日中韓辭典研究所
The CJK Dictionary Institute

Automatic Glossary Generator for LLM Translation

Jack Halpern
The CJK Dictionary Institute, Inc.
August 8, 2024

1. LLM Translation

In recent years, **large language models** (LLMs) such as GPT-4 have come into widespread use and are now being employed for machine translation (MT). We will refer to this as **LLM Machine Translation** or **LLM Translation**. LLMs, trained on large quantities of multilingual data, can perform high-precision translation comparable to neural machine translation (NMT). However, domain-specific data such as proper nouns and points of interest (POIs) tend to be underrepresented or missing entirely in commonly used training corpora, which leads to reduced translation quality.

2. Enhancing MT accuracy

Some researchers have developed techniques to enhance domain-specific MT accuracy, such as few-shot example fine-tuning (supervised) and prompt-oriented fine-tuning (unsupervised). Another approach is to utilize prompting-based techniques such as in-context learning (ICL) and instruction tuning to enhance translation quality at inference time. One particularly promising technique is **Retrieval Augmented Generation** (RAG) based prompt-augmentation, which typically utilizes external data sources for additional context.

These techniques have various advantages (such as domain adaptability and increased contextual relevance) and disadvantages (such as training and inference costs), which we will report in a forthcoming paper. However, while these techniques can improve the overall translation quality, they do not solve the issue of the lack of domain-specific data such as proper noun and POI data.

3. LRAG and Glossary Generation

At our institute, we decided to take a fresh, novel approach. We have developed a tool that addresses the core issue of insufficient domain-specific data by applying RAG principles at a lexical level and combining it with our **large-scale multilingual databases**. It automatically generates **domain-specific glossaries** whose terms are extracted exclusively from the source text to be translated. We call this "domain-specific, source-specific" process **Lexical Retrieval Augmented Generation (LRAG)**. An auxiliary tool called the **LRAG Glossary Generator** retrieves data via API access from our large-scale multilingual databases of proper nouns and technical terms, referred to as **LRAG Databases,** and automatically creates the **LRAG Glossary**, which can be optionally customized by supplementing it with **User Dictionaries**.

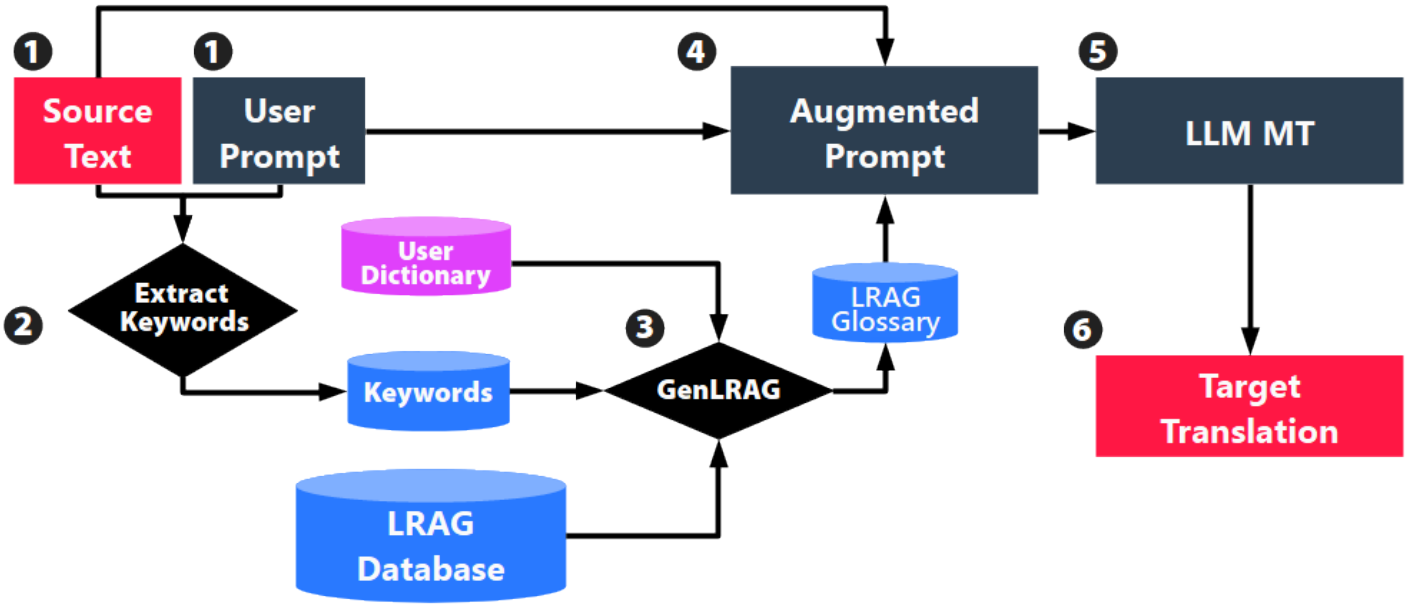
This allows the LLM to leverage the contents of large-scale multilingual terminology databases as an external data source without the need to retrain or fine-tune the LLM itself, while retaining the ability to adapt to specific user source texts and domains. As a result, translation errors are reduced and translation accuracy can be significantly improved.

4. Distinctive Features

The **LRAG Glossary Generator** offers unique features that enable highly efficient glossary generation.

1. **Keywords** such as technical terms and proper nouns are **automatically extracted**.
2. Users can specify the domain or the domain is **automatically inferred.**
3. **Real-Time access** to the LRAG Databases, which consist of tens of millions of entries
4. Multiple translation equivalents are prioritized by context.
5. **Optional User Dictionaries** can be added.
6. An **Augmented Prompt** including the LRAG Glossary is automatically generated.

5. Process Flow



The LRAG Glossary is generated seamlessly as shown in the flow chart above. The generation process is as follows:

1. The **Source Text** and the **User's Prompt** are provided by the user. The prompt specifies the source language, target language and optionally the domain of the source text.
2. The **Extract Keywords** module extracts keywords such as proper nouns and technical terms from the source text.
3. The **GenLRAG** module searches each keyword in the large-scale **LRAG Databases** and the optional **User Dictionary**, then selects the appropriate equivalents by calculating the priority among multiple candidates. The keywords and their equivalents are compiled into a small-scale **LRAG Glossary** specific to the source text.
4. The **Augmented Prompt** is created by combining the **source text** with the **LRAG Glossary**.
5. The Augmented Prompt is passed to the LLM MT system.
6. The target translation text with improved translations of proper nouns and technical terms is generated.

6. LRAG Databases

For several decades, **The CJK Dictionary Institute** (CJKI) has been actively developing ultra-large-scale dictionary and lexical databases that are being repurposed for use in LLM MT systems, known as **LRAG Databases.** They contain tens of millions of named entities (proper nouns) such as personal names and POIs (points of interest) in the CJK (Japanese, Chinese, and Korean) languages and Arabic. They are designed specifically for MT applications, significantly contributing to MT systems based on LLMs. Below is a description of some of the major databases (more details at cjk.org/all).

1. Chinese Personal Names Variant Database (CNV)

CNV contains approximately 10 million entries, including 1.6 million basic Chinese personal names and major Romanized variants, supporting standard Chinese and four dialects.

2. Japanese Orthographic Variants Database (JOD)

JOD contributes to information retrieval and MT by identifying orthographic variants of the same word. For example, /neko/ (cat) can be written as 猫, ねこ, ネコ; /kakiarawasu/ (write out publish) can be written as 書き著す, 書著す, 書き著わす, 書著わす, and recognizing these variants can improve accuracy.

3. Japanese Personal Names Variant Database (JNV)

Contains approximately 3.5 million entries, including 550,000 basic Japanese personal names (surname and given name) and Romanized variants, covering a wide range of standard and common Romanized variants.

4. Arabic Personal Names Database (DAN)

A comprehensive database of approximately 6.5 million entries, including Arabic personal names and their Romanized variants, with both vocalized and unvocalized Arabic entries.

5. Japanese Multilingual Place Name POI Database (JMP)

A large database of Japanese place names and POIs (points of interest such as stations, schools, airports, etc.) in Chinese, Japanese, Korean, European, and Asian languages, containing 3.1 million items in 14 languages.

These databases are used by the world's largest IT companies for NLP and AI applications such as speech technology, morphological analysis, and machine translation.

7. LRAG in Action

The following is a sample of experiments we have conducted with GPT-4 (OpenAI) and Qwen1.5-110B (Alibaba). When presented with the task of translating an English text containing Japanese POIs and proper nouns into Simplified Chinese, both GPT-4 and Qwen performed poorly:

Source Text	Golden Standard	GPT-4 Translation
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 ...	当我走过**卡通公园**，我痛苦地想起了上一次家庭旅行像是把我的钱包弄了个大洞。这让我想起来；我身上没有任何现金。我走到**饭田信用金库**的 ATM，想赶紧取些钱。让我震惊的是，它显示“停止服务”。我不能再迟到了，所以我只得向**东海林**或**魏**借一些钱，但我该问谁呢？一方面，**东海林**住在**鱼金山**附...	当我走过 Toontown 时，我痛苦地回想起上次家庭旅行时它给我的钱包留下的巨大空洞。这让我想起来；我身上没有现金。我在**飯田信用金庫的**当地 ATM 进行了个快速停留以取些钱。让我震惊的是，它显示“停止服务”。我不能再迟到了，所以我只能从**翔二**和**伟**那里借一些，但我应该问谁呢？Shoji 一方面住在靠近 Yoganezan 山的...

Red: Error **Green**: Correct

Overall, GPT-4 had an error rate of 54%, meaning that it got the translation wrong for more than half of all POIs and proper nouns. Although Qwen performed slightly better with an error rate of 42%, the result is still less than desirable. When provided with the LRAG Glossary, the error rate went to 0% for both GPT-4 and Qwen.

LRAG Glossary

English	Chinese
Wei	魏
Shoji	东海林
Iida Shinkin Bank	饭田信用金库
Toontown	卡通公园
Mt. Yoganezan	鱼金山

The reason for this drastic improvement is two-fold:

1. By providing the LLMs with the correct translations and appropriate instruction, the LLMs ability to follow instructions became the main determining factor in producing the correct translation, rather than it's knowledge of any given term.
2. The extensive coverage of our proper noun and POI Databases ensures that personal names and place names can be reliably translated without relying on the training data containing such proper nouns and POIs.

8. Target Users

Different types of users can benefit from this novel tool.

8.1 Individual Users

Individuals, especially translators, can benefit from the LRAG Glossaries for personal and professional translation tasks, achieving higher translation accuracy. The ability to customize glossaries ensures that translations are more reliable and personalized.

8.2 Language Service Providers

By leveraging our large-scale multilingual databases combined with customer-specific user dictionaries, language service providers (LSPs) (especially translation companies) can use the LRAG Glossary Generator to create more accurate and consistent translation drafts for post-editing before delivering translation products to customers.

The flexibility of this advanced tool enable seamless integration into MT systems, significantly reducing the time and effort required by LSPs to produce high-quality translations.

8.3 LLM Developers

Developers of LLM-based systems can utilize the **LRAG Glossary Generator** and the **LRAG Databases** to fine-tune their models in the following ways:

1. Incorporate domain-specific glossaries and user dictionaries to create more robust and accurate translation solutions.
2. Use the entire contents of the LRAG Databases (tens of millions of entries) as training data by considering the domain-specific terminologies as a corpus.

The LRAG Databases and LRAG Glossaries enable developers to meet diverse user needs efficiently and to enhance translation quality.

The CJK Dictionary Institute

The CJK Dictionary Institute (CJKI) was founded in 1993. Its principal activity is the compilation of large-scale dictionary databases of proper nouns and technical terms for CJK (Chinese, Japanese, Korean) and Arabic, currently with over 50 million entries. CJKI has become the world's prime source for CJK lexical resources for the IT industry and software developers, providing high-quality comprehensive dictionary data, educational tools, and consulting services. Based in Saitama, Japan, CJKI is headed by Jack Halpern, editor in chief of The Kodansha Kanji Learner's Dictionary and several other dictionaries that have become standard works for learning Japanese.

Jack Halpern

Jack Halpern (春遍雀來), CEO of **The CJK Dictionary Institute**, is a lexicographer by profession, specializing in Japanese and Chinese. His work as an editor in chief of learner's dictionaries resulted in various renowned standard reference works. He has been a resident of Japan for over 40 years but was born in Germany and has lived in France, Brazil, Japan, and the United States. He is an avid polyglot who has studied 18 languages (speaks 12).

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