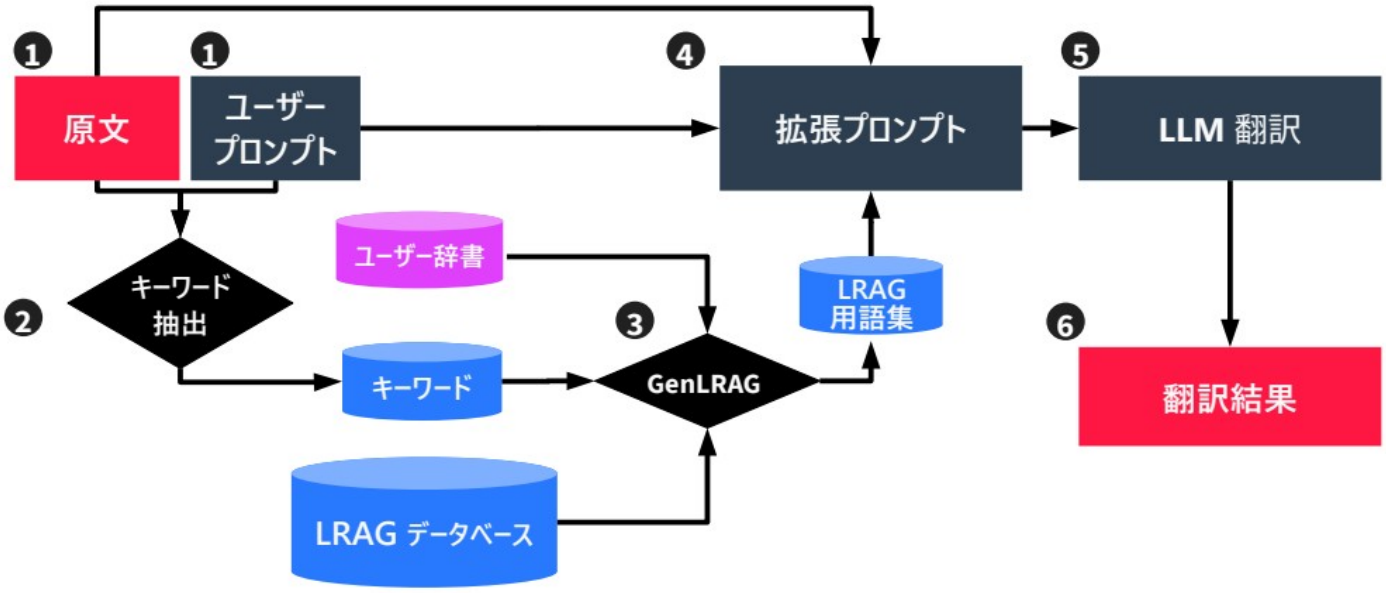
これにより、LLM を再学習や微調整しなくても、大規模多言語用語データベースの内容を外部データソースとして参照し、個々のユーザーの原文や分野に適応させることが可能になる。その結果、誤訳が減り翻訳の精度が大幅に向上する。

4. 特徴的な機能

LRAG 用語集生成ツールは、非常に効率良く用語集を生成するための独自の機能を持つ。

1. 専門用語や固有名詞等の**キーワード**を**自動的に抽出**する。
2. ユーザーが指定した分野にしたり、分野を**自動的に推測**する。
3. 数千万件のエントリが含まれる LRAG データベースへ**リアルタイムにアクセス**する。
4. 訳語の複数の候補を文脈に応じて優先順位付けする。
5. オプションとして**ユーザー辞書**を追加できる。
6. LRAG 用語集を含む**拡張プロンプト**を自動的に生成する。

5. 処理の流れ



LRAG 用語集は上のフローチャートのようにシームレスに生成される。生成プロセスは次の通りである。

1. ユーザーは翻訳の**原文**と**ユーザープロンプト**を用意する。プロンプトでは、原語（翻訳元）、対象言語（翻訳先）、原文の分野等を指定できる。
2. **Extract Keywords** モジュールが、固有名詞や専門用語等のキーワードを原文から抽出する。
3. **GenLRAG** モジュールが、各キーワードを大規模な **LRAG データベース**（とオプションで**ユーザー辞書**）から検索し、訳語の複数の候補の中から、優先順位を計算して適切な訳語を選ぶ。そしてキーワードとその訳語を纏め、今回の原文専用の小規模な **LRAG 用語集**を作成する。
4. **原文**と **LRAG 用語集**をプロンプトに追加して、**拡張プロンプト**を作成する。
5. 拡張プロンプトを LLM 翻訳システムに渡す。
6. 固有名詞や専門用語の訳が改善された翻訳文が生成される。

6. LRAG データベース

日中韓辭典研究所（CJKI）は、超大規模な辞書や語彙データベースの開発を数十年にわたり積極的に進めてきた。**LRAG データベース**は、それらのデータベースを LLM 翻訳システムで用いるために再構成したもので、日中韓およびアラビア語の人名や Point of Interest（POI）等の固有表現を何千万件も収録している。機械翻訳アプリケーション向けに特化して設計されており、LLM 翻訳システムに大きく貢献している。それらのデータベースの内、主なものを以下に紹介する（詳細は https://www.cjk.org/data/all/ を参照）。

1. **中国人名異表記データベース（CNV）**
   中国人の基本人名 160 万項目と主なローマ字異表記を合わせた約 1,000 万項目を収録しており、標準中国語と四つの方言に対応する。

2. **日本語異表記データベース（JOD）**
   JOD は同一語の表記の揺れを識別することで情報検索と機械翻訳に貢献する。例えば、/neko/ (cat) には、猫、ねこ、ネコ；/kakiarawasu/ (write out, publish) には、書き著す、書著す、書き著わす、書著わす等の表記があるが、この揺れを認識することで精度をあげることができる。

3. **日本人名異表記データベース（JNV）**
   日本人の基本人名（姓・名）55万項目とローマ字異表記を合わせた約350万項目を収録しており、ローマ字異表記は、標準的なローマ字表記からその他の一般的な表記、混合型表記まで幅広く網羅する。

4. **アラブ人名データベース（DAN）**
   アラブ人名とそのローマ字異表記を合わせた約650万項目を収録する包括的なデータベースで、母音付きと母音無しのアラビア語表記を共に収録する。

5. **日本語多言語地名 POI データベース（JMP）**
   日本の地名とPOI（駅、学校、空港等の場所）の名称を、中国語、日本語、韓国語、ヨーロッパ諸言語、アジア諸言語に翻訳した大規模データベースである。14言語による多様な分野のPOIを310万項目収録する。

これらのデータベースは、世界最大級のIT企業により、音声技術、形態素解析、機械翻訳等の自然言語処理およびAI（人工知能）アプリケーションにおいて活用されている。

7. LRAG の実例

GPT-4（OpenAI）と Qwen1.5-110B（Alibaba）を用いて行った実験のサンプルを以下に示す。日本の POI や固有名詞が含まれている英語のテキストを日本語に翻訳させた時、GPT-4 と Qwen のいずれも結果は芳しくなかった。

原文	GPT-4 による和訳	LRAG を用いた和訳
As I walk past **Toontown**, I am painfully reminded of the hole it had left in my wallet on my last family trip. Which reminds me; I don't have any cash on me. I make a quick stop at the local ATM of the **Iida Shinkin Bank** to withdraw some money. To my shock, it says out of service. I can't be late again, so I'll just have to borrow some from either **Shoji** or **Wei**, but who should I ask? **Shoji**, on one hand, lives in a small town near **Mt. Yoganezan** and ...	私が**トゥーンタウン**を通り過ぎるとき、前回の家族旅行で私の財布に残した穴を痛感する。それを思い出させてくれます。私は現金を一切持っていません。**飯田信用金庫**の地元 ATM で急いで立ち寄り、少しお金を引き出します。驚いたことに、「運行停止中」だと表示されています。また遅れるわけにはいかないので、**障地**さんか**ウェイ**さんから借りるしかありませんが、誰に頼むべきでしょうか？**ショージ**は一方で、**ヨガネザン山**の近くの小さな町に住んでいます...	私が**トゥーンタウン**を歩き過ぎると、前回の家族旅行で私の財布に残された大きな穴を痛感します。それを思い出させてくれますが、私は現金を全く持っていません。私は**飯田信用金庫**の地元の ATM で素早く立ち寄って、お金を引き出します。驚いたことに、「故障中」だって。また遅刻はできないから、**東海林**か**魏**から借りるしかない。でも、どっちにお願いした方がいいのだろうか？**東海林**さんは一方で、**魚金山**近くの小さな町に住んでおり...

赤: 誤り　**緑**: 正解

全体として、GPT-4 はエラー率 50%で、約半数の POI や固有名詞を誤訳した。Qwen はエラー率 76%というさらに悪い結果を示した。LRAG 用語集を与えた場合、エラー率は GPT-4 が 0%、Qwen が 6%に減少した。

LRAG 用語集

英語	日本語
Wei	魏
Shoji	東海林
Iida Shinkin Bank	飯田信用金庫
Toontown	トゥーンタウン
Mt. Yoganezan	魚金山

この劇的な改善の理由は二つある。

1. LLMに正しい訳語と適切な指示を与えた場合には、LLMがどれだけ用語を知っているかよりも、LLMがどれだけ指示に従うかが、正しい翻訳結果を生成するための主な決定要因となる。

2. 我々の固有名詞やPOIのデータベースは広い範囲をカバーしているので、固有名詞やPOIを含む訓練データに頼らなくても、人名や地名を確実に翻訳できる。

8. 対象ユーザー

この革新的なツールは、様々なタイプのユーザーにとって役立つ。

8.1 個人ユーザー

個人、特に翻訳者が専門的で個別の対応が必要な翻訳作業をする際に、LRAG用語集は翻訳精度を向上させる。カスタマイズ可能な用語集のおかげで、より信頼性が高く個々のニーズに合わせた翻訳が可能になる。

8.2 言語サービス提供会社

言語サービス提供会社（LSP）、特に翻訳会社は、LRAG用語集生成ツールを用いて大規模多言語データベースとユーザー辞書を組み合わせることで、より正確で一貫性のある翻訳のドラフトを顧客への納品前に作成できる。

この高度なツールは柔軟にできているため、機械翻訳システムにシームレスに統合でき、LSPが高品質な翻訳を作成するのに必要な時間と労力を大幅に削減する。

8.3 LLM開発者

LLMを用いたシステムの開発者は、**LRAG用語集生成ツール**や**LRAG用語集**を用いて、以下のような方法でモデルを微調整できる。

1. より堅牢で正確な翻訳ソリューションを作成するために、特定の分野の用語集やユーザー辞書を組み込む。

2. 特定の分野の専門用語集をコーパスとみなし、LRAGデータベースの全内容（数千万件）を訓練データとして用いる。

LRAGデータベースとLRAG用語集により、開発者はユーザーの多様な要望に効率的に対応し、翻訳の品質を向上させることができる。

日中韓辭典研究所（CJKI）

日中韓辭典研究所は辞書の編纂を主たる業務とし、保有する包括的な辞書データベースは約5,000万項目に上る。高品質な日中韓・アラビア語の辞書データやコンサルティングの提供によって高成長を続けるIT産業界を支援する他、ソフトウェア開発にも貢献、先端的な計算辞書学の手法で構築・開発・維持された辞書データベースは、固有名詞抽出・処理・機械翻訳・音声処理技術、情報検索システム等多方面で利用されている。またiPhone/iPad用に、講談社『漢英学習字典』のiOS版や漢字語彙学習アプリ等、各言語の多種多様なソフトを開発、提供している。

春遍雀來（Jack Halpern）

春遍雀來は日中韓辭典研究所の取締役社長である。職業は辞書編纂家で、日本語と中国語を専門としている。学習辞典の編集長として、多くの著名な標準となる辞書を作り上げた。ドイツで生まれ、フランス、ブラジル、日本、アメリカに住んだ経験があり、日本には40年以上住んでいる。熱心なポリグロットで、18言語を学び、12言語を話せる。

株式会社日中韓辭典研究所
〒352-0001 埼玉県新座市東北 2-34-14 小峰ビル
電話：048－473－3508 FAX：048－486－5032

The CJK Dictionary Institute, Inc.
Komine Building 34-14, 2-chome,Tohoku, Niiza-shi
Saitama 352-0001 Japan

E-mail: **jack@cjki.org** URL: **http://www.cjk.org** Phone : 048-473-3508 Fax : 048-486-5032