

大语言模型驱动的语料库翻译研究 方法创新*

北京外国语大学 许家金 孙铭辰

提 要：语料库翻译研究历经30余年的发展，研究议题和成果数量总体呈现收拢趋势。然而，以ChatGPT为代表的大语言模型可以有效助推语料库翻译研究的方法升级和学科创新。本文通过精编的提示语，展示如何借助大语言模型提升语料库翻译研究在双语文本采集加工、数据深度分析、理论建构等方面的创新成效。

关键词：大语言模型；语料库翻译研究；后经典时代；大模型语言学研究方法；提示语

Advancing Corpus-based Translation Studies Through Large Language Models

Jiajin Xu and Mingchen Sun

Abstract: Corpus-based translation studies have undergone over 30 years of development, and the number of research topics and achievements generally shows a converging trend. The advent of large language models, represented by ChatGPT, however, holds the potential to effectively boost the methodological progress and introduce novel perspectives within the discipline. This article highlights, through carefully-written prompts, how to leverage large language models to enhance the innovation of corpus-based translation studies in areas like bilingual text collection, in-depth data analytics, and theory construction.

Keywords: Large Language Model; corpus-based translation studies; post-classic era; LLM Linguistics Methodology; prompt

* 本文系国家社科基金重大项目“围绕汉语的超大型多语汉外平行语料库集群研制与应用研究”（项目编号：21&ZD290）的阶段性研究成果。

作者贡献：

许家金：选题构思、数据收集、研究方法、初稿撰写、修改润色、字数占比（30%）；

孙铭辰：数据收集、数据分析、初稿撰写、字数占比（70%）

1 引言

以 ChatGPT 为代表的大语言模型已在社会生活各个场景得到广泛应用，同时也引发人们关于通用人工智能时代来临的讨论。在此背景下，语言学界已开始探讨如何有效利用大语言模型助力语言学和应用语言学研究（冯志伟等，2023；袁毓林，2023）。语料库翻译研究有条件利用大语言模型开启后经典时代（许家金，2023）研究范式的新发展阶段。本文尝试结合案例，对大语言模型驱动的语料库翻译研究方法创新作初步探讨。

随着大语言模型的广泛运用，在文本生成、数据分析等方面我们已经进入“智能后编辑”时代。大模型的出现显著降低了语料库建设、加工、整理、分析的技术门槛。研究者只需明确选题需求，即可用自然语言“提示语”（prompt）作为指令，引导大语言模型对文本采集加工和数据分析进行高效处理。研究者在此基础上再作核验即可。遵循大模型驱动的语料库分析思路，研究效率大大提高。在研究方法层面，研究者对分析工具的利用，从强调“搜商”的时代进入到比拼“问商”（prompting quotient）（许家金等，2024，2025）的时代。如今，大语言模型俨然已成为研究人员的高级助理。

2 现有语料库翻译研究方法的高效替代

在“智能后编辑”时代，研究者可通过编写指令，引导大语言模型采集文本、加工文本和分析文本，从而实现“经典时代”语料库研究方法的技术替代。现有语料库翻译研究方法的基础替代技术，可分为语料库建设方法替代和语料库分析方法替代两个层级。为展示方法、重在示例，并控制篇幅，本文中所用大模型提示语和结果都在一定程度上作了简化。文中省略号处所略去部分的完整信息及其他相关资料可通过访问北京外国语大学语料库团队“PromptBank 提示库”页面^①获取。

2.1 语料库建设方法的替代技术

语料库建设方法替代主要分为语料获取和语料对齐两个方面。在传统的平行语料库建设中，语料的来源多采用人工收集或软件爬取的方式（如 BootCat、HTTrack 等）；双语语料也多采用单机或在线软件（如 ABBYY Aligner、TMXmall、OmegaT 等），来实现基于可扩展标记语言（XML/TMX）的平行文本句对齐或段对齐。大语言模型的问世，将帮助研究者以更加“用户友好型”的方式，快速、准确地实现上述过程。本文将双语平行语料的网页爬取和 TMX 对齐为例进行介绍。

^① “PromptBank 提示库” 页面详见 <https://promptbank.unipus.cn>。

2.1.1 双语平行语料的网页爬取

网页爬取是双语平行语料获取的重要方式。可基于GPT-5编写Python代码，调用“requests”和“BeautifulSoup”配置爬虫，以获取《中国日报》“双语新闻播报”栏目的平行语料。

(1) 提示语 (GPT-5):

```
Please write Python code that uses the requests and BeautifulSoup libraries to crawl China Daily's 'Bilingual News Broadcast' section and scrape each article's title, date, URL, and full text.
```

结果:

```
# here's a clean, robust scraper for China Daily's "Bilingual News Broadcast / 新闻播报" section
```

```
import re
import time
import json
import csv
import random
from urllib.parse import urljoin
import requests
from bs4 import BeautifulSoup
...
```

2.1.2 实现双语平行文本句对齐

针对已爬取的《中国日报》“双语新闻播报”语料，可利用Claude Sonnet 4模型实现句对齐，并以翻译记忆交换格式 (Translation Memory eXchange, 简称TMX) 输出结果，建立平行语料库。

(2) 提示语 (Claude Sonnet 4):

```
CD_en.txt contains the original text, and CD_zh.txt is its corresponding Chinese translation. Could you align the two uploaded texts at the sentence level, and then output the results in the Translation Memory Exchange (TMX) format?
```

结果:

```
I'll align them as Chinese-English sentence pairs and create the TMX format.
```

```
<xml version="1.0" encoding="UTF-8" >
<tmx version="1.4" >
<header><prop type="x-filename" >CD_alignment.tmx</prop></header>
<body>
<tu>
<tuv xml:lang="en"><seg>Despite public concern and raging opposition from
both home and abroad, the Japanese government announced on Tuesday that it
has decided to start releasing nuclear-contaminated wastewater from the
crippled Fukushima Daiichi Nuclear Power Plant into the ocean on the August
24.</seg></tuv>
<tuv xml:lang="zh" ><seg>日本政府不顾国内外公众的担忧和强烈反对,22
日宣布决定于8月24日启动福岛核污染水排海。</seg></tuv>
</tu>
...
```

2.2 语料库分析方法的替代技术

语料库分析方法替代主要体现在文内、文外的共选生义。在语料库研究方法的“经典时代”，翻译语言特征研究、翻译规范研究和译者风格研究多聚焦于单特征，通过 AntConc、ParaConc 等软件，统计源语和译语的词频、平均句长、型次比等，进行语言特征的计量对比，开展翻译转换特征、翻译文体、语言对比等方面的研究。在大语言模型驱动的语料库“后经典时代”研究中，可通过编辑指令驱动大模型完成文内、文外的多因素统计分析。本文将词汇特征统计、翻译共性研究和译者风格研究为例进行介绍。

2.2.1 译语词汇特征统计

大语言模型可替代原有软件，对多种译语词汇特征的统计给出方法。以下为基于 Claude Sonnet 4 编写的 R 语言代码，以实现输入语料进行词数、型次比、标准化型次比、移动平均型次比、动词词组统计、名词/代词比、实词/虚词比、词汇密度、词汇多样性、词汇复杂性等的批量统计。

(3) 提示语 (Claude Sonnet 4):

Please write R code to perform comprehensive statistical analysis of multiple texts, analyzing the following linguistic features: word count, type-token ratio, standardized type-token ratio, moving-average type-token ratio, verb phrase count, noun-pronoun ratio, content/function ratio, lexical density, lexical variation, lexical sophistication.

结果:

Here's comprehensive R code to perform statistical analysis of multiple texts for the specified linguistic features:

```
# Install and load required packages
required_packages <- c("tidyverse", "quanteda", "quanteda.textstats",
                      "udpipe", "spacyr", "koRpus", "readtext",
                      "ggplot2", "corrplot", "knitr")

# Install missing packages
new_packages <- required_packages[!(required_packages % in% installed.
packages()[, "Package"])]
if(length(new_packages)) install.packages(new_packages)
...
```

2.2.2 《道德经》英译的翻译共性

大语言模型可考虑文内、文外多种因素，对翻译共性开展研究，且大语言模型可在学习经典文章、重要概念的基础上对输入语料进行分析。在以下案例中，本文在 GPT-5 模型网页交互界面上上传 Baker (1993, 1995, 1996) 三篇语料库翻译学研究的经典论文供大语言模型学习；同时，提取 Baker (1993, 1996) 中有关翻译共性（简化、显化、范化、整齐化）方面的重要论述编写入提示语中。基于上传文档和提示语，引导 GPT-5 从翻译共性的角度对许渊冲版《道德经》英译本的前 10 章进行分析。

(4) 提示语 (GPT-5):

Please read the uploaded File I, File II and File III. These articles are Baker's (1993, 1995, 1996) three classic articles concerning corpus-based translation studies.

Please note (some explanations of "translation universals"):

Universal features of translation: It might be useful at this point to give a few examples of the type of translation universals I have in mind ...

Explication: I take "explication" to mean that there is an overall tendency to spell things out rather than leave them implicit in translation ...

Based on the input File I, File II, File III, and the notes presented above, please

study the concept of “translation universals,” with a special focus on explicitation, simplification, normalization and levelling-out. According to your study of “translation universals,” please analyze the first 10 chapters of Yuanchong Xu’s translation of “Tao Te Ching” from the perspective of translation universal.

结果:

Here’s a compact, evidence-based read of Xu Yuanchong’s Chapters 1 - 10 through Baker’s “translation universals.” I first anchor each universal in Baker, then point to concrete places in Xu’s English where that universal plausibly surfaces.

What “translation universals” are (per Baker)

Corpus-based work in translation studies looks for recurring features of translated text that are not mere source-language interference—e.g., explicitation, simplification, normalisation/conservatism, and levelling-out.

...

Xu’s Chapters 1 - 10: where the universals show up

1) Explication

Chapter 1 “道可道，非常道；名可名，非常名。” → “The divine law may be spoken of, but it is not the common law. ... So we should be free from desires ... in order to understand ...”

Xu supplies category labels (“divine law,” “common law”) and explicit causal/purposive connectors (“so,” “in order to”), making implicit logic overt.

...

2.2.3 《红楼梦》霍译本和杨译本译者风格比较

大语言模型也可开展译者风格的研究。可在GPT-5模型网页交互界面，将《红楼梦》原文、霍克斯与闵福德译本（简称霍译本）和杨宪益与戴乃迭译本（简称杨译本）全文上传，并基于GPT-5模型，对霍译本和杨译本进行译者风格的比较。在编写提示语时，本文输入了Baker（2000）中245—246页中关于译者风格的解释以供GPT-5参考。

(5) 提示语（GPT-5）:

File I is the original text of Honglouloumeng. File II is Hawkes’ translation of

Honglouloumeng. File III is Yangs' translation of Honglouloumeng. Please compare File II (Hawkes' translation) and File III (Yangs' translation) concerning translator style with File I (original text) as the reference.

Please note Mona Baker's (2000) understanding of translator style:

I understand style as a kind of thumb-print that is expressed in a range of linguistic — as well as non-linguistic — features ... But like both branches of stylistics, I am interested in patterns of choice (whether these choices are conscious or subconscious) rather than individual choices in isolation.

结果:

Here's a focused, Baker-style comparison of Hawkes (File II) and Yangs (File III), using the Chinese original (File I) as the baseline and treating "style" as recurring, measurable patterns of linguistic habit — not one-off choices. In Baker's terms, we're looking for a translator's "thumb-print" in their preferred patterns of expression, including (but not limited to) degree of explicitation, sentence shaping, naming practices, and narrator stance....

Explication of author's preface (自述“梦幻……真事隐去”)

Hawkes (File II): Compresses/angles the metafiction into urbane English; less meta-commentary spelled out.

Yangs (File III): Retells the preface step by step (“In writing this story... he tried to hide the true facts... Hence... names like Zhen Shiyin.”).

What this suggests (vs. File I): Yangs systematically unpack implicit cues; Hawkes tends to recreate tone rather than explicate every link.

...

3 语料库翻译研究方法的提质升级

在“智能后编辑”时代，大语言模型的出现可显著提升原有语料库翻译研究技术的处理质量（如3.1节），可较大程度代替人工的标注和判断过程（如3.2节、3.3节），并助力低资源多语种平行语料库的建设（如3.4节）。

3.1 基于大语言模型的情感分析——以“被”字句为例

大语言模型可为情感分析研究提质升级。基于大语言模型的情感分析同传统情感分析相比，前者数据集、训练参数和迭代次数更为庞大，具有更强的语境理解能

力，能够处理更加复杂的语义结构，显著提升情感分析的质量。以“被”字句为例，本文基于 Claude Sonnet 4 模型，对“浙江大学翻译汉语语料库 ZCTC”小说部分，以及原创汉语“兰卡斯特汉语语料库 LCMC”小说部分，进行了翻译汉语和原创汉语“被”字句使用的情感分析比较。

(6) 提示语 (Claude Sonnet 4):

Bei_Fiction_ZCTC.txt contains 346 concordance lines extracted from the translated Chinese corpus ZCTC fiction section using the word ‘被’ (‘bei’) as the passive marker. Bei_Fiction_LCMC.txt contains 376 concordance lines extracted from original Chinese corpus LCMC fiction section using the word ‘被’ (‘bei’) as the passive marker. Please perform sentiment analysis on these two files, and compare these two files with examples and explanations. Please output the results in English.

结果:

1. Sentiment Distribution Patterns:

LCMC (Original Chinese Fiction):

Higher prevalence of negative sentiment in passive constructions

Common patterns: 被抛弃 (abandoned), 被打伤 (injured), 被杀害 (killed)

More emotionally intense passive constructions related to suffering and harm

ZCTC (Translated Chinese Fiction):

More neutral/procedural passive constructions

Examples: 被送到 (sent to), 被安排 (arranged), 被称为 (called)

Less emotionally charged passive voice usage

2. Semantic Field Analysis:

Violence/Harm Domain:

LCMC: 被砸、被撞、被碾裂 (smashed, hit, crushed), 被烈火烧伤 (burned by fire)

ZCTC: 被一条鱼割伤 (cut by a fish), 被车撞 (hit by car)

Social Relations:

LCMC: 被抛弃 (abandoned), 被轻视 (despised), 被人淡忘 (forgotten)

ZCTC: 被人理解 (understood), 被邀请 (invited), 被允许 (allowed)

...

3.2 翻译策略判断

大语言模型可为翻译策略研究提质升级。本节以“文化负载词及翻译策略”“隐喻及翻译策略”和“口译策略”为例展开讨论。在以往研究中，“文化负载词”“隐喻”等的识别多采取基于形式进行提取的方式，识别准确率较低且漏识率较高；而翻译策略的判断更需人工进行。在“智能后编辑”时代，研究者可编写提示语引导大语言模型去进行标注判断，研究者基于大语言模型的结果再进行修改优化。

3.2.1 文化负载词及其翻译策略

本文通过GPT-5模型，对《2023年国务院政府工作报告》中英平行语料进行了中国文化负载词的提取及其翻译策略的分析（如音译、音译+注释、直译、直译+注释、音译+直译、意译等）。在编写提示语时，本文参考了郑德虎（2016）对中国文化负载词的定义及其常用翻译策略的解释。

(7) 提示语（GPT-5）：

The following text is an excerpt from the 2023 REPORT ON THE WORK OF THE GOVERNMENT, provided in both its original Chinese and its English translation. Please analyze the translation strategy used in the translation of the Chinese culture-loaded terms, such as transliteration, transliteration + annotations, literal translation, literal translation + annotations, transliteration + literal translation, free translation. Please present the results in a tabular format.

Please note:

Culture-loaded terms: Culture-loaded terms refer to ...

Literal Translation: Translating according to the literal order or meaning. For example: 三思而后行 think thrice before you take any action; 春节 spring festival; 龙舟 dragon boat

Free Translation: Translation method that retains the content but not the form of the original text. For example: 颐和园 the summer palace; 诗经 The Book of Songs

...

结果：

Here's a focused, term-by-term look at the culture-loaded items in your excerpt and the strategies used in their English renderings.

Legend (strategies)

T = Transliteration (pinyin) • T+A = Transliteration + Annotation • L = Literal translation • L+A = Literal + Annotation • T+L = Transliteration + Literal • F = Free translation (functional/interpretive)

Chinese term (source)	English rendering (target)	Strategy	Quick rationale / how it maps	Suggested alternative (if helpful)
“十四五” 规划	the 14th Five-Year Plan	L	Standard literal policy term; widely conventionalized	—
专项债务结存 限额	carryover quotas for special-purpose bonds	F	Interprets “结存限额” as “carryover quotas”; not literal	“remaining quota for special-purpose debt” (L)
政策性开发性 金融工具	development and policy-backed financial instruments	L	Term of art rendered literally (order slightly smoothed)	“policy-based and development-oriented...” (closer L)
专项再贷款	targeted re-lending	L	Accepted literal macro-finance term	—
...

3.2.2 隐喻及其翻译策略

本文通过GPT-5模型，基于Newmark（1981）的七种隐喻翻译策略，对《2020年国务院政府工作报告》中的部分平行语料进行了隐喻的标注及其翻译策略的解析。在编写提示语时，本文参考了Newmark（1981）及Schäffner（2004）中关于隐喻翻译策略的解释。

(8) 提示语（GPT-5）:

The following texts are the original Chinese texts and their corresponding English Translations extracted from 2020 REPORT ON THE WORK OF THE GOVERNMENT. Can you annotate the metaphor cases in the original texts and analyze the translation strategies of the metaphor cases based on Newmark’s theory? Please output the results in a table format.

Here are Newmark’s 7 classifications of translation strategies of the metaphor:

1. Reproducing the same image in the target language, e.g., golden hair - goldenes Haar...

5. Converting metaphor to sense, e.g., sein Brot verdienen - to earn one’s living ...

结果:

Here’s a concise, research-style annotation of metaphor cases in the Chinese

source (SL) with their English renderings (TL), analyzed using Newmark’s seven procedures.

#	SL metaphor (Chinese)	Literal gloss	Context (SL snippet)	TL rendering (English)	Source domain	Newmark strategy	Rationale (1 - 2 lines)
1	打赢脱贫攻坚战	“win the tough battle of poverty alleviation”	坚决打赢脱贫攻坚攻坚战。	We will win the fight against poverty.	War	1. Same image	War metaphor “battle/fight” is kept intact in TL.
2	攻坚（加大……攻坚力度）	“assault a fortified position”	加大……贫困县和贫困村攻坚力度……	...channel greater energy into eliminating poverty...	War / Siege	5. Metaphor →sense	Military image dissolved into neutral “eliminating / greater energy.”
3	兜底保障	“catch-all underlay; safety net”	强化对特殊贫困人口兜底保障。	ensure that the basic needs of the most vulnerable...	Safety net	5. Metaphor →sense	Idiomatic metaphor “兜底” is explicated as meeting basic needs.
...

3.2.3 口译策略

本文通过GPT-5模型，参考中国口译学习者语料库CILC（张威，2015）口译策略的标注体系（代码转换、词性转移/句法转移、主动省略/被动省略、替代、简化、语义解释/语素增补、切分和重复），对“记者招待会平行语料库PICME”的2017年子库进行了口译策略的标注和分析。在编写提示语时，本文参考了张威（2015）对CILC语料库口译策略标注的解释及其示例。

(9) 提示语（GPT-5）:

Please annotate the interpretation strategy employed in the texts in terms of ... rank shift — lexical ... ellipsis — active ...

Please note the interpretation strategy and its annotation format:

...

Rank shift — lexical (<RSL> ... </RSL>) refers to the change in the grammatical nature of a word from the source text in the translation to facilitate the immediate conveyance of interpretation information.

Original: That change is a recognition of how seriously Europe takes our relations with Cambodia.

Interpretation: 这个变化<RSL>反映出</RSL>欧洲在很认真地对待与柬埔寨的关系。

Ellipsis — active (<ESA> ... </ESA>): is when an interpreter deliberately adjusts the information volume and structure of the original text.

Original: 威胜仪表2005年度仍然取得骄人成绩。全年营业额达人民币4.5亿元,较2004年度增长39%。

Interpretation: In 2005, Wei Sheng Meter achieved a very good turnover, reaching <ESA> 450 million RMB </ESA>, 39% more than that of 2004.

...

结果:

Here's an annotated table of interpretation strategies identified in your sentence-aligned excerpts. Each row shows the tagged source (CN) and target (EN), the strategy label, and a brief rationale.

Annotation (CN / EN)	Strategy	Explanation
CN: 我的回答是, 不管谁当选美国总统, 虽然中美关系经历过<SBT>风风雨雨</SBT>, 但是一直前行, 我对此持乐观态度。EN: Well, I said that China - U.S. relations have been going forward in spite of <SBT>various twists and turns</SBT> in the past several decades. <SPL>So I am optimistic about this relationship, no matter who gets elected.</SPL>	<SBT>, <SPL>	Idiom “风风雨雨” rendered as “twists and turns” (idiomatic equivalence, <SBT>); one CN sentence split into two EN sentences (<SPL>).
CN: 有了这样一个政治基础, 中美合作的前景是<SBT>广阔的</SBT>。 EN: With that foundation in place, <MPA>we believe</MPA> there are <SBT>bright prospects</SBT> for China - U.S. cooperation.	<SBT>, <MPA>	“广阔的” → “bright prospects” (<SBT>); insertion of stance marker “we believe” (<MPA>).
...

3.3 翻译错误判断

大语言模型可为翻译错误研究提质升级。错误分析是大语言模型较为擅长的功能, 可由此进行学习者的口笔译的偏误分析, 开展翻译教学研究。本文基于 Claude Sonnet 4 模型, 对“中国大学生英汉汉英口笔译语料库”中五名学习者的汉英笔译语料进行了错误分析, 并提出修改建议。

(10) 提示语 (Claude Sonnet 4):

Below are some Chinese-English translations of Chinese EFL students from the Parallel Corpus of Chinese EFL Learners (PACCEL). Please proceed with

error analysis. Please output the results with error explanations and suggestions in a table format.

Here are the original texts:

<CE1>除了电脑，电影电视也在散布不良信息。

<CE2>在这方面，我的态度是坚定的：我赞成严加管理。

...

Here are Chinese EFL Learners' translations from 5 students:

<WRITTEN><GRADE3><YEAR04><CE><SEX=M><TEM4=70><CET6=><TEM8=><F101><Y06>

<CE1>Beside computers, movies and television programmes also spread the negative information.

<CE2>To this point, I extremely insist on my attitude: I agree to strengthen the management.

...

结果:

Error Analysis of Chinese EFL Learners' Translation

Sentence	Student	Error Type	Error	Correction	Explanation / Suggestion
CE1	S1	Spelling	nelgative	negative	Check spelling carefully; use spell-check tools
CE1	S1	Preposition	Beside	Besides	"Besides" means "in addition to"; "Beside" means "next to"
CE1	S3	Word Order	something bad is spreading over movies and televisions	bad information is spreading through movies and televisions	Maintain logical subject - verb - object order
CE2	S1	Redundancy	extremely insist on	insist on / am extremely firm about	Avoid redundant intensifiers
CE2	S3	Word Choice	rigid conductions	strict management	Use appropriate collocations
...

3.4 生成多语平行语料库——中、英、汤加语平行语料库的创建

大语言模型可助力低资源多语种平行语料的建设。用大语言模型生成式语料库 (AIGC Corpus) 建设低资源多语种平行语料库是现有条件下的次优解。大语言模

型所生成的语料虽不属于现实交际中真实运用的语料，但其生成水平一定程度上接近本族语者水平，对于一些极度欠发达或者人口规模极小国家语种的语料库建设，可通过大语言模型短期内快速实现。本节以中、英、汤加语平行语料库的创建为例，生成“日常技艺及消遣爱好”（Brown 语料库E类）的100个中、英、汤三语平行句对。

(11) 提示语（Gemini 2.5 Pro）：

Please generate 100 parallel sentences in Chinese, English, Tongan related to “Skills and Hobbies.” Please create a new line for each aligned sentence triplet.

结果：

Here are a few examples to demonstrate the structure:

我喜欢弹钢琴。| I enjoy playing the piano. | ‘Oku ou fiefia ke tā piano.

我的爱好是摄影。| My hobby is photography. | Ko e hoku ‘ofa‘angá ko e tā fakatātā.

他擅长游泳。| He is skilled at swimming. | ‘Okú ne taukei ‘i he kakaú.

她正在学习编织。| She is learning to knit. | ‘Okú ne akoako ke lalanga.

我会说三种语言。| I can speak three languages. | ‘Oku ou lava ‘o lea ‘aki ha lea fakafonua ‘e tolu.

4 语料库翻译研究方法的应用升级

在与语料库翻译学研究相关的应用场景中，大语言模型的出现带来技术的升级变革，双语术语抽取、翻译质量评估、语义相似度计算、双语词典编纂等过程也在大语言模型的帮助下能够快速准确地实现。

4.1 双语术语抽取

基于大语言模型的双语术语抽取（bilingual terminology extraction）是重要的应用技术升级，可为翻译教学提供丰富资源。现有的双语术语抽取软件，大多基于形式进行抽取，在对齐文本的基础上考虑词频、长度等因素，准确率相对较低，且误抽率较高。大语言模型以语义为核心，基于深度学习对平行语料进行分析，可严谨地实现双语术语的抽取。本文以《2020年国务院政府工作报告》为例，通过DeepSeek-reasoner进行双语术语的抽取。

(12) 提示语（DeepSeek-reasoner）：

Please perform bilingual terminology extraction from the following parallel

texts. Here are the texts.

<seg><zh>二、今年发展主要目标和下一阶段工作总体部署</zh>

<en>II. Main targets for development in 2020 and the overall plan for the next stage of work</en></seg>

...

结果:

Macroeconomic Goals and Philosophy

Chinese (中文)

English

全面建成小康社会

building a moderately prosperous society in all respects

稳中求进工作总基调

the general principle of pursuing progress while ensuring stability

新发展理念

the new development philosophy

供给侧结构性改革

supply-side structural reform

...

4.2 翻译质量评估

基于大语言模型的翻译质量评估 (translation quality assessment) 是另一个重要的应用技术升级。大语言模型可通过多因素分析, 考虑多个文内文外因素, 实现翻译文本质量的综合评估。本文以“TED 英汉平行演讲语料库”为例, 通过 GPT-5, 对该库中以“众包”形式完成的中文翻译进行翻译质量评估。本文在编写提示语时, 上传了 House (2014) 248—258 页关于翻译质量评估模式 (linguistic, functional-pragmatic model of translation criticism) 的论述以供 GPT-5 参考。

(13) 提示语 (GPT-5):

The document describes the linguistic, functional-pragmatic model of translation criticism discussed in pages 248-258 of House (2014). Below is a TED English speech and its Chinese translation. Please perform translation quality assessment on the translation.

Here are the texts: ...

结果:

Here's a concise House-style Translation Quality Assessment (TQA) of the Chinese version, using House's linguistic, functional-pragmatic model (Field

- Tenor - Mode; overt vs. covert; dimensional vs. non-dimensional mismatches).

1) Text profile & translation brief (House)

Genre: Public science communication / persuasive TED talk.

Register (ST):

Field: Links ocean health to human health; uses a historical pollution vignette (Cannery Row), ecology concepts (food web/pyramid).

Tenor: Marine ecologist to lay audience; informal, humorous, rapport-building (colloquialism “ain’t,” rhetorical questions, quotes).

Mode: Written-to-be-spoken monologue with audience involvement cues (aphorisms, anaphora, conversational asides).

...

4.3 原文、译本语义相似度计算

大语言模型可应用于语义相似度的计算。基于词嵌入对文本进行向量化，是语义相似度计算的常用方法。在翻译研究中，基于大语言模型，可生成对应的计算语言代码，以实现平行语料、可比语料的语义相似度比较。在该小节中，本文通过 Claude Sonnet 4 生成相应的 Python 代码，以调用 Sentence Transformer 模型，从而对 TMX 格式的英汉平行语料库的原文与译文进行语义相似度的比较。

(14) 提示语 (Claude Sonnet 4):

Based on Python, how can I compare the semantic similarity between the source text and the translation in an English-Chinese parallel corpus stored in TMX format using Sentence Transformer?

结果:

Here’s a comprehensive Python solution to compare semantic similarity between source and target texts in a TMX parallel corpus using Sentence Transformers:

Install Required Dependencies

pip install sentence-transformers lxml pandas numpy scikit-learn

...

4.4 双语词典编纂

大语言模型可应用于双语词典编纂。在大语言模型的帮助下，基于“概率语境

共选”（许家金，2020）和“四用原则”（许家金，2022），可进行高职阶段行业英语的双语词典编纂。以酒店行业英语的双语词典编纂为例，分为以下四个步骤：1）使用GPT-5生成高职酒店管理专业学生所需掌握的常用词表（500个词）；2）对词表基于应用场景（如客房服务、餐饮服务等）进行聚类；3）以词表中的concierge一词为例，基于GPT-5生成该词汇的常用词组表达、难度适中的例句及其对应翻译；4）基于通义万相大模型为concierge一词生成对应配图。

(15) 提示语 (GPT-5):

Please generate 500 frequently used vocabulary words in hotel industry scenarios for vocational college students majoring in hotel management.

Please cluster the 500 words in the hotel industry into categories such as general scenarios, room services, dining services, front desk services, recreational services, business center, and gift services.

Please list 10 common expressions (bigram, trigram, four-gram etc.) with the word “concierge” in the context of hotel management. Please output the results and the expressions’ translation in a table format.

Please create 5 example sentences for the word “concierge” in the context of hotel management. As it’s targeted at Chinese vocational college EFL students, keep the sentence difficulty moderate. Please output the sentences and their translation in a table format.

提示语（通义万相）：中国宾馆前台礼宾

结果：

Concierge 礼宾

Concierge service 礼宾服务

Concierge desk 礼宾台

Contact concierge 联系礼宾

...

The concierge can help you with directions.

礼宾可以帮助你找路。

I’ll ask the concierge for restaurant recommendations.

我会问礼宾餐厅推荐。

...



5 结语

从前文介绍的语料库翻译研究替代和升级技术来看，大语言模型在翻译研究相关的语料的采集与清洗、语料的对齐与转换、常用标注、翻译语言特征的统计分析、术语库建设、翻译质量评估、双语词典编纂等方面，都具备全面处理语料库翻译研究议题的能力，甚至还具有提升语料库翻译研究方法论的潜能。

不难想象在大语言模型技术持续进步的背景下，语料库翻译研究一定会迎来全新发展。包括此前只能小规模开展的翻译策略相关的研究以及多模态翻译相关的语料库研究，都有可能发挥突出作用，极大提升翻译研究的综合水平。

当然，大语言模型大批量处理语料库的局限、数据分析结果的准确性问题、潜在的数据伦理问题，也值得在后续运用大模型中加以重点关注。

参考文献

- Baker M, 1993. Corpus linguistics and translation studies: implications and applications[C]//Baker M, Francis G, Tognini-Bonelli E. Text and technology. Amsterdam: John Benjamins: 233-250.
- Baker M, 1995. Corpora in translation studies: an overview and some suggestions for future research[J]. Target, 7(2): 223-243.
- Baker M, 1996. Corpus-based translation studies: the challenges that lie ahead[C]//Somers H. Terminology, LSP and translation. Amsterdam: John Benjamins: 175-186.
- Baker M, 2000. Towards a methodology for investigating the style of a literary translator[J]. Target, 12 (2): 241-266.

- House J, 2014. Translation quality assessment: past and present[C]//House J. Translation: a multidisciplinary approach. London: Palgrave Macmillan: 241-264.
- Newmark P, 1981. Approaches to translation[M]. Oxford: Pergamon Press.
- Schäffner C, 2004. Metaphor and translation: some implications of a cognitive approach[J]. Journal of Pragmatics, 36(7): 1253-1269.
- 冯志伟, 张灯柯, 2023. GPT与语言研究[J]. 外语电化教学 (2): 3-11.
- 许家金, 2020. 多因素语境共选: 语料库语言学新进展[J]. 外语与外语教学 (3): 1-10.
- 许家金, 2022. 高职阶段行业英语词汇表编制与“四用”教学原则的实施[J]. 外语教育研究前沿 (1): 43-49.
- 许家金, 2023. 后经典时代语料库研究方法及其理论启示[J]. 外语教学与研究 (3): 442-454.
- 许家金, 赵冲, 孙铭辰, 2024. 大语言模型的外语教学与研究应用[M]. 北京: 外语教学与研究出版社.
- 许家金, 赵冲, 孙铭辰, 2025. 大语言模型的外语教学与研究应用[M]. 2版. 北京: 外语教学与研究出版社.
- 袁毓林, 2023. 人工智能大飞跃背景下的语言学理论思考[J]. 语言战略研究 (4): 7-18.
- 张威, 2015. 中国口译学习者语料库的口译策略标注: 方法与意义[J]. 外国语 (5): 63-73.
- 郑德虎, 2016. 中国文化走出去与文化负载词的翻译[J]. 上海翻译 (2): 53-56.

通信地址

100089 北京市 北京外国语大学中国外语与教育研究中心

(责任编辑: 刘若冰)