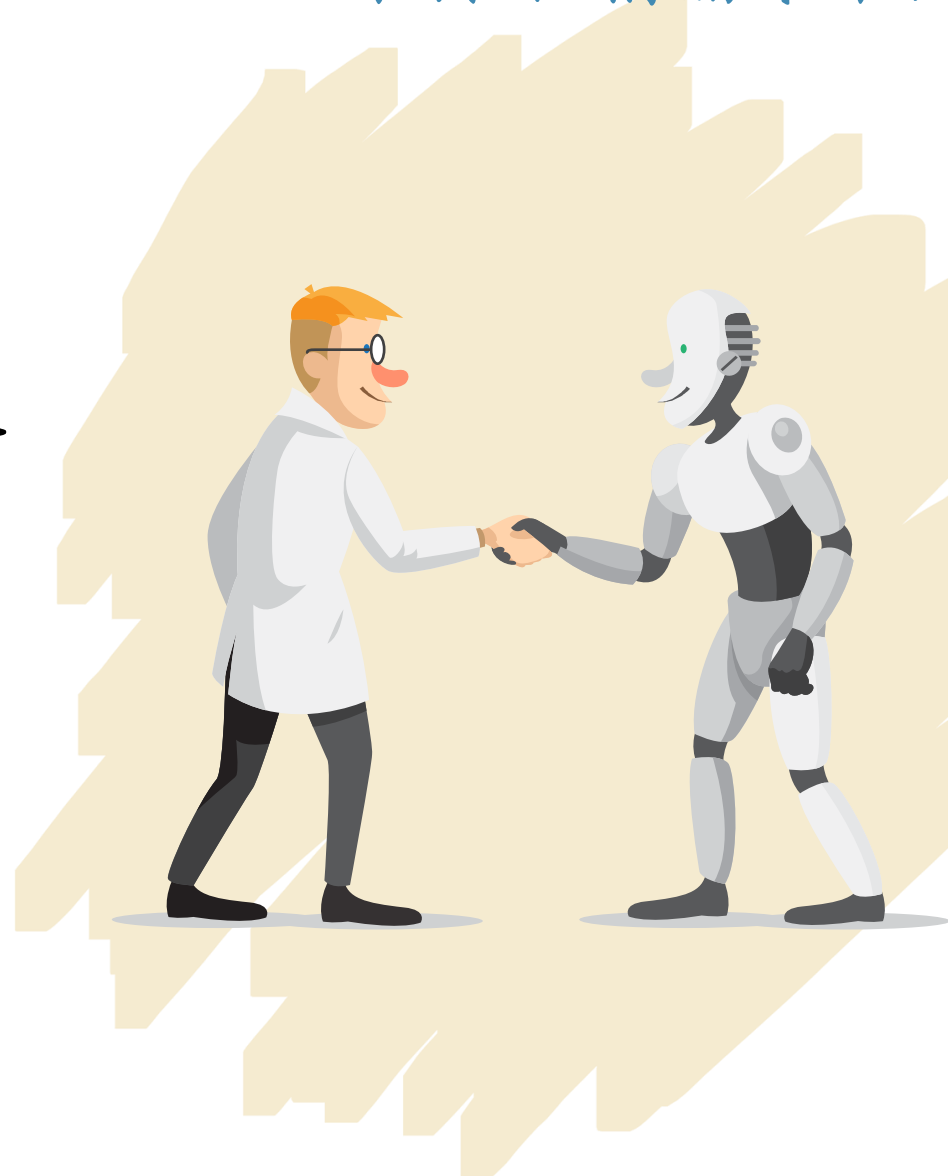


第5讲 语料库工具及 语料库创建入门

黄婕

2025-10-17 (3班)

2025-10-21 (4班)



任务场景

为了完成一篇关于人工智能的科技文章的翻译，你需要搜索在线资料，然后使用语料库工具来提高翻译效率。具体包括：

- 1) 建立一个双语的语料库
- 2) 筛选出其中的术语，生成术语表
- 3) 基于双语的语料，生成双语翻译记忆库

请思考：你需要进行什么操作？过程中用到哪些工具？

本课学习的软件

语料库工具

Sketch Engine

强大的云端语料库工具

AntConc 4.1.1

https://pan.baidu.com/s/1cdPOFUbrCwx2BxPYv-g_oQ?pwd=wcjh

提取码: wcjh

北外ParaConc (tools)

过程工具

中文分词工具

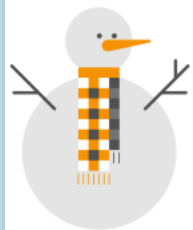
英文词性标注Tree Tagger

Abbyy Aligner 2.0

<https://pan.baidu.com/s/1erALATbT28VwbcJT9HwdTg>

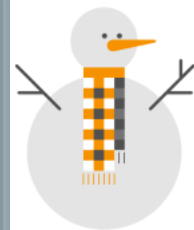
提取码: zj96

本节 内容



1. 语料库基础

2. 语料库的创建



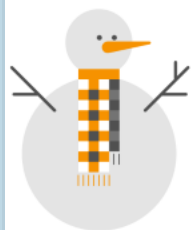
3. 术语的导出和对齐

4. 本地单语、双语语料库工具



5. 中文分词、英文词性标注工具

6. 语料库对齐工具



7. 翻译记忆库的创建和导出

番外： SketchEngine 注册 和使用指南





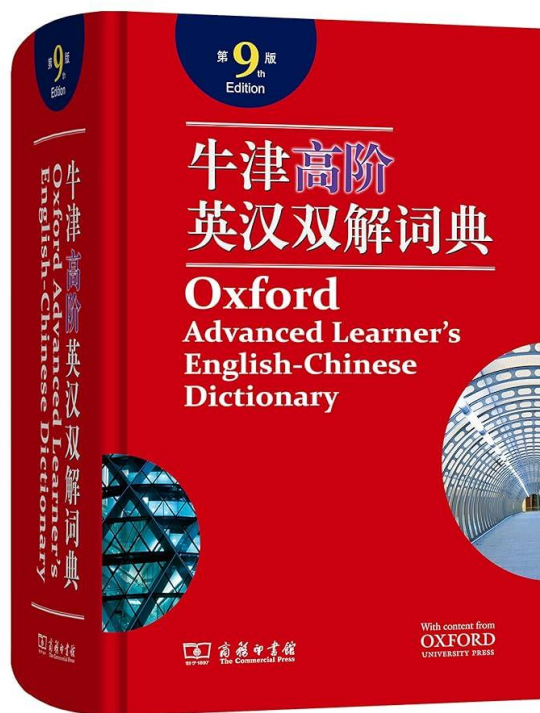
1. 语料库基础

语料库是什么？ Corpus/Corpora

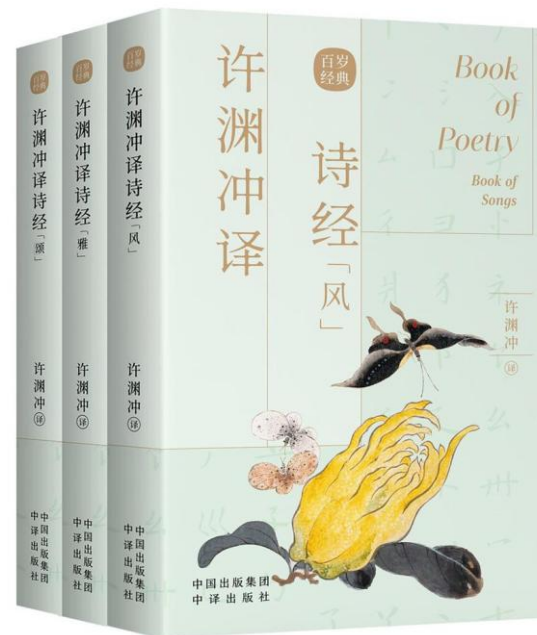
按照明确的目的和设计要求，根据语言学或翻译学原则，运用科学合理的技术方法，将一定规模的语言文本汇总而成的电子文本库。（管新潮，陶友兰，《语料库与翻译》，2017）

运用计算机技术，按照一定的语言学规则，根据特定的语言研究目的而大规模收集并存储在计算机中的真实语料，这些语料经过一定程度的标注，便于检索，可用于描述研究和实证研究。（王克非，《语料库翻译学探索》，2012）

语料库长什么样子？



双语词典



古诗词英译本

语料库长什么样子? (续)

ICS 01.080.10
A 22



中华人民共和国国家标准

GB/T 30240.9—2017

服务领域英文译写规范
第9部分: 餐饮住宿

for the use of English in public service areas—
Part 9: Accommodation and catering

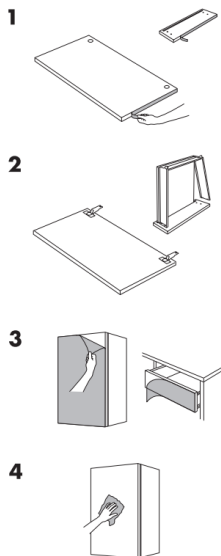
国家标准的
菜单翻译

ENGLISH

CARE AND CLEANING INSTRUCTIONS

- 1 Immediately after having removed the protective film the surface is especially sensitive to scratches.
- 2 To increase the surface's resistance, wash it with a soft cloth damped in a mild soap solution (max. 1%). Note! Do not use cleaners containing alcohol or abrasives.
- 3 Wipe the surface dry with a soft cloth.

For daily cleaning, see point 2.



中文

保养和清洁说明

- 1 去掉保护膜之后, 产品表面极易受到刮划。
- 2 为了增强产品表面的强度, 可用软布在中性皂液(最大1%)中蘸湿, 擦洗表面。注意! 不要使用含酒精或磨擦剂的清洁剂。
- 3 用软布将表面擦干。

关于日常清洁, 见第2点。

繁体中文

保養及清潔說明

- 1 移除保護膜後, 產品表面特別容易刮傷。
- 2 為增加產品表面的耐用性, 可用沾有溫和肥皂液(濃度最大1%)的軟布清洗。注意! 不可使用含酒精或磨砂劑的清潔劑。
- 3 用軟布擦乾表面。

為了解每天的清潔說明, 請看第二點。

한국어

관리 세척 방법

- 1 표면에 부착된 보호필름을 제거하면 긁힘에 유의해야 합니다.
- 2 표면 저항력을 강화하려면 순한 비눗물 (최대 1% 비누)로 적신 부드러운 천으로 표면을 닦아주세요. 주의! 알코올 또는 연마제가 들어있는 세제를 사용하지 마세요.
- 3 마른 천으로 표면의 물기를 닦아내세요.

일상적인 관리에 2번을 참고하세요.

序号	中文	英文
41	麻辣烫	Spicy Hot Pot
42	药膳	Tonic Diet 或 Herbal Cuisine
43	烧烤	Grill〔在平底锅里烤〕; Barbecue〔直接在火上烤〕
	(口味)	
44	酸	Sour
45	甜	Sweet
46	辣	Hot 或 Spicy
47	三分熟	Rare
48	五分熟	Medium
49	七分熟	Medium Well
50	十分熟〔全熟〕	Well Done
	(主食)	
51	米饭	Rice
52	面条	Noodles
53	拉面	Lamian Noodles
54	刀削面	Daoxiao Noodles
55	米线	Rice Noodles
56	馄饨	Huntun 或 Wonton
57	馅饼	Pie
58	熟食	Delicatessen〔也可简作 Deli〕或 Cooked Food
59	糕点	Cakes and Pastries

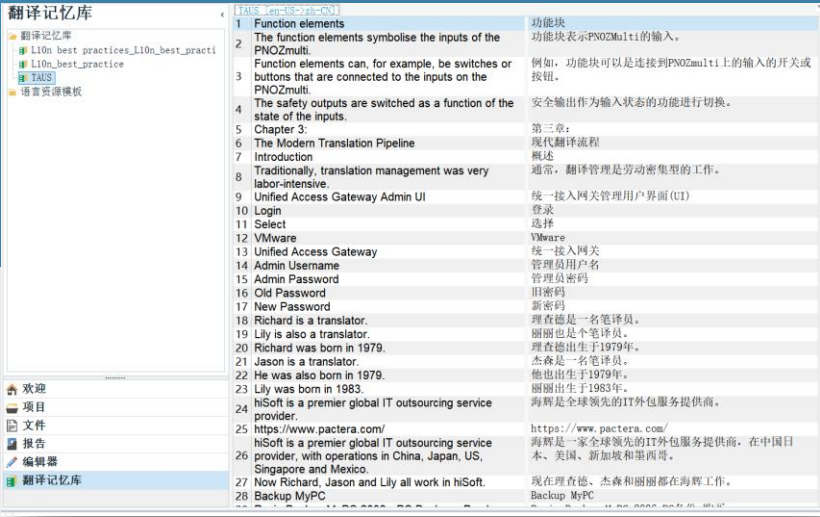
产品说明书

语料库的形成过程

纸质笔记、
资料

数据库

专业 TM



Excel 表格

CAT 工具

英文术语	标准译法	禁用译法	备注	首次出现
Dashboard	仪表板	控制面板/看板	统一使用"仪表板"	UI主界面
Log in	登录	登入	按钮文本	登录页面
Settings	设置	设定	菜单项	主菜单
Notification	通知	提醒/通告	系统消息	消息中心

语料库是什么？（天马行空版本）



语言记忆银行：

储存各种语言现象、
表达方式和用法实
例，随时提取、比
较和借鉴



跨文化对话空
间：

不同语言、方言和
文化表达方式的交
汇点，揭示语言间
微妙的表达差异



翻译决策档案
馆：

记录翻译专家如何
处理棘手问题的智
慧结晶



术语知
识花园：

专业领域词汇和表
达的生长地，不断
更新和扩展

语料库是翻译人的
“数字记忆花园”

语料库基础：分类



常见的语料库

语言学或翻译学研究类语料库

- 布朗（Brown）美国英语语料库（F. Nelson和H. Kucera，世界最早的计算机语料库）
- 上海交通大学科技英语语料库（JDEST，上海交通大学杨惠中）
- 翻译英语语料库（世界第一个翻译语料库，英国曼彻斯特大学，Mona Baker）
- 英国国家语料库（BNC）
- 美国当代英语语料库（COCA，美国杨百翰大学Mark Davies）
- 现代汉语语料库（中国国家语言文字委员会）
- 通用汉英对应语料库（北京外国语大学，王克非）
- 两岸三地英汉科普历时平行语料库（上海交通大学）
- 英汉医学平行语料库（上海交通大学，管新潮）
- 中国法律法规汉英平行语料库（绍兴文理学院）
- 汉学文史著作英汉平行语料库（山东师范大学，徐彬）

常见的语料库

翻译实践应用类语料库

- 欧洲议会平行语料库（12亿单词，21种欧洲语言）
- TAUS Data (700亿单词，2200个语言对)
- MyMemory (<https://mymemory.translated.net/>)



语料库的功能

在教学中的应用

词法、句法查证
教学案例准备
课堂作业选材
学生作文分析
课文分析研究与课程设计

教学

在科研中的应用

词典编撰
句法、语义研究
语用研究
口语研究
语言变异/变化研究
翻译策略
翻译技巧
译者风格
.....

科研

翻译

利用翻译记忆原理，将相似度较高的句对提供给译者参考，提升翻译效率

在翻译中的应用

技术

神经网络机器翻译的翻译效果
需要借助语料库提升。

在技术（MT训练）中的应用



语料库的应用场景: 1 翻译实习与专业实践

术语管理与一致性保障:

- 从双语语料库中提取专业领域术语对照, 建立个人术语库, 确保翻译一致性

翻译方案参考:

- 查询特定表达、难句、文化特定项的处理方案, 避免“重复发明轮子”

翻译记忆库构建:

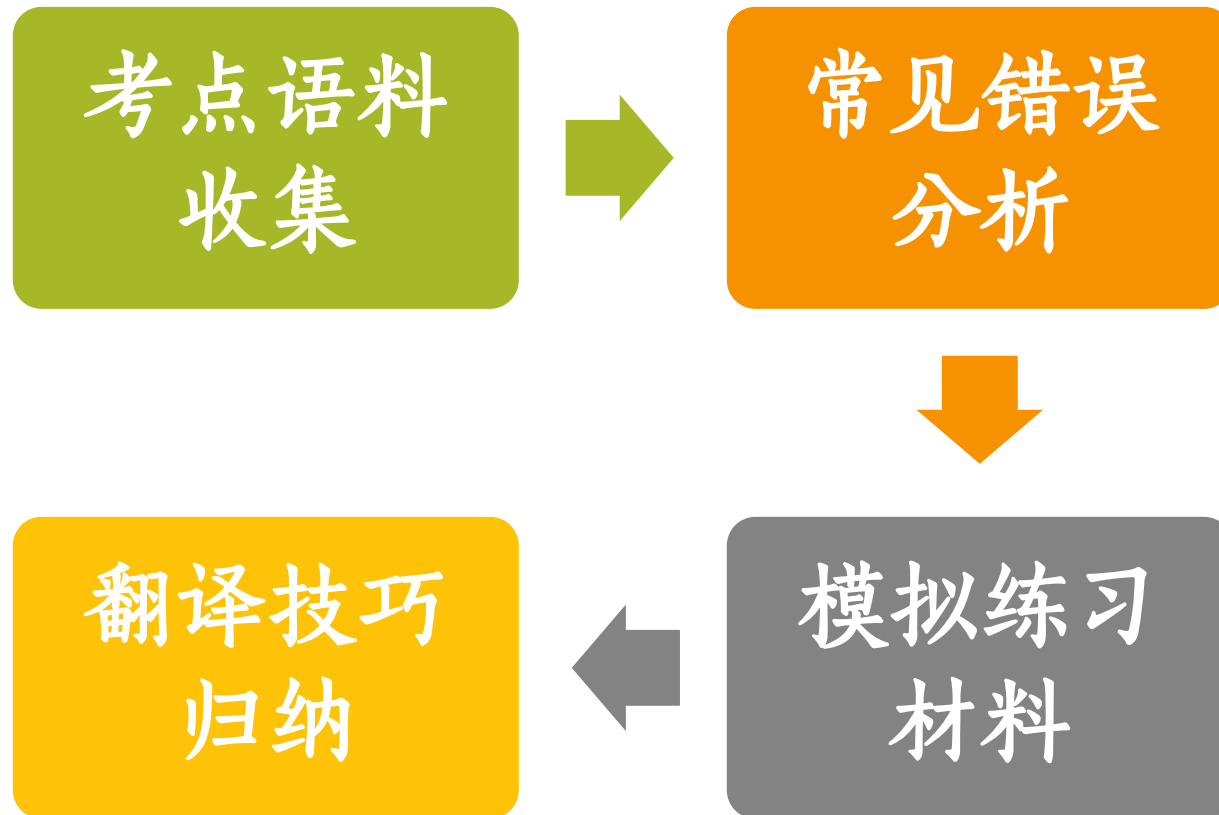
- 将高质量的双语对齐文本转化为TM (Translation Memory), 提高翻译效率

机器翻译训练数据:

- 为定制化机器翻译系统提供训练数据, 尤其是垂直领域的专业翻译



语料库的应用场景: 2 翻译资格证备考



语料库的应用场景: 3 毕业论文研究

数据驱动的翻译研究:

- 基于双语语料库进行翻译普遍特征、翻译策略或翻译规范研究

翻译质量评估:

- 对比不同译者或不同翻译方法产生的译文质量差异

跨语言对比研究:

- 探究源语言和目标语言的结构差异、表达习惯和修辞特点

翻译教学研究:

- 分析学生译文与专业译文的差异, 探索翻译教学改进方法

领域特色研究:

- 探究特定专业领域(法律、医学、技术等)的翻译特征和规律



- **KWIC/concordance search 关键词索引**: 特定词/短语的上下文使用案例
- **Word Frequency lists 词频列表**: 词或词组频繁出现的次数
- **Collocation analysis 搭配分析**: 某个词的前后搭配关系
- **WordList 词表**: 按照字母或词频顺序列出给定文本的词表, 统计其文本特征。
可以用于翻译准备阶段的术语提取
- **POS tagging 词性标注**: (Part-of-speech tagging) 对语料库内的单词按照语义和上下文进行标记, 即标注其词性
- **Keyword analysis 关键词分析**
- **Term extraction 术语提取**

双语平行语料库的应用

语料是基础

翻译风格研究

句法结构研究

跨文化传播研究、法律/商务话语研究、文化交际和形象研究、翻译研究以及翻译实践

构建语料检索、匹配、加工、利用、交易、开放的系统平台

制定专业词表、术语（单语、双语）

统计字频、词频，编写教材

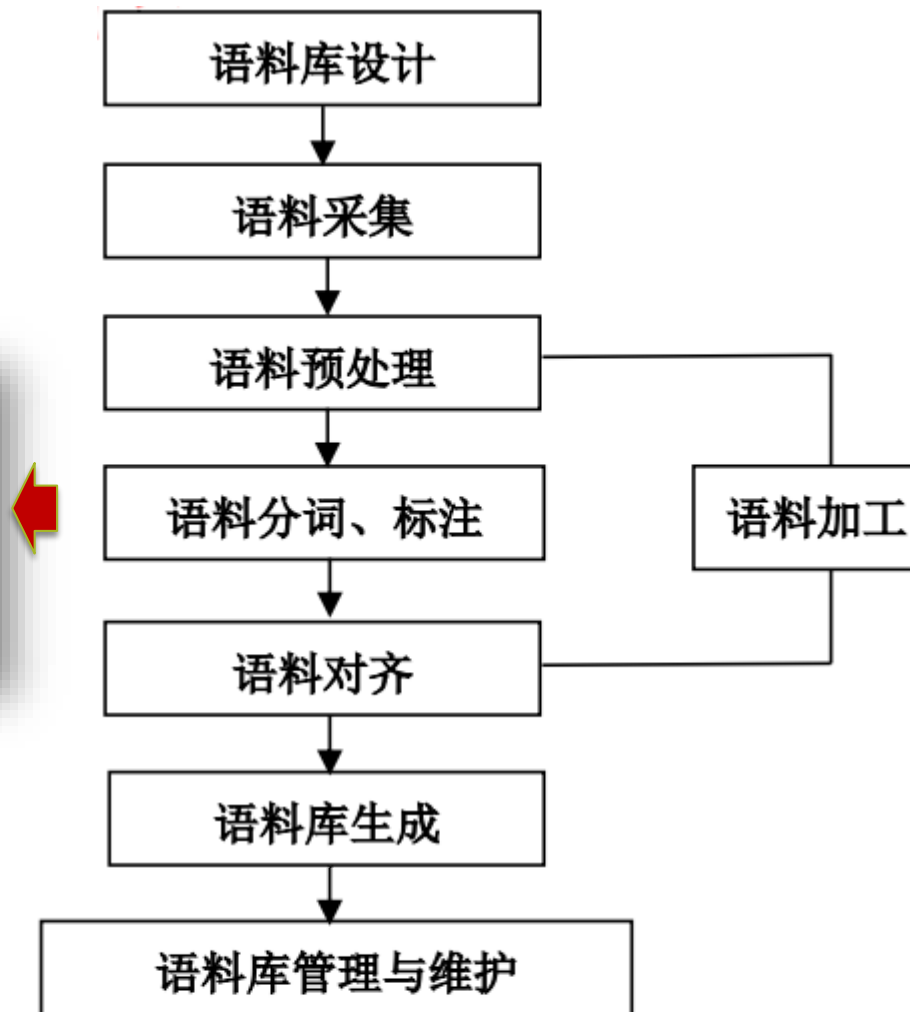
辅助写作，辅助翻译

训练机器翻译

语料库建设流程

Corpora	Li	Le	reaches	Benz	Inc
pku	李	乐	到达	奔驰	公司
msr	李乐		到达	奔驰公司	
as	李樂		到達	賓士	
cityu	李樂		到達	平治	

Table 1: Illustration of different segmentation criteria of SIGHAN bakeoff 2005.



什么是“对齐”？

- 在源语文本和目的语文本具体单位之间建立的对应关系。
- 可分为**词汇**、**语块**、**语句**、段落和篇章等层次对齐。

图 1：语料库建设流程图

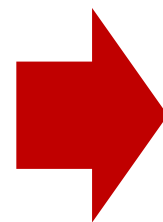
2. 语料库的创建



任务场景

为了完成一篇关于人工智能的科技文章的翻译，你需要搜索在线资料，然后使用语料库工具来提高翻译效率。具体包括：

- 1) 建立一个双语的语料库
- 2) 筛选出其中的术语，生成一份术语对齐表。



先获取语料：

1. 客户提供参考文件
2. 网络搜索可靠资料

Step 1: 获取双语专业文档——搜索WIPO数据库

https://patentscope.wipo.int/

WIPO

IP Portal

帮助 中文 知识产权门户登录

主页 > PATENTSCOPE > 检索

反馈 检索 浏览 工具 设置

PATENTSCOPE 简单检索

您可以通过PATENTSCOPE检索1.25亿专利文件，其中包含523万已公布的国际专利申请（PCT）。[具体信息](#)

国际公布第41/2025期（2025年10月9日）现可从[这里](#)查阅。下一次国际公布第42/2025期定于2025年10月16日 星期四发布。[更多信息](#)

[查看PATENTSCOPE的最新新闻和功能](#)

PATENTSCOPE在线聊天：每个星期一从13:00至17:00（CET）

字段
首页

检索内容.....
large language models

1. **118939848** 基于大语言模型和向量数据库的非结构化数据处理系统

CN 12.11.2024

国际分类 G06F 16/903 申请号 202411407258.4 申请人 BOYUN VISION [BEIJING] TECHNOLOGY CO., LTD. 发明人 CHEN JIE

本发明提供一种基于大语言模型和向量数据库的非结构化数据处理系统，系统中预先设置多个大语言模型，以非结构化数据形态呈现的私域数据输入这些大语言模型，形成多个向量矩阵以向量形式对私域数据进行向量化梳理，用户的问题语句也输入至这些大语言模型各自形成查询向量，将向量矩阵与查询向量对应求取向量距离，判断模块依照向量距离筛选出终端大语言模型，并在此输入在向量距离计算过程中获取的提示向量，最终生成问题答案。本发明初始利用多个大语言模型进行后期筛选，做到了不同大语言模型之间的扬长避短，在此过程中所生成的提示向量又能确保最终答案的精确，整个系统的运行既保证结果的精确又保证运算的快速。

2. **119443049** 文本处理方法、装置、电子设备以及存储介质

CN 14.02.2025

国际分类 G06F 40/157 申请号 202411417310.3 申请人 PENG CHENG LABORATORY 发明人 FENG XIAOCHENG

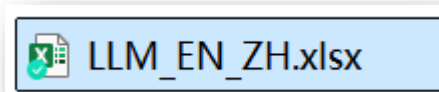
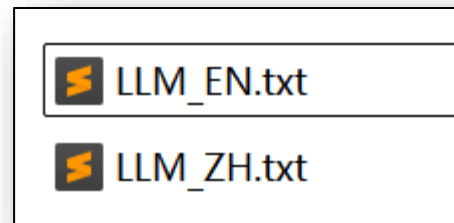
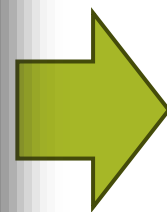
本申请实施例提供了一种文本处理方法、装置、电子设备以及存储介质，属于人工智能技术领域。该方法包括：将获取到的待处理的初始文本信息分别输入至预设的大语言模型中，得到各个大语言模型对应输出的初始特征，大语言模型包括模型结构不同的目标大语言模型和多个异构大语言模型；基于与目标大语言模型以及每个异构大语言模型的模型结构对应的各个向量转换矩阵，对任意一个初始特征进行映射处理，得到在同一向量空间下各个大语言模型对应的映射特征；基于多个不同的映射特征确定目标特征，并基于目标大语言模型对目标特征进行逆映射处理，得到初始文本信息对应的目标文本信息。本申请能够提高输出的目标文本信息的准确度。

Step 1: 获取双语专业文档——保存为本地文档

摘要

[EN] The embodiment of the invention provides a text processing method and device, electronic equipment and a storage medium, and belongs to the technical field of artificial intelligence. The method comprises the steps that obtained to-be-processed initial text information is input into preset large language models, initial features correspondingly output by all the large language models are obtained, and the large language models comprise target large language models of different model structures and a plurality of heterogeneous large language models; based on each vector conversion matrix corresponding to the target large language model and the model structure of each heterogeneous large language model, performing mapping processing on any initial feature to obtain a mapping feature corresponding to each large language model in the same vector space; and determining a target feature based on the plurality of different mapping features, and performing inverse mapping processing on the target feature based on the target large language model to obtain target text information corresponding to the initial text information. The accuracy of the output target text information can be improved.

[ZH] 本申请实施例提供了一种文本处理方法、装置、电子设备以及存储介质，属于人工智能技术领域。该方法包括：将获取到的待处理的初始文本信息分别输入至预设的大语言模型中，得到各个大语言模型对应输出的初始特征，大语言模型包括模型结构不同的目标大语言模型和多个异构大语言模型；基于与目标大语言模型以及每个异构大语言模型的模型结构对应的各个向量转换矩阵，对任意一个初始特征进行映射处理，得到在同一向量空间下各个大语言模型对应的映射特征；基于多个不同的映射特征确定目标特征，并基于目标大语言模型对目标特征进行逆映射处理，得到初始文本信息对应的目标文本信息。本申请能够提高输出的目标文本信息的准确度。



Step 2 创建语料库，导入双语文档

- <https://auth.sketchengine.eu/#login>
- 邮箱注册 sketch engine 账号 free trial, 免费使用30天

CREATE CORPUS

CREATE CORPUS > ALIGNMENT > UPLOAD DATA > COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name

LLM_EN_ZH

Corpus type

☐ Single language corpus

☒ Multilingual corpus

Storage used: 645 of 1,000,000 words (0%).

BACK


NEXT

Step 2 创建语料库，导入双语文档

CREATE CORPUS

AI-tech, Chinese Simplified


CREATE CORPUS > ALIGNMENT > UPLOAD DATA > COMPILE



Aligned documents

Parallel corpus from aligned texts.

.tmx, .xliff 2.0+, .xlf 2.0+, .xls?, .xlsx?, .zip



Non-aligned documents

Parallel corpus from texts which are not aligned but are translations of each other. Sketch Engine will align them automatically.

.doc, .docx, .htm, .html, .pdf, .txt, .zip

需要提交对齐的双语文件，比如 xlsx 表格

LEARN TO BUILD

BACK

	A	B
1	English The embodiment of the invention provides a text processing method and device, electronic equipment and a storage medium, and belongs to the technical field of artificial intelligence.	Chinese 本申请实施例提供了一种文本处理方法、装置、电子设备以及存储介质，属于人工智能技术领域。
2	The method comprises the steps that obtained to-be-processed initial text information is input into preset large language models, initial features correspondingly output by all the large language models are obtained, and the large language models comprise target large language models of different model structures and a plurality of heterogeneous large language models; based on each vector conversion matrix corresponding to the target large language model and the model structure of each heterogeneous large language model,	该方法包括：将获取到的待处理的初始文本信息分别输入至预设的大语言模型中，得到各个大语言模型对应输出的初始特征，大语言模型包括模型结构不同的目标大语言模型和多个异构大语言模型；
3	performing mapping processing on any initial feature to obtain a mapping feature corresponding to each large language model in the same vector space; and determining a target feature based on the plurality of different mapping features,	基于与目标大语言模型以及每个异构大语言模型的模型结构对应的各个向量转换矩阵，对任意一个初始特征进行映射处理，得到在同一向量空间下各个大语言模型对应的映射特征；
4	and performing inverse mapping processing on the target feature based on the target large language model to obtain target text information corresponding to the initial text information.	基于多个不同的映射特征确定目标特征，并基于目标大语言模型对目标特征进行逆映射处理，得到初始文本信息对应的目标文本信息。
5	The accuracy of the output target text information can be improved.	本申请能够提高输出的目标文本信息的准确度。
6		

以句子为单位对齐的 excel 表格

Step 2 创建语料库，导入双语文档

Build your own private corpus from texts on the web or from your own documents.

Name LLM_EN_ZH

Corpus type ☐ Single language corpus
☒ Multilingual corpus

Storage used: 360 of 1,000,000 words (0%).

BACK

NEXT

标注两个语料库的语言

CREATE CORPUS

type to search



CREATE CORPUS > ALIGNMENT > UPLOAD DATA > SETTINGS > COMPILE

Each language in the source file will be processed into a separate monolingual corpus and aligned with the corresponding corpus in the other language(s). Below you can change the corpus names and/or the automatically detected languages.

Corpus name (English) LLM_EN_ZH, English

Corpus language (English) English

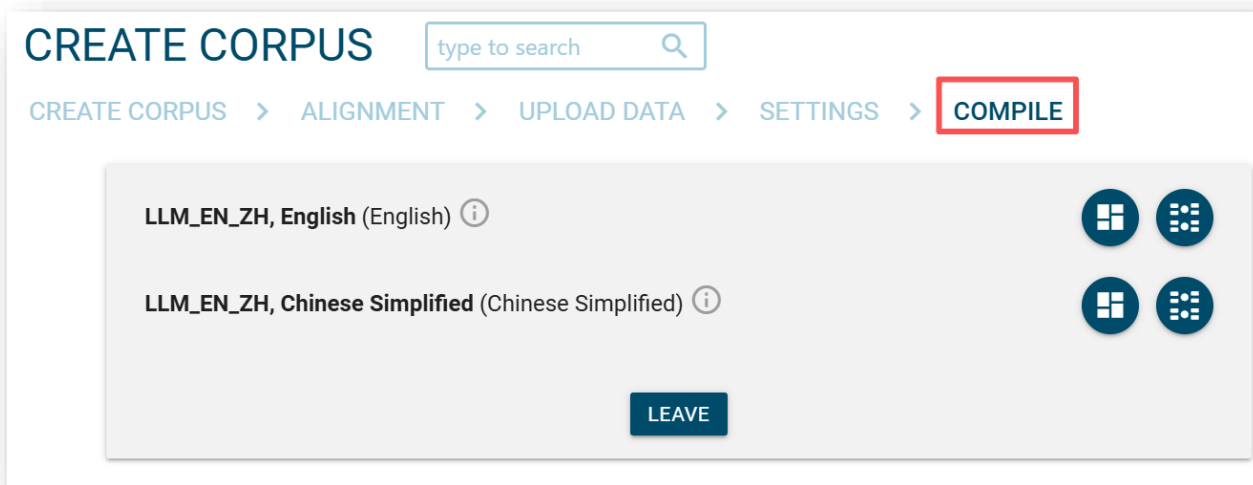
Corpus name (Chinese) LLM_EN_ZH, Chinese Si...

Corpus language (Chinese) Chinese Simplified

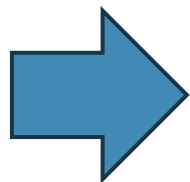
BACK

NEXT

Step 2 创建语料库，导入双语文档



创建成功



RECENTLY USED CORPORA			NEW CORPUS
LLM_EN_ZH, English	English	176	
LLM_EN_ZH, Chinese Simplified	Chinese Simplified	164	

Sketch engine 的基本功能

DASHBOARD

LLM_EN_ZH, English



LLM_EN_ZH, ENGLISH

CORPUS INFO

MANAGE CORPUS



Word Sketch

Collocations and word combinations



Word Sketch Difference

Compare collocations of two words



Thesaurus

Synonyms and similar words



Concordance

Examples of use in context



Parallel Concordance

Translation search



Wordlist

Frequency list



N-grams

Multword expressions (MWEs)



Keywords

Terminology extraction



Trends

Diachronic analysis, neologisms



Text type analysis

Statistics of the whole corpus



OneClick Dictionary

Automatic dictionary drafting



Bilingual terms

Bilingual terminology extraction

- Word sketch 词汇素描:

- 展示一个单词的语法和搭配行为，包括常见的修饰词、宾语、主语等。

- Thesaurus 同义词:

- 用于查找同义词，特别适用于写作时遇到词穷的情况。

- Wordlist 词频列表

- Keywords 关键词术语表

Sketch engine 的基本功能

- Parallel Concordance 双语对应检索

DASHBOARD LLM_ZH

LLM_ZH

Word Sketch
Collocations and word combinations

Thesaurus
Synonyms and similar words

Parallel Concordance
Translation search

N-grams
Multiword expressions (MWEs)

Trends
Diachronic analysis, neologisms

OneClick Dictionary
Automatic dictionary drafting

CORPUS INFO MANAGE CORPUS

Word Sketch Difference
Compare collocations of two words

Concordance
Examples of use in context

Wordlist
Frequency list

Keywords
Terminology extraction

Text type analysis
Statistics of the whole corpus

Bilingual terms
Bilingual terminology extraction

RECENTLY USED CORPORA

LLM_ZH	Chinese Simplified	127
LLM_EN	English	158

NEW CORPUS

ParlaTalk corpora
22 corpora of parliamentary debates in 20 languages updated.

OPEN CORPORA

可双语检索的部分

PARALLEL CONCORDANCE LLM_EN_ZH, Chinese Simplified

simple 映射 4
22,346.37 per million tokens 2.2%

doc#0

<s> 基于与目标大语言模型以及每个异构大语言模型的模型结构对应的各个向量转换矩阵,对任意一个初始特征进行映射处理,得到在同一向量空间下各个大语言模型对应的映射特征;</s>

<s> based on each vector conversion matrix corresponding to the target large language model and the model structure of each heterogeneous large language model, performing mapping processing on any initial feature to obtain a mapping feature corresponding to each large language model in the same vector space; and determining a target feature based on the plurality of different mapping features, </s>

doc#0

<s> 基于与目标大语言模型以及每个异构大语言模型的模型结构对应的各个向量转换矩阵,对任意一个初始特征进行映射处理,得到在同一向量空间下各个大语言模型对应的映射特征;</s>

<s> based on each vector conversion matrix corresponding to the target large language model and the model structure of each heterogeneous large language model, performing mapping processing on any initial feature to obtain a mapping feature corresponding to each large language model in the same vector space; and determining a target feature based on the plurality of different mapping features, </s>

Sketch engine: 关键词、词频 (自动词性标注)

KEYWORDS

LLM_EN_ZH, English



SINGLE-WORDS ✓

MULTI-WORD TERMS ✓



reference corpus: English Web 2021 (enTenTen21)

Lemma		Lemma	
1 to-be-processed	...	11 processing	...
2 heterogeneous	...	12 initial	...
3 correspondingly	...	13 comprise	...
4 plurality	...	14 target	...
5 mapping	...	15 invention	...
6 inverse	...	16 language	...
7 preset	...	17 matrix	...
8 corresponding	...	18 obtain	...
9 vector	...	19 artificial	...
10 embodiment	...	20 accuracy	...

WORDLIST

LLM_EN_ZH, English



word (73 items | 187 total frequency)

Word	Frequency ? ↓	Word	Frequency ? ↓
1 the	16 ...	14 initial	4 ...
2 language	9 ...	15 mapping	4 ...
3 large	9 ...	16 feature	4 ...
4 and	8 ...	17 information	4 ...
5 target	7 ...	18 corresponding	3 ...
6 of	7 ...	19 .	3 ...
7 model	6 ...	20 based	3 ...
8 to	6 ...	21 each	3 ...
9 ,	6 ...	22 processing	3 ...
10 text	5 ...	23 different	2 ...
11 a	5 ...	24 method	2 ...
12 models	5 ...	25 obtained	2 ...
13 on	5 ...	26 obtain	2 ...



3. 术语的导出和对齐

Step 3 导出术语、导出双语记忆库文件

DASHBOARD

LLM_EN_ZH, English



点 Manage corpus

LLM_EN_ZH, ENGLISH



Word Sketch

Collocations and word combinations



Word Sketch Difference

Compare collocations of two words



Thesaurus

Synonyms and similar words



Concordance

Examples of use in context

CORPUS INFO

MANAGE CORPUS

Download corpus



Plain text

Without part-of-speech tags
and lemmas



Vertical

One token per line with part-of-
speech tags and lemmas



TMX

For aligned multilingual
corpora

More settings ▾

CLOSE

Step 3 导出术语、导出双语记忆库文件

vert 文件：
术语的词性标注

```
llm_en_zh_english.vert x
1 <doc id="file37040153" filename="LLM_EN_ZH.xlsx" parent_folder="upload">
2 <align>
3 <s>
4 The DT the-x
5 embodiment NN embodiment-n
6 of IN of-1
7 the DT the-x
8 invention NN invention-n
9 provides VVZ provide-v
10 a DT a-x
11 text NN text-n
12 processing NN processing-n
13 method NN method-n
14 and CC and-c
15 device NN device-n
16 <g/>
17 , , , -x
18 electronic JJ electronic-j
19 equipment NN equipment-n
20 and CC and-c
21 a DT a-x
22 storage NN storage-n
23 medium NN medium-n
24 <g/>
```

Step 3 导出术语、导出双语记忆库文件

tmx 文件： 双语翻译记忆库

```
llm_en_zh_english.vert x llm_en_zh_english.tmx x
1 <?xml version="1.0" encoding="UTF-8" standalone="no" ?>
2 <tmx version="1.4"><header /><body>
3 <tu>
4 <tuv xml:lang="en"><seg>The embodiment of the invention provides a text processing method and device, electronic
equipment and a storage medium, and belongs to the technical field of artificial intelligence.</seg></tuv>
5 <tuv xml:lang="zh-Hans"><seg>本 申 请 实 施 例 提 供 了 一 种 文 本 处 理 方 法 、 装 置 、 电 子 设 备 以 及 存 储 介 质 ， 属 于 人 工
智 能 技 术 领 域 。</seg></tuv>
6 </tu>
7 <tu>
8 <tuv xml:lang="en"><seg>The method comprises the steps that obtained to-be-processed initial text information is
input into preset large language models, initial features correspondingly output by all the large language models are
obtained, and the large language models comprise target large language models of different model structures and a
plurality of heterogeneous large language models;</seg></tuv>
9 <tuv xml:lang="zh-Hans"><seg>该 方 法 包 括 ： 将 获 取 到 的 待 处 理 的 初 始 文 本 信 息 分 别 输 入 至 预 设 的 大 语 言 模 型 中 ，
得 到 各 个 大 语 言 模 型 对 应 输 出 的 初 始 特 征 ， 大 语 言 模 型 包 括 模 型 结 构 不 同 的 目 标 大 语 言 模 型 和 多 个 异 构 大 语 言
模 型 ；</seg></tuv>
10 </tu>
11 <tu>
12 <tuv xml:lang="en"><seg>based on each vector conversion matrix corresponding to the target large language model and
the model structure of each heterogeneous large language model, performing mapping processing on any initial feature
to obtain a mapping feature corresponding to each large language model in the same vector space; and determining a
target feature based on the plurality of different mapping features,</seg></tuv>
13 <tuv xml:lang="zh-Hans"><seg>基 于 与 目 标 大 语 言 模 型 以 及 每 个 异 构 大 语 言 模 型 的 模 型 结 构 对 应 的 各 个 向 量 转 换
矩 阵 ， 对 任 意 一 个 初 始 特 征 进 行 映 射 处 理 ， 得 到 在 同 一 向 量 空 间 下 各 个 大 语 言 模 型 对 应 的 映 射 特 征 ；</seg></tuv>
14 </tu>
15 <tu>
16 <tuv xml:lang="en"><seg>and performing inverse mapping processing on the target feature based on the target large
language model to obtain target text information corresponding to the initial text information.</seg></tuv>
17 <tuv xml:lang="zh-Hans"><seg>基 于 多 个 不 同 的 映 射 特 征 确 定 目 标 特 征 ， 并 基 于 目 标 大 语 言 模 型 对 目 标 特 征 进 行 逆
映 射 处 理 ， 得 到 初 始 文 本 信 息 对 应 的 目 标 文 本 信 息 。</seg></tuv>
18 </tu>
```

4. 本地单语、双语语 料库工具



单语语料库工具：AntConc

<http://www.laurenceanthony.net/software/antconc/>

UTF-8编码：
通用、国际化

AntConc

File Edit Settings Help

Target Corpus
Name: temp
Files: 1
Tokens: 337042
UNcorpora.EN.txt

KWIC Plot File Cluster N-Gram Collocate Word Keyword Wordcloud

Total Hits: 98 Page Size 100 hits 1 to 98 of 98 hits

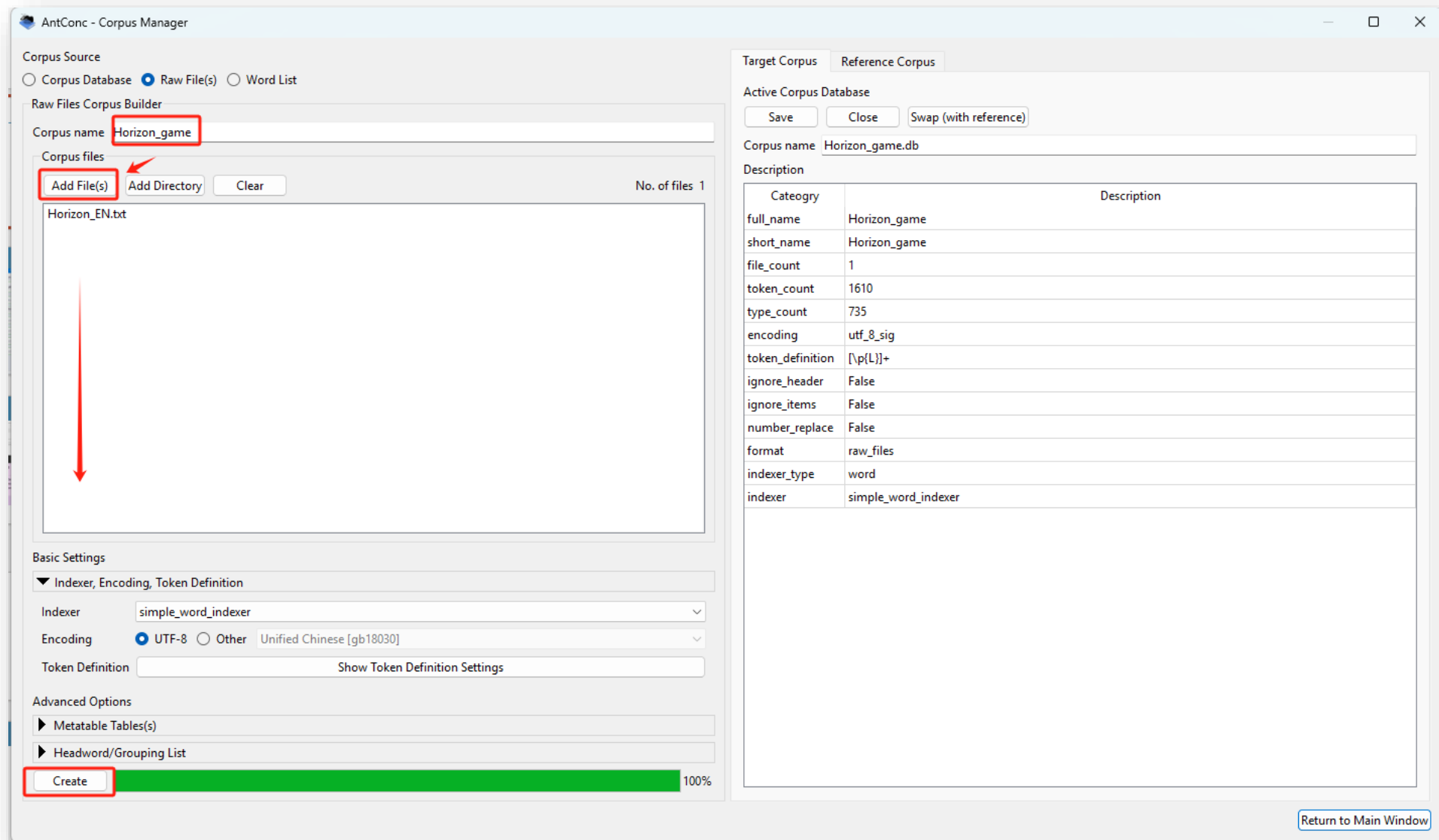
	File	Left Context	Hit	Right Context
1	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
2	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
3	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
4	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
5	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
6	UNcorpora.EN.txt	ria, Burkina Faso, Burundi, Cambodia, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
7	UNcorpora.EN.txt	ria, Burkina Faso, Burundi, Cambodia, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
8	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
9	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Cuba,
10	UNcorpora.EN.txt	Burkina Faso, Burundi, Cambodia, Cameroon, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Cuba,
11	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
12	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Cuba,
13	UNcorpora.EN.txt	Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Cuba,
14	UNcorpora.EN.txt	am, Bulgaria, Burkina Faso, Cambodia, Cameroon, Canada, Chad, Chile,	China,	Colombia, Comoros, Costa Rica, Croatia, Cuba, Cyprus, C
15	UNcorpora.EN.txt	salam, Burkina Faso, Burundi, Cambodia, Cameroon, Cape Verde, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Cuba,
16	UNcorpora.EN.txt	russalam, Bulgaria, Burkina Faso, Burundi, Cambodia, Cape Verde, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,
17	UNcorpora.EN.txt	urkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chile,	China,	Colombia, Comoros, Costa Rica, C ? te d ' Ivoire, Croatia,

Search Query ☒ Words ☐ Case ☐ Regex Results Set All hits Context Size 10 token(s)

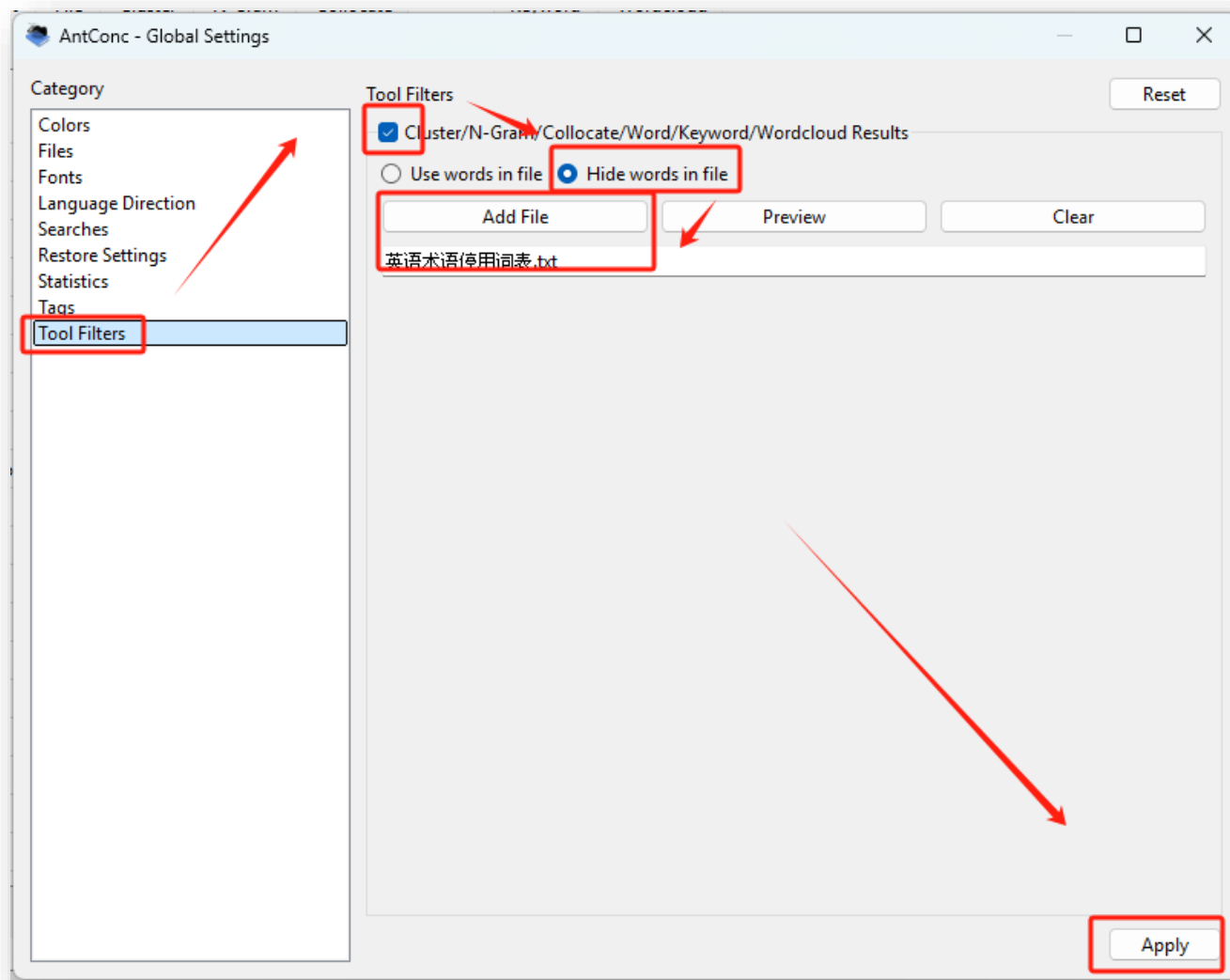
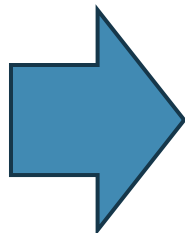
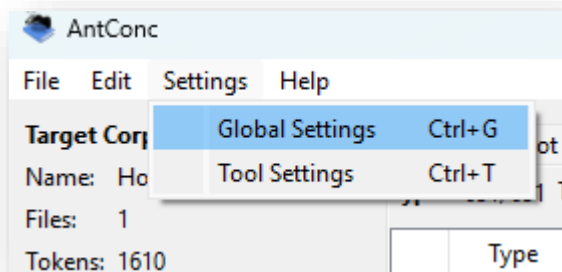
China Start ☐ Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

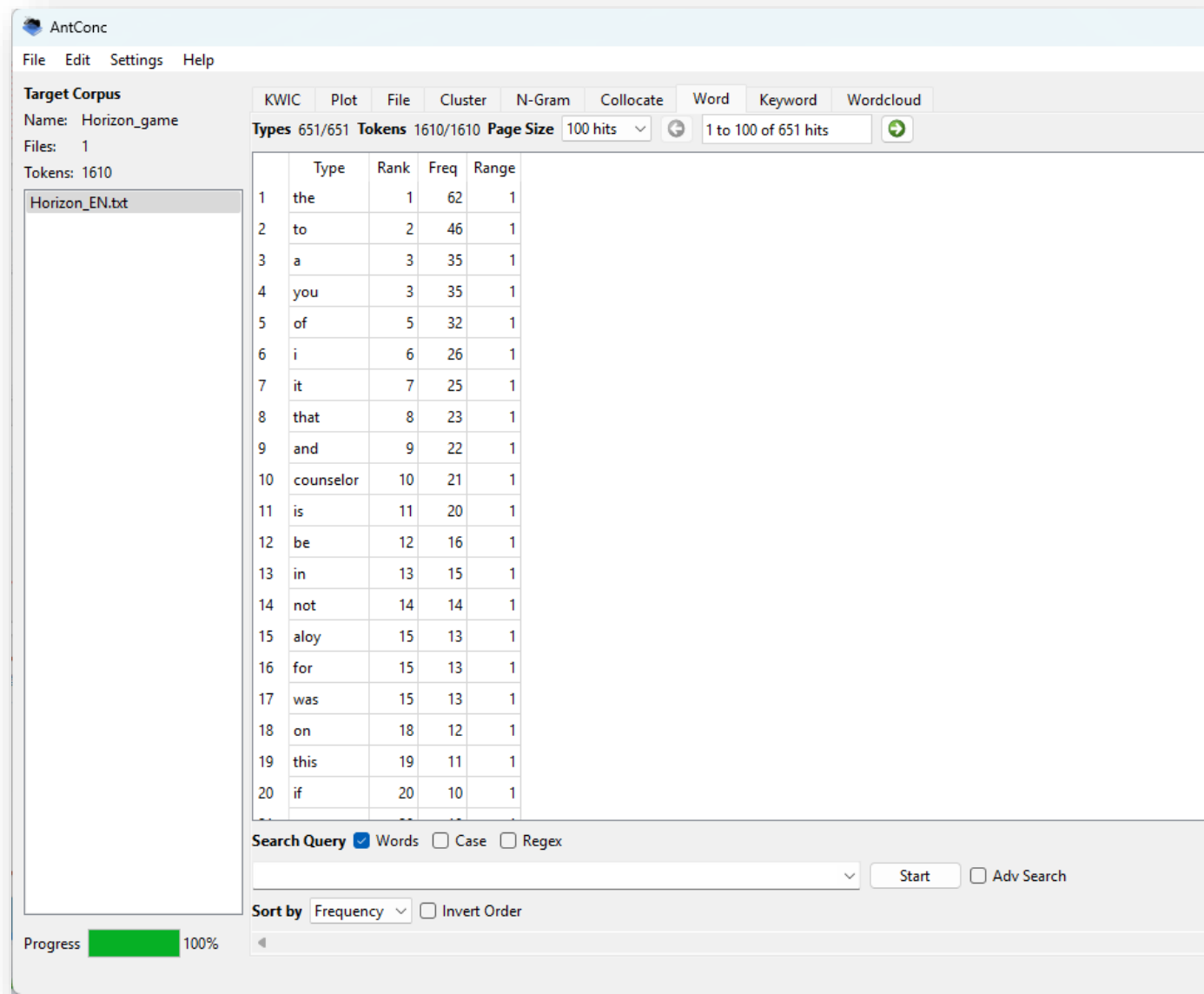
AntConc 步骤1: 创建corpus



AntConc 步骤2: 设置停用词表



AntConc 步骤3: 查询词频、检索KWIC等



The screenshot displays the AntConc application window. On the left, the 'Target Corpus' section shows 'Horizon_game' with 1 file and 1610 tokens. The main area is the 'Word' tab, which lists the top 20 words by frequency. The table includes columns for Rank, Type, Freq, and Range. The words listed are: the, to, a, you, of, i, it, that, and, counselor, is, be, in, not, aloy, for, was, on, this, and if. The bottom of the window shows a search query field with 'Words' selected, a 'Start' button, and a progress bar at 100%.

	Type	Rank	Freq	Range
1	the	1	62	1
2	to	2	46	1
3	a	3	35	1
4	you	3	35	1
5	of	5	32	1
6	i	6	26	1
7	it	7	25	1
8	that	8	23	1
9	and	9	22	1
10	counselor	10	21	1
11	is	11	20	1
12	be	12	16	1
13	in	13	15	1
14	not	14	14	1
15	aloy	15	13	1
16	for	15	13	1
17	was	15	13	1
18	on	18	12	1
19	this	19	11	1
20	if	20	10	1

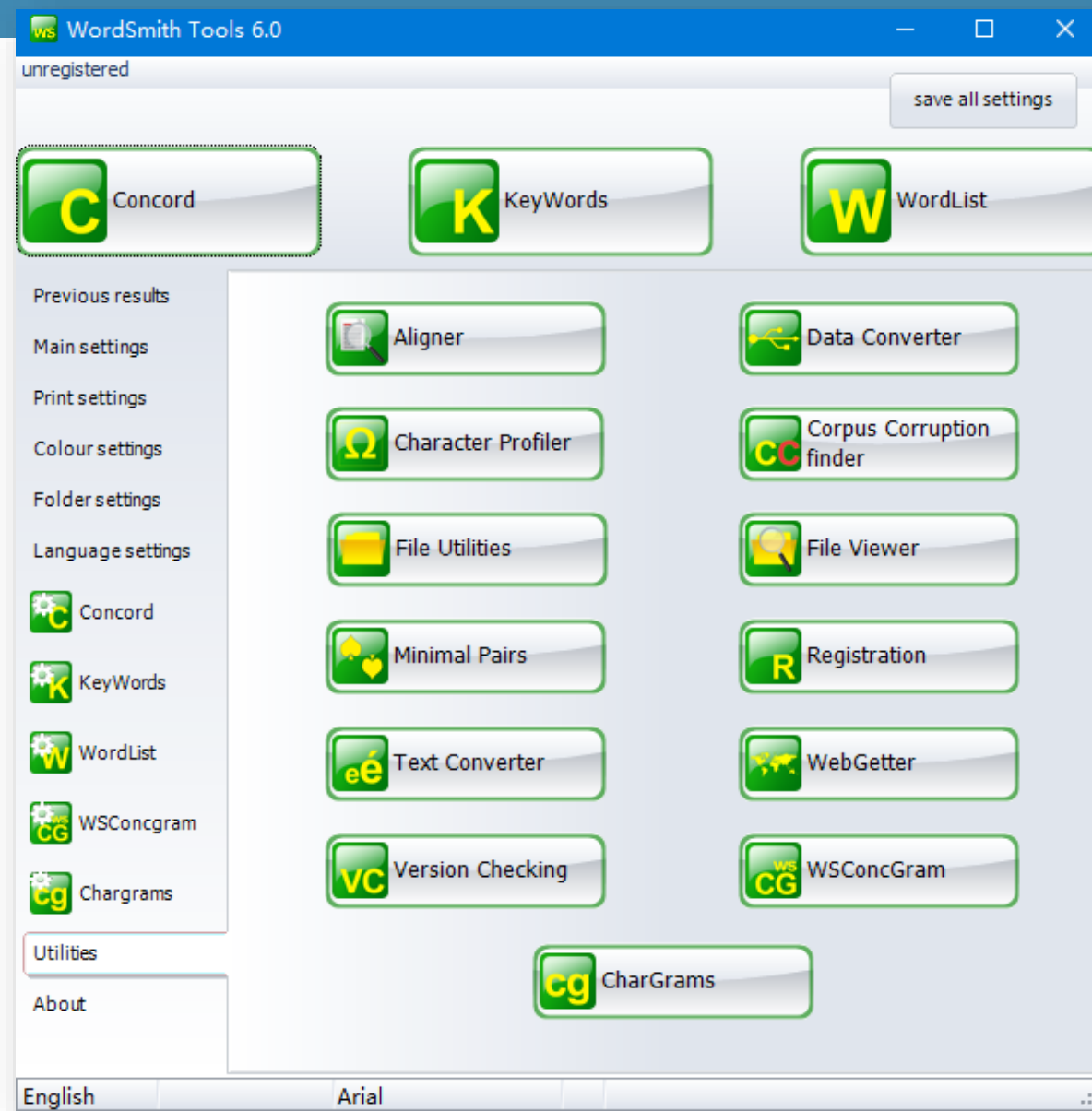
单语语料库工具：WordSmith

优点：

- 1.功能丰富：词汇检索、词频统计、共现分析等
- 2.界面友好：软件界面布局合理，操作相对简便
- 3.支持大文件处理：可以处理大型语料库。

缺点：

- 1.商业软件：需要付费使用
- 2.功能细节不尽完善：虽然功能丰富，但细节方面可能不如其他免费的AntConc。



双语语料库工具：ParaConc

ANSI编码：
需转换、有局
限性

但是可以创建
双语语料库



The screenshot shows the BFSU ParaConc 1.2 application window. It has a menu bar with 'Settings', 'Query', and 'About'. Below the menu bar, there is a search input field containing the Chinese word '海洋' (Ocean), a 'Search' button, and an 'Exclude' checkbox. The main area displays a table of search results.

No	Text
1	7. 重申违反联合国有关决议进行破坏性开采和掠夺非自治领土的 海洋 资源及其他自然资源,对这些领土的完整和繁荣构成威胁; (UNcorporata.ZH.txt, line 1386)
	7 . Reiterates that the damaging exploitation and plundering of the marine and other natural resources of the Non-Self-Governing Territories , in violation of the relevant resolutions of the United Nations , is a threat to the integrity and prosperity of those Territories ; (UNcorporata.EN.txt, line 1386)
	(d) 这些领土的 海洋 资源被非法掠夺和需要按照人民的利益利用这些资源的问题; (UNcorporata.ZH.txt, line 1431)



5. 中文分词、 英文词性标注工具

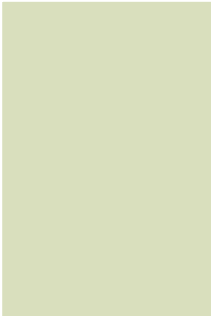
中文分词的目的

将自然语言分解为最小的单元（词语），为以下活动做准备：

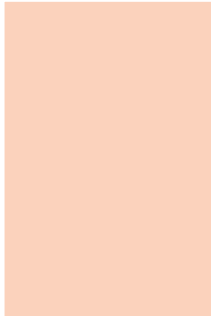
生成词典

多语语料对齐

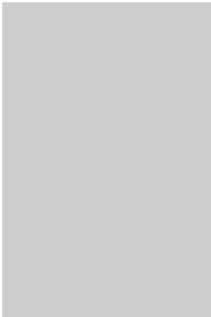
常见中文分词的工具 (Python包)




Jieba: 中文分词+词性标注
(热门工具)



SnowNLP: 中文分词+词性标注、情感分析、文本分类等

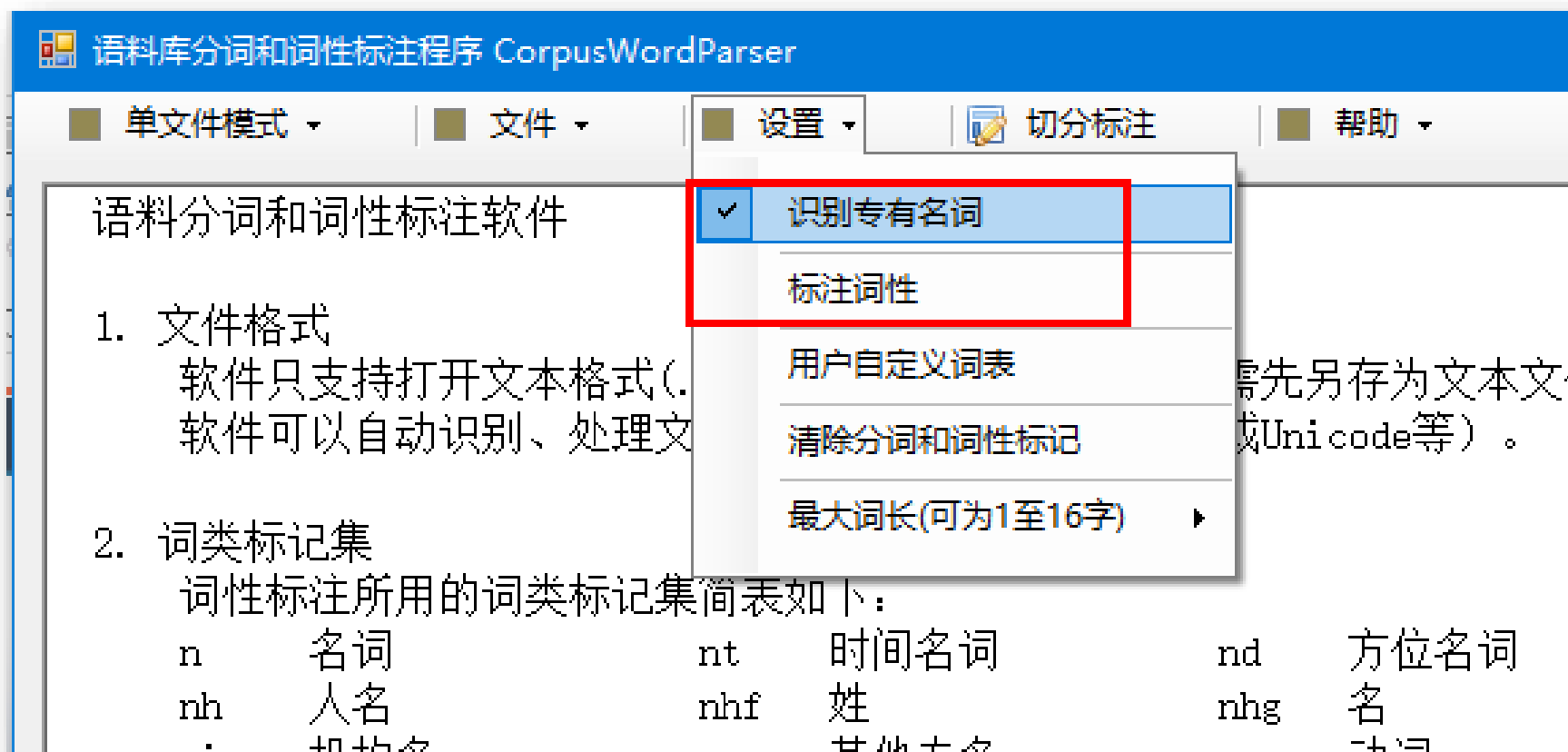


PkuSeg: 多领域分词, 个性化的预训练模型。

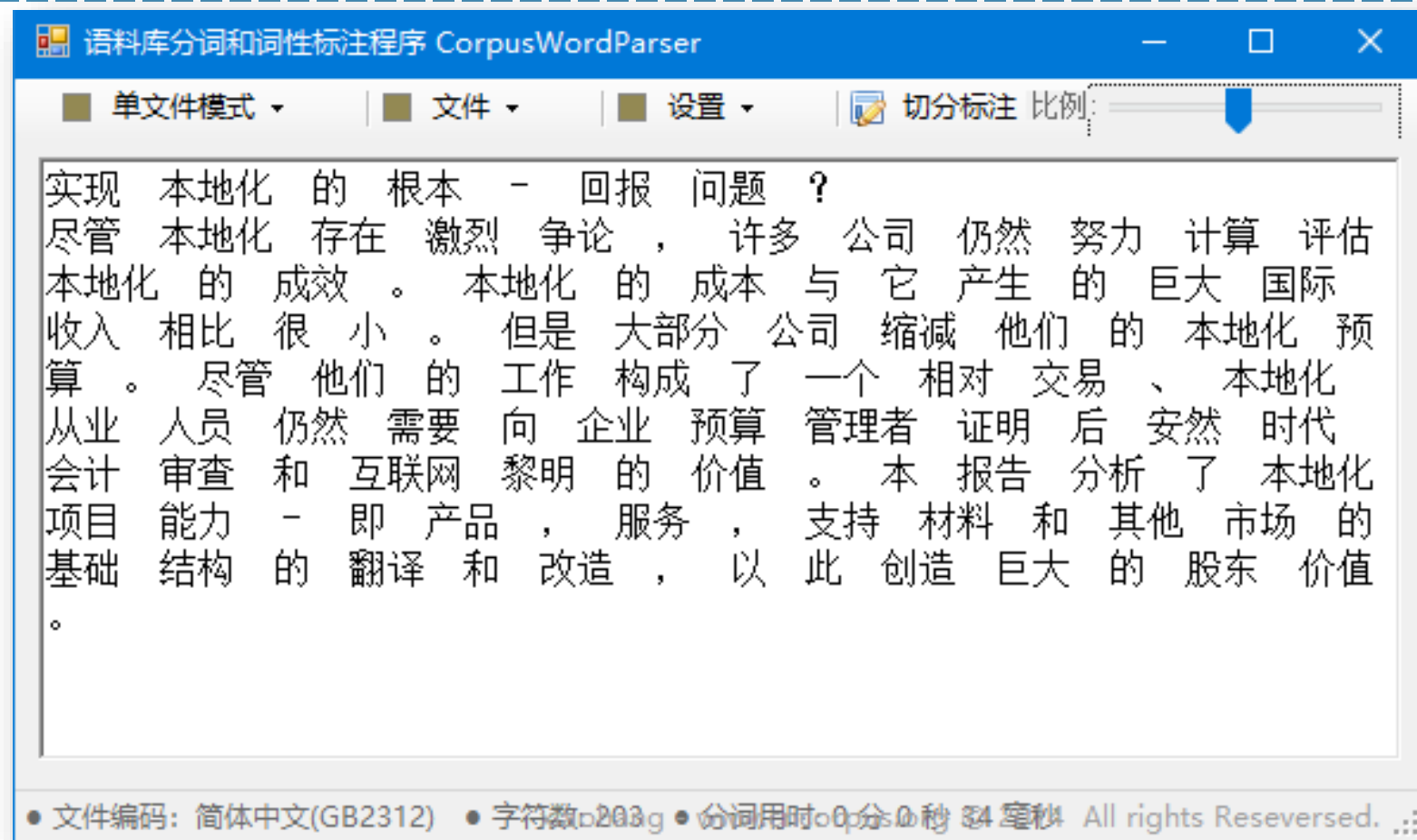


THULAC: 中文分词+词性标注。清华大学自然语言处理与社会人文计算实验室。

中文自动分词工具：CorpusWordParser.exe

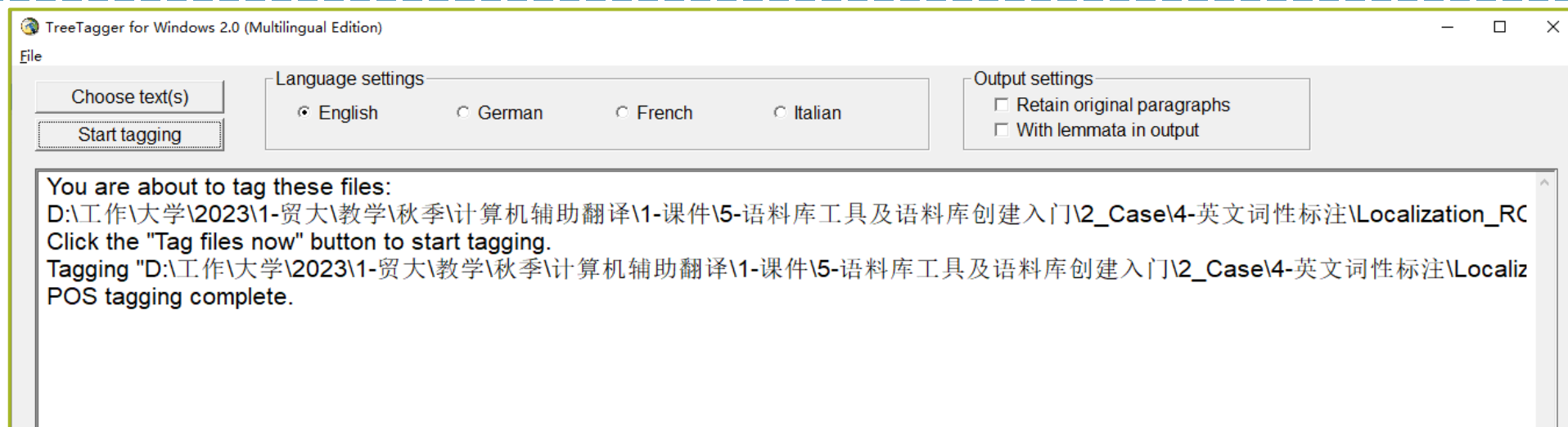


中文自动分词工具：CorpusWordParser.exe



中文语料库在线分词与标注工具，分词后的文件为ANSI编码的TXT文件

英文词性标注工具: Tree Tagger



```
Getting_VVG to_TO the_DT Bottom_NP of_IN Localization_NP -_NN  
What_WP Is_VBZ the_DT Payback_NN ?_SENT  
Despite_IN compelling_JJ arguments_NNS for_IN localization_NN ,_  
many_JJ firms_NNS are_VBP still_RB struggling_VVG with_IN  
figuring_VVG out_RP how_WRB to_TO justify_VV the_DT effort_NN  
._SENT  
The_DT cost_NN of_IN localization_NN happens_VVZ to_TO be_VB  
very_RB small_JJ compared_VVN to_TO the_DT big_JJ  
international_JJ revenue_NN it_PP can_MD help_VV generate_VV  
._SENT  
But_CC most_JJS firms_NNS shortchange_VV their_PP$  
localization_NN budgets_NNS ._SENT  
Even_RB though_IN their_PP$ work_NN constitutes_VVZ a_DT  
relative_JJ bargain_NN ,_, localization_NN practitioners_NNS  
still_RB have_VHP to_TO prove_VV their_PP$ value_NN to_TO  
corporate_JJ budgeters_NNS mindful_JJ of_IN post-Enron_NN  
accounting_NN scrutiny_NN and_CC the_DT morning-after_JJ  
internet_NN malaise_NN . SENT
```

ANSI编码的
TXT文件

英文词性标注工具：Tree Tagger

其他调用 tree
tagger 的方式

Sketch grammar ?

☒ English 3.3 for TreeTagger pipeline v2 (recommended)



☐ Universal-generic-1.0

☐ None (no word sketches)

The following steps are necessary to install the TreeTagger (see below for the [Windows version](#)). Download the files by right-clicking on the link. Then select "save file as". All files should be stored in the same directory.

1. Download the tagger package for your system ([PC-Linux](#), [Mac OS-X \(Intel\)](#), [Mac OS-X \(M1\)](#), [ARM64](#), [ARMHF](#), [ARM-Android](#), [PPC64le-Linux](#)).
If you have problems with your Linux kernel version, download this [older Linux version](#) and rename it to tree-tagger-linux-3.2.5.tar.gz.
2. Download the [tagging scripts](#) into the same directory.
3. Download the installation script [install-tagger.sh](#).
4. Download the [parameter files](#) for the languages you want to process.
5. Open a terminal window and run the installation script in the directory where you have downloaded the files:
`sh install-tagger.sh`
6. Make a test, e.g.
`echo 'Hello world!' | cmd/tree-tagger-english`
or
`echo 'Das ist ein Test.' | cmd/tagger-chunker-german`
7. You also might want to have a look at my new part-of-speech tagger [RNNTagger](#).

6. 语料库对齐工具



语料对齐

- **什么是对齐 (Alignment) ?**

- 通过比较和关联源语言文档和目标语言文档创建双语平行语料库的过程。

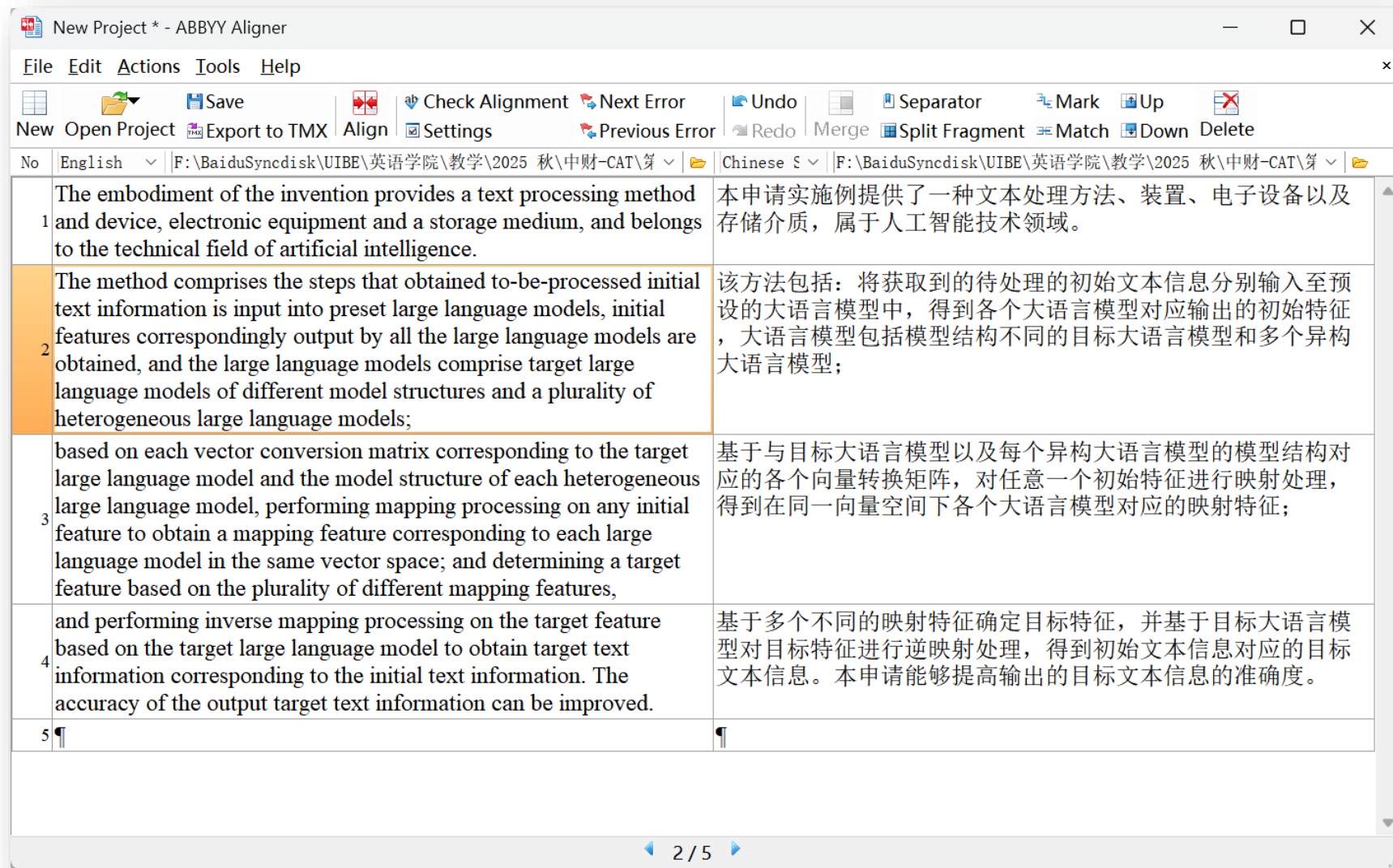
- **对齐方法**

- 使用CAT工具的对齐功能
- 使用特定工具 (Abbyy Aligner, TMXmall的在线对齐)

- **对齐注意事项**

- 原文和译文段落数相同
- 手工处理断句问题


使用Abbyy Aligner对齐














Abbyy_Aligner.zip 链接:

https://pan.baidu.com/s/1s_l6kuCn1Dig33PFC6qa7w?pwd=tarb
提取码: tarb

使用TMXMail在线对齐

 在线对齐

首页 |  公有云 |  私有云

 导出为TMX |  对齐 |  合并 |  拆分 |  上移 |  下移 |  插入 |  删除 |  帮助

No.	检测语言: 英文 Localization_ROI.doc	检测语言: 中文 Localization_ROI_ZH-CN.doc
1	Getting to the Bottom of Localization – What Is the Payback?	实现本地化的根本 – 回报问题？
2	Despite compelling arguments for localization, many firms are still struggling with figuring out how to justify the effort.	尽管本地化存在激烈争论，许多公司仍然努力计算评估本地化的成效。
3	The cost of localization happens to be very small compared to the big international revenue it can help generate.	本地化的成本与它产生的巨大国际收入相比很小。
4	But most firms shortchange their localization budgets.	但是大部分公司缩减他们的本地化预算。
5	Even though their work constitutes a relative bargain, localization practitioners still have to prove their value to corporate budgeters mindful of postEnron accounting scrutiny and the morningafter internet malaise.	尽管他们的工作构成了一个相对交易、本地化从业人员仍然需要向企业预算管理者证明后安然时代会计审查和互联网黎明前的价值。
6	This report analyzes the ability of localization projects – that is, the translation and adaptation of products, services, supporting materials, and infrastructure for other markets – to create demonstrable shareholder value.	本报告分析了本地化项目能力 – 即产品，服务，支持材料和其他市场的基础结构的翻译和改造，以此创造巨大的股东价值。

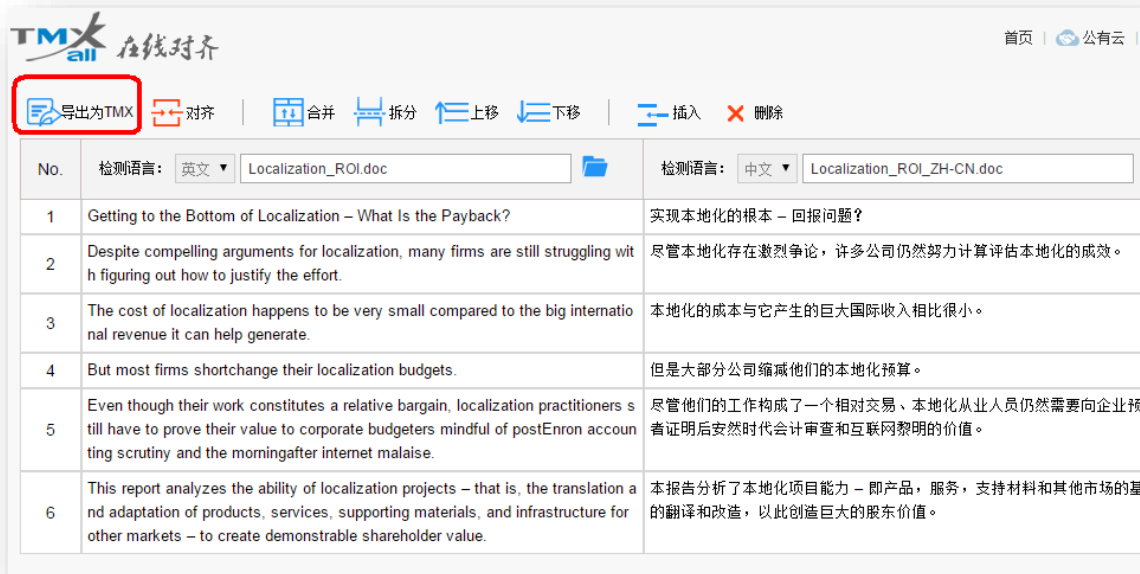
<http://www.tmxmail.com>



7. 翻译记忆库的创建和导出

翻译记忆库的导出

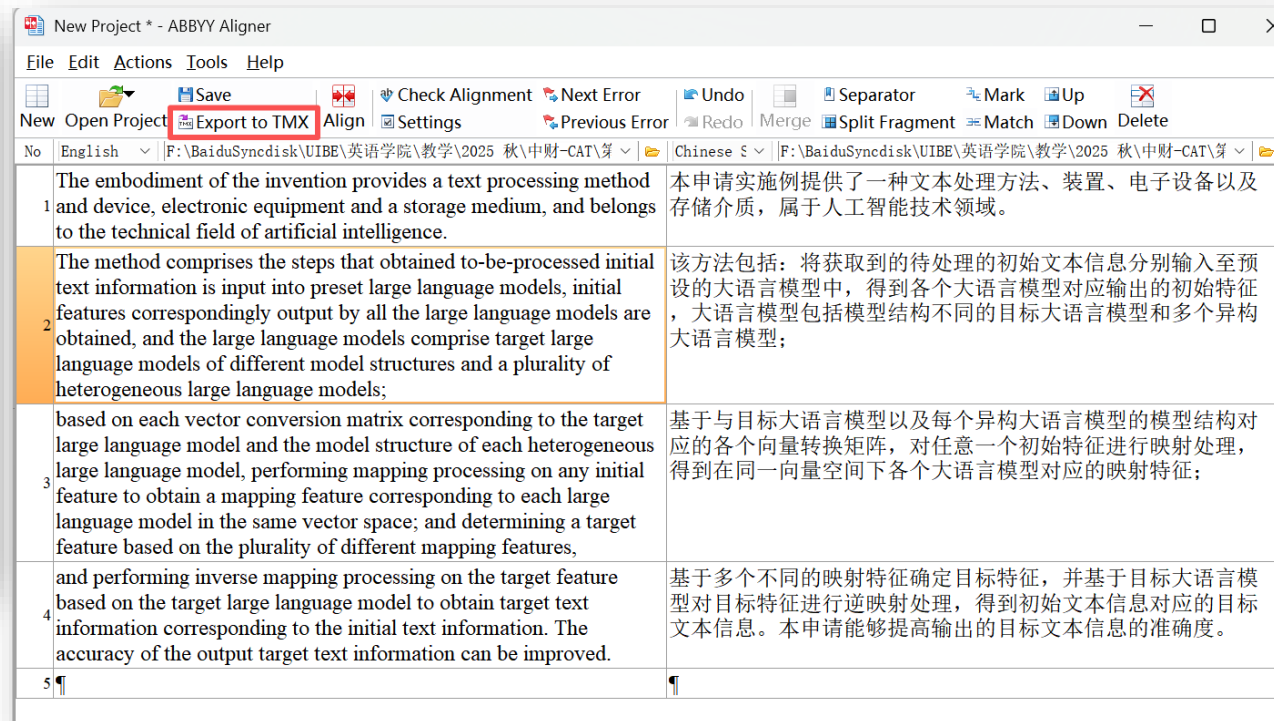
1. 从TMXMail在线对齐工具导出



The screenshot shows the TMXMail Online Alignment Tool interface. The top navigation bar includes "导出为TMX" (Export to TMX), "对齐" (Align), "合并" (Merge), "拆分" (Split), "上移" (Move Up), "下移" (Move Down), "插入" (Insert), and "删除" (Delete). The main area displays a table with two columns: "检测语言: 英文" (Source Language: English) and "检测语言: 中文" (Target Language: Chinese). The table contains six rows of aligned text segments.

No.	检测语言: 英文	检测语言: 中文
1	Getting to the Bottom of Localization – What Is the Payback?	实现本地化的根本 – 回报问题?
2	Despite compelling arguments for localization, many firms are still struggling with figuring out how to justify the effort.	尽管本地化存在激烈争论, 许多公司仍然努力计算评估本地化的成效。
3	The cost of localization happens to be very small compared to the big international revenue it can help generate.	本地化的成本与它产生的巨大国际收入相比很小。
4	But most firms shortchange their localization budgets.	但是大部分公司缩减他们的本地化预算。
5	Even though their work constitutes a relative bargain, localization practitioners still have to prove their value to corporate budgeters mindful of post-Enron accounting scrutiny and the morning-after internet malaise.	尽管他们的工作构成了一个相对交易、本地化从业人员仍然需要向企业预算者证明后安然时代会计审查和互联网黎明明的价值。
6	This report analyzes the ability of localization projects – that is, the translation and adaptation of products, services, supporting materials, and infrastructure for other markets – to create demonstrable shareholder value.	本报告分析了本地化项目能力 – 即产品, 服务, 支持材料和其他市场的翻译和改造, 以此创造巨大的股东价值。

2. 从Abbyy Aligner 导出



The screenshot shows the Abbyy Aligner New Project dialog box. The "Export to TMX" option is highlighted in the "Actions" menu. The main area displays a table with two columns: "English" and "Chinese". The table contains four rows of aligned text segments.

No	English	Chinese
1	The embodiment of the invention provides a text processing method and device, electronic equipment and a storage medium, and belongs to the technical field of artificial intelligence.	本申请实施例提供了一种文本处理方法、装置、电子设备以及存储介质, 属于人工智能技术领域。
2	The method comprises the steps that obtained to-be-processed initial text information is input into preset large language models, initial features correspondingly output by all the large language models are obtained, and the large language models comprise target large language models of different model structures and a plurality of heterogeneous large language models;	该方法包括: 将获取到的待处理的初始文本信息分别输入至预设的大语言模型中, 得到各个大语言模型对应输出的初始特征, 大语言模型包括模型结构不同的目标大语言模型和多个异构大语言模型;
3	based on each vector conversion matrix corresponding to the target large language model and the model structure of each heterogeneous large language model, performing mapping processing on any initial feature to obtain a mapping feature corresponding to each large language model in the same vector space; and determining a target feature based on the plurality of different mapping features,	基于与目标大语言模型以及每个异构大语言模型的模型结构对应的各个向量转换矩阵, 对任意一个初始特征进行映射处理, 得到在同一向量空间下各个大语言模型对应的映射特征;
4	and performing inverse mapping processing on the target feature based on the target large language model to obtain target text information corresponding to the initial text information. The accuracy of the output target text information can be improved.	基于多个不同的映射特征确定目标特征, 并基于目标大语言模型对目标特征进行逆映射处理, 得到初始文本信息对应的目标文本信息。本申请能够提高输出的目标文本信息的准确度。
5		

大模型创建翻译记忆库

将英文和中文以句子为单位对齐成标准TMX格式的翻译记忆库文件，下面第一段是英文原文，第二段是中文译文：

Getting to the Bottom of Localization – What Is the Payback?

Despite compelling arguments for localization, many firms are still struggling with figuring out how to justify the effort. The cost of localization happens to be very small compared to the big international revenue it can help generate. But most firms shortchange their localization budgets. Even though their work constitutes a relative bargain, localization practitioners still have to prove their value to corporate budgeters mindful of post-Enron accounting scrutiny and the morning-after internet malaise. This report analyzes the ability of localization projects – that is, the translation and adaptation of products, services, supporting materials, and infrastructure for other markets – to create demonstrable shareholder value.

实现本地化的根本 – 回报问题？

尽管本地化存在激烈争论，许多公司仍然努力计算评估本地化的成效。本地化的成本与它产生的巨大国际收入相比很小。但是大部分公司缩减他们的本地化预算。尽管他们的工作构成了一个相对交易、本地化从业人员仍然需要向企业预算管理者证明后安然时代会计审查和互联网黎明的价值。本报告分析了本地化项目能力 – 即产品，服务，支持材料和其他市场的基础结构的翻译和改造，以此创造巨大的股东价值。

大模型创建翻译记忆库

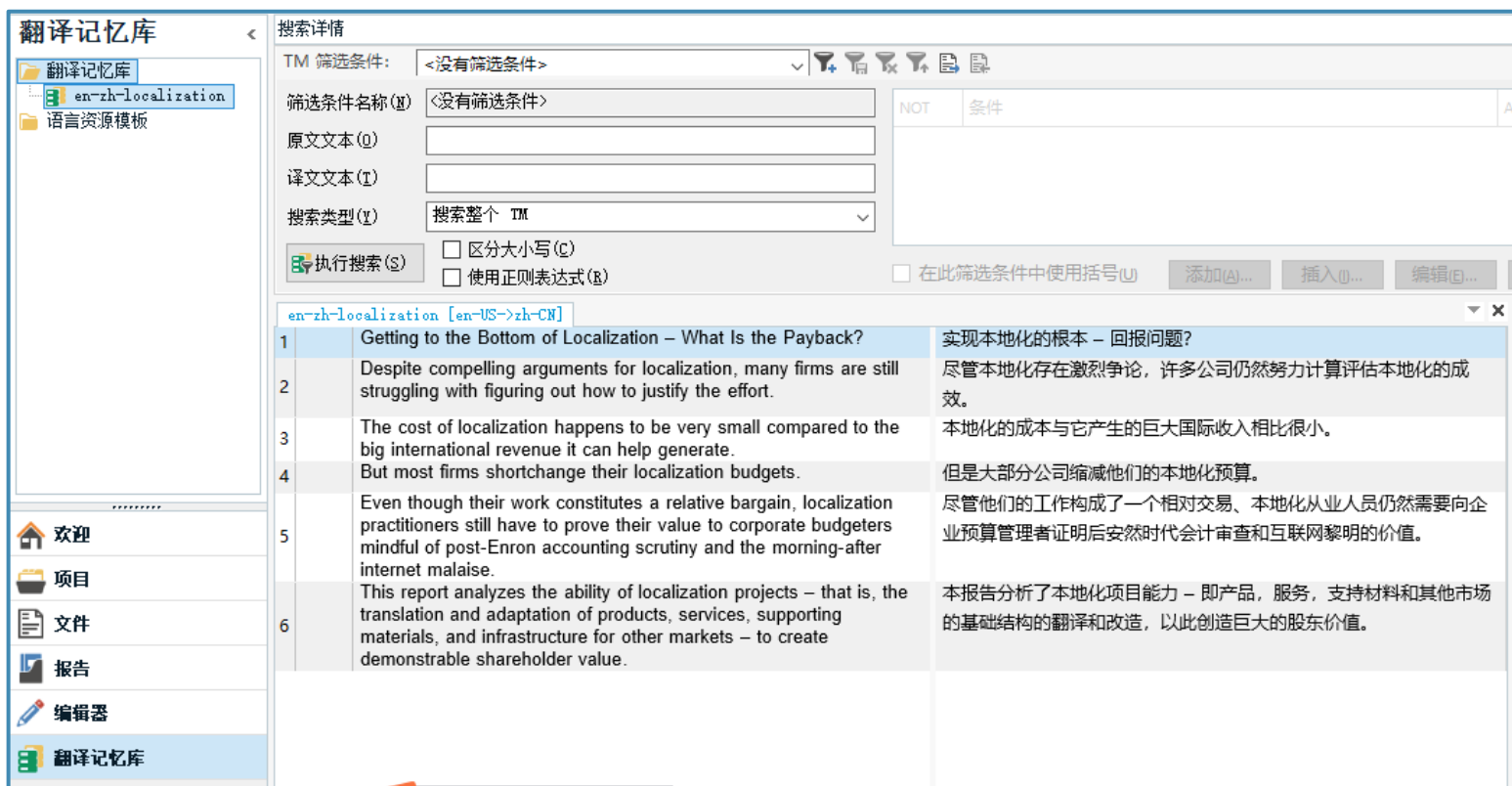
tmx ^



```
<?xml version="1.0" encoding="UTF-8"?>
<tmx version="1.4">
  <header creationtool="Manual Creation" creationtoolversion="1.0" segtype="sentence" o-tmf="unknown" adminla
  <body>
    <tu>
      <tuv xml:lang="en">
        <seg>Getting to the Bottom of Localization - What Is the Payback?</seg>
      </tuv>
      <tuv xml:lang="zh-CN">
        <seg>实现本地化的根本 - 回报问题? </seg>
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="en">
        <seg>Despite compelling arguments for localization, many firms are still struggling with figuring out
      </tuv>
      <tuv xml:lang="zh-CN">
        <seg>尽管本地化存在激烈争论, 许多公司仍然努力计算评估本地化的成效。</seg>
      </tuv>
    </tu>
    <tu>
      <tuv xml:lang="en">
        <seg>The cost of localization happens to be small compared to the big international revenue it
```

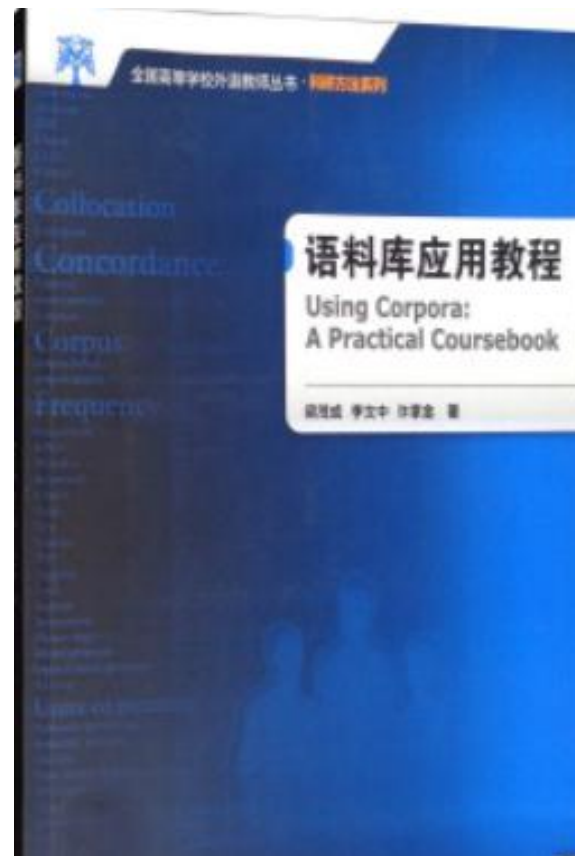
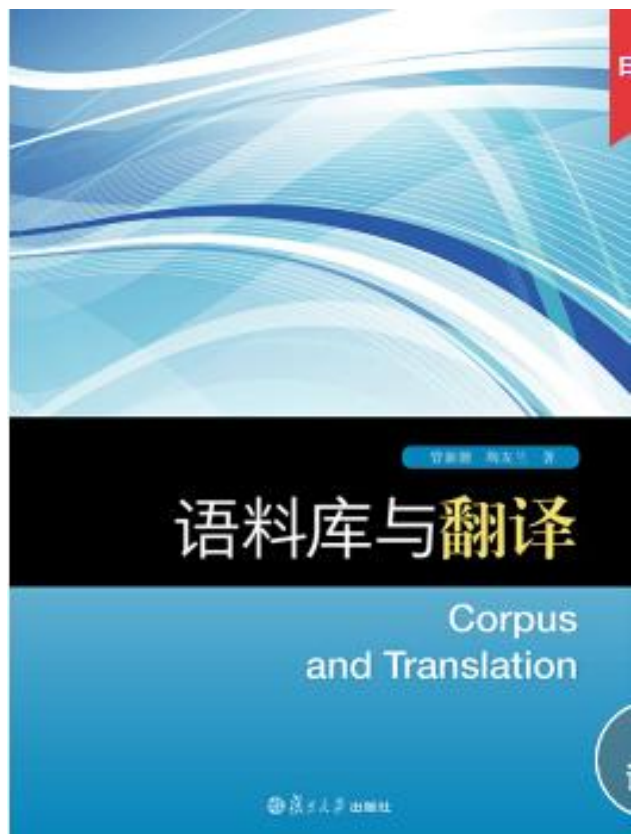
大模型创建翻译记忆库

将文本内容保存到TXT文件，修改文件扩展名为TMX，在Trados Studio新建翻译记忆库文件，导入TMX文件。



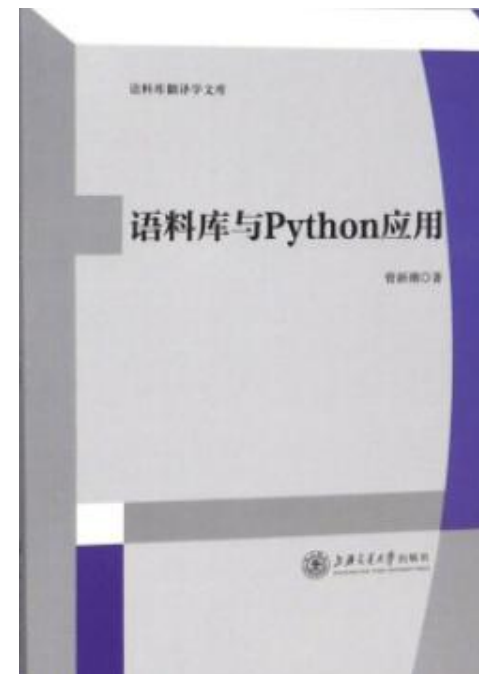
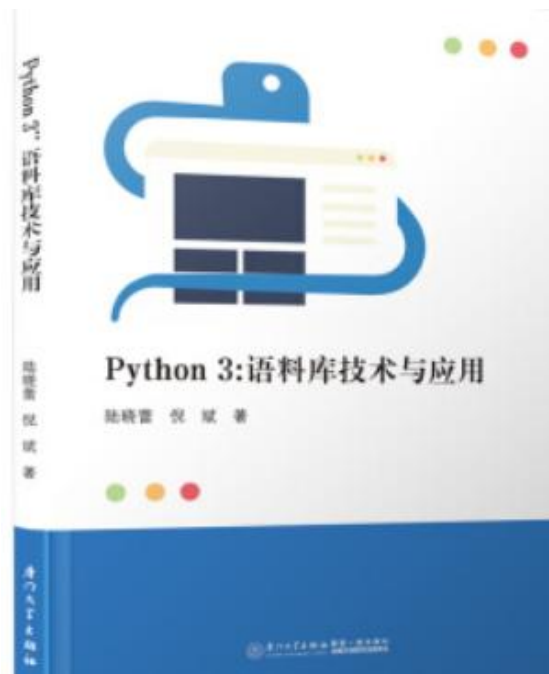
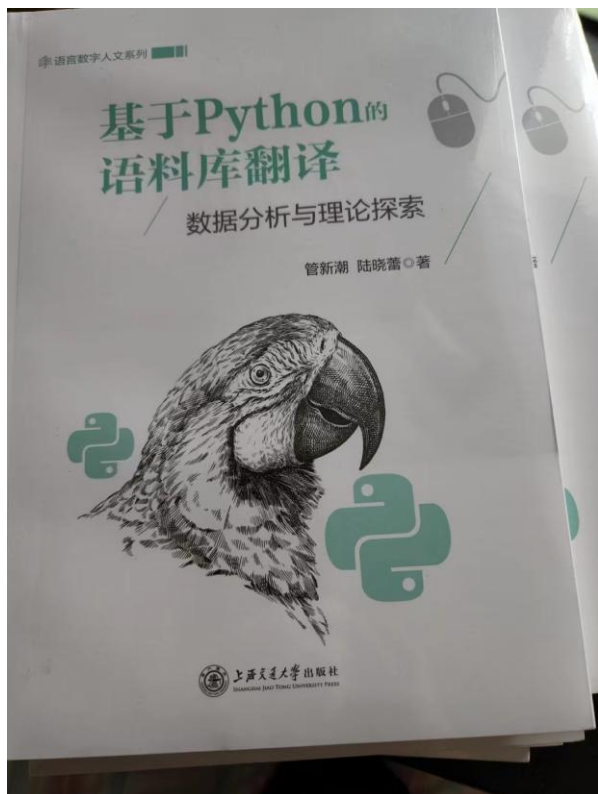
课外阅读材料

- Excel文件与TMX文件的相互转换方法
- 来源：“本地化世界”微信公众号



参考材料

- Python语料库编程





番外：
SketchEngine 注册
和使用指南




注册 30-day free trial

<https://auth.sketchengine.eu/#register>


 SKETCH
ENGINE

Jie Huang 


Sign up




Free 30-day trial
The complete functionality, 300+ corpora, 90+ languages. May contain advertising.



Individual user account
An academic or commercial use conducted by a single person.



Multi-user account
An academic or commercial use conducted by an institution.




Join a multi-user account
An access code is required.



创建语料库—create corpus

The screenshot displays the 'SELECT CORPUS' interface. On the left, a sidebar contains several icons, with the third icon (a list) highlighted by a red box and labeled with a red '1'. The main area is divided into 'BASIC' and 'ADVANCED' tabs. Under the 'BASIC' tab, the 'LANGUAGES' section shows buttons for ARABIC, ENGLISH, GERMAN, ITALIAN, and PORTUGUESE. Below these buttons, it says 'More languages'. A red '1' is placed next to the text 'Select a language and we will pick the best one for you.' On the right, a search bar labeled 'type to search' is present. Below it, a table lists various corpora. The table has three columns: a checkbox, the corpus name, the language, and the number of words. The corpora listed are: Arabic Web 2018 (arTenTen18), Chinese Web 2017 (zhTenTen17) Simplified, Dutch Web 2020 (nlTenTen20), English Web 2021 (enTenTen21) (checked), French Web 2023 (frTenTen23), German Web 2020 (deTenTen20), Hindi Web 2021 (hiTenTen21), Italian Web 2020 (itTenTen20), Japanese Web 2011 (jaTenTen11), and Korean Web 2018 (koTenTen18). At the bottom right, there is a checkbox for 'Show description', a link for 'ADVANCED SEARCH', and a button labeled 'CREATE CORPUS' which is highlighted by a red box and labeled with a red '2'. Above this button, it says '842 corpora'.

	Corpus Name	Language	Words
<input type="checkbox"/>	Arabic Web 2018 (arTenTen18)	Arabic	4,637,956,234
<input type="checkbox"/>	Chinese Web 2017 (zhTenTen17) Simplified	Chinese	13,531,331,169
<input type="checkbox"/>	Dutch Web 2020 (nlTenTen20)	Dutch	5,890,009,964
<input checked="" type="checkbox"/>	English Web 2021 (enTenTen21)	English	52,268,286,493
<input type="checkbox"/>	French Web 2023 (frTenTen23)	French	23,874,070,858
<input type="checkbox"/>	German Web 2020 (deTenTen20)	German	17,512,733,172
<input type="checkbox"/>	Hindi Web 2021 (hiTenTen21)	Hindi	792,395,313
<input type="checkbox"/>	Italian Web 2020 (itTenTen20)	Italian	12,451,734,885
<input type="checkbox"/>	Japanese Web 2011 (jaTenTen11)	Japanese	8,432,294,787
<input type="checkbox"/>	Korean Web 2018 (koTenTen18)	Korean	1,668,851,720

设置语料库属性

 **CREATE CORPUS**


English Web 2021 (enTenTen21)  

CREATE CORPUS > ADD TEXTS > COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name 2024_spring_CAT

Corpus type ☒ Single language corpus
☐ Multilingual corpus

Language English 

Description Game

Storage used: 0 of 1,000,000 words (0%)

选择语料来源——网页/本地文件

CORPUS: 2024_spring_CAT (English)

CREATE CORPUS > ADD TEXTS > COMPILE



Find texts on the web

Automatically find and download relevant texts



I have my own texts

Upload your own files (.txt, .pdf,...) or paste text

网页搜索关键词：至少**3**个

Input type

☒ Web search 

☐ URL 

Input some words and phrases that define the topic of the new corpus. Words will be randomly selected and groups of 3 will be sent to the Bing search engine. The web pages that Bing returns will be downloaded and processed into a corpus. Input between 3 and 20 words or phrases.

Hit ENTER after each one.

Select web pages to download

The selected web pages will be downloaded. Deselect those that should be skipped. A page may be removed after the download if it does not match your denylist settings, allowlist settings or size restrictions. [Check the settings now](#) (Your current selection will be lost.)

Filter

type to search

SELECT VISIBLE

DESELECT VISIBLE

EXPAND ALL

COLLAPSE ALL

✓ game terminology • game localization • video game (23/23 selected) ^

- ✓ blog.andovar.com/games-translation-ultimate-guide ↗
- ✓ ehlion.com/magazine/gaming-terminology/ ↗
- ✓ en.wikipedia.org/wiki/Glossary_of_video_game_terms ↗
- ✓ gengo.com/industry-translation/video-game-translation-services/ ↗
- ✓ j-entranslations.com/what-skills-do-i-need-to-be-a-game-translator-part-1-translation-skills/ ↗
- ✓ link.springer.com/chapter/10.1007/978-3-030-42105-2_15 ↗
- ✓ link.springer.com/chapter/10.1007/978-3-030-88292-1_3 ↗
- ✓ multiplatform.com/news/demystifying-game-terminology-in-video-game-localization-ptw-s-experience/ ↗
- ✓ research-information.bris.ac.uk/en/publications/terminology-management-in-game-localization ↗
- ✓ smartcat.com/blog/game-localization/ ↗
- ✓ academia.edu/6639017/Challenges_in_video_game_localization_An_integrated_perspective ↗

语料库创建成功

2024_spring_CAT 

user/jie.huang/corpus_2024_spring_cat • created March 29, 2024 at 5:56:47 PM

Game

MANAGE CORPUS

MANAGE SUBCORPORA

COMPARE CORPORA

TEXT TYPE ANALYSIS

GENERAL INFO

Language: English

CORPUS DESCRIPTION & BIBLIOGRAPHY

TAGSET


WORD SKETCH GRAMMAR

TERM GRAMMAR

COUNTS

Tokens	80,452
Words	61,108
Sentences	3,448
Paragraphs	859
Documents	22

LEXICON SIZES

word?	11,186
tag	62
lempos?	8,427
pos	9
lemma	7,960
lempos_lc 	8,039

COMMON TAGS

adjective	J.*
adverb	RB.?
conjunction	CC
determiner	DT
noun	N.*
numeral	CD

TEXT TYPES

TEXT TYPE ANALYSIS

<doc> (7)	22	▼
Domain name , doc.urldomain	16	≡
File ID , doc.id	22	≡
File name , doc.filename	22	≡
Folder , doc.parent_folder	1	≡
Top level domain , doc.tld	5	≡
URL , doc.url	22	≡
Website , doc.website	16	≡
<g> (0)	16,718	▼
<s> (0)	3,448	▼
<p> (0)	859	▼
<im1> (0)	2	▼
<Image> (0)	2	▼
<txt2> (0)	1	▼
<txt3> (0)	1	▼

术语列表

KEYWORDS



SINGLE-WORDS ✓

MULTI-WORD TERMS ✓



reference corpus: English Web 2021 (enTenTen21) (items: 7,271)

Lemma	Lemma	Lemma	Lemma	Lemma
1 llm ...	11 openai ...	21 generative ...	31 instructibility ...	41 eva-clip ...
2 llms ...	12 multi-modal ...	22 ai-native ...	32 vision-language ...	42 xu ...
3 lmms ...	13 llava ...	23 pre-training ...	33 cvf ...	43 chatbot ...
4 arxiv ...	14 pre-trained ...	24 llm-based ...	34 cogvlm ...	44 human-like ...
5 lmm ...	15 modality ...	25 gpt-4v ...	35 rlhf ...	45 llama ...
6 chatgpt ...	16 peft ...	26 imagebind ...	36 fine-tuning ...	46 zhao ...
7 gpt-4 ...	17 sft ...	27 vit ...	37 blip-2 ...	47 neuro-symbolic ...
8 multimodal ...	18 mm-llm ...	28 next-gpt ...	38 image-text ...	48 encoder ...
9 preprint ...	19 gpt-3 ...	29 q-former ...	39 vicuna ...	49 zhang ...
10 mm-llms ...	20 models ...	30 webvoyager ...	40 instructblip ...	50 tangibility ...

Rows per page: 50 ▼

1-50 of 100



1

/ 2



语料库管理

MANAGE CORPUS

2024_spring_CAT



CORPUS: 2024_spring_CAT (English)

Game



Browse

View documents and folders, edit metadata



Make bigger

Add texts to corpus



Share

Share corpus with other users



Download

Download corpus to your drive



Compile

Compile corpus or change compiler settings



Delete

Remove corpus permanently



Subcorpora

Manage subcorpora



Configure

Change corpus configuration



Logs


View corpus logs



New corpus

Create new corpus

Pricing

 SKETCH
ENGINE

Home News & Events Pricing Guide About us Contact

ACADEMIC PERSONAL SUBSCRIPTION

I am in an [academic environment](#) and I do not conduct lexicography or non-academic activities.

Your country:

Subscription

billing period

yearly

quarterly

monthly

82.68 €

24.24 €

8.81 €

82.68 €
per year

+

Space

for your own [user corpora](#)

1

million words for

0.00 €
per year

No quota is needed to access the [preloaded corpora](#). 1 million words are included free for you to try the [corpus building tools](#).


=

Total


82.68 €
per year excluding VAT

Subscribe now

Sketch engine vs. English-corpora



English-Corpora.org



corpora guides related resources users my account upgrade help

English-Corpora.org and SketchEngine are probably the two largest sites for online corpora. We believe that both sites provide valuable resources for linguists, lexicographers, and language learners and teachers.

The following is a comparison of the two sites, for those who are already family with Sketch Engine, but are new to English-Corpora.org. Admittedly (because this list is at English-Corpora.org), it is probably biased towards English-Corpora.org, and we invite you to look more in depth at what Sketch Engine has to offer as well. Finally, if there is incomplete / incorrect information below, please [let us know](#).

Feature	Sketch Engine	English-Corpora.org
Corpora	<ul style="list-style-type: none">- Extremely wide (90+) range of languages, and hundreds of corpora- For English, very large web-based corpora, as well as many other specialized corpora	<ul style="list-style-type: none">• Mostly English, as well as some for Spanish and Portuguese• For English, perhaps the best suite of corpora for looking at variation: genre-based, historical, and dialectal• Largest corpora are iWeb (14 billion words) and NOW (14.6 billion words and growing by ~250 million words each month)
Users / research	<ul style="list-style-type: none">- Linguistics and lexicographers, teachers and learners, etc <p>(For those with information on Sketch Engine, please send us more detailed / verifiable information on number of users, researchers, universities with licenses, number of publications, etc)</p>	<ul style="list-style-type: none">- ~130,000 distinct users each month, including about 80,000 registered users- ~300 universities have academic (group) licenses, as well as large government-funded licenses- More than 16,500 registered "researchers" (professors or graduate students) in linguistics or language studies- Cited in more than 10,000 academic publications, including more than 5,000 in the past five years- The data (e.g. full-text, word frequency) is used by hundreds of companies, including Google, Amazon, Microsoft, IBM, Samsung, Merriam-

END