**平行语料管理系统**

# 1、项目描述

本项目是一个平行语料管理系统，支持机器翻译系统开发人员管理海量平行语料。平行语料是构建机器翻译系统或其他多语言自然处理系统的基础，该系统可以方便地管理语料库，从而提高翻译的质量和效率。

**成员信息：**

项目负责人：

陈怀宇

项目成员：

李思成 丁隽 严子涵 赵展

# 2、需求文档

## 2.1用户需求

**1. 创建语料库：**在提交请求后，系统应该立即创建一个新的语料库，并返回创建成功的信息。创建语料库前，系统应询问用户新语料库的名称以及是否需要导入已有的语料库。若需要导入，在创建过程中系统应该自动执行合并语料库的行为。在语料库创建成功之后，应该由明晰的提示，来告知用户接下来可以进行哪些操作，例如查看语料库、编辑语料库等。但同时应考虑到因为特殊原因导致的语料库创建失败的情况，例如用户本地空间不足导致的创建失败等。这些信息需要被系统识别并捕获，然后清晰地呈现给用户。在提示用户创建失败后，应该有一个或多个按钮供用户进行回溯性质的工作，诸如关闭软件、重新创建、返回主页面等。

**2. 更新语料库：**在提交新语料后，系统应立即更新语料库，并返回更新成功的信息。从逻辑上来说，提示更新成功后，下方应该有至少两个按钮，一个用于返回用户上一步的界面，另一个用于查看语料库。同时应该考虑到因为各种原因导致的新语料添加失败的情况，系统需要捕获导致添加失败的原因，并以文字和弹窗的方式明晰的呈现给用户。对于不同的失败原因，系统可给出一些通用性建议。在用户了解完失败原因，关掉弹窗后，系统应该自动回到用户上一步的界面。

**3. 合并语料库：**在提交请求后，系统应立即合并两个指定的语料库，并返回合并成功的信息。另外，还应该有更加详细的信息反馈，诸如合并后新语料库的大小。而且，在合并语料库的过程中，系统应该在完成初步合并后自动对语料库“去重”。合并成功后，应该有明晰的操作提示，提醒用户“查看新语料库”以及“返回”等。合并前，系统应询问用户是否需要删除原来的语料库以及新语料库的名称。

**4. 拆分语料库：**在提交请求之后，系统应立即拆分指定语料库，并返回拆分成功的信息。

拆分操作主要是将语料库拆分为不同的块，用于机器语言学习的不同部分，例如：将一个大语料库拆分为两个小的语料库，一半语料库用于机器学习的训练，一半用于对训练结果的效果测试。比如，用户可以提供需要拆分的块数，2块、3块，或者提供拆分的比例，20%、30%、50%，并依据此来拆分语料库。

**5. 过滤语料库：**在执行过滤操作之后，系统应该去除有问题的平行语料、更新语料库。现在有很多语料库都来自于网络，这就导致其中可能存在很多有问题的平行语料，例如翻译严重有问题、对应错误等，我们在使用语料库之前，需要先把这样的语料过滤删去，以免影响训练使用。

**6. 去重语料库：**在执行去重语料库操作之后，系统应该删去语料库中的重复项，更新语料库。在语料库中，可能存在重复的对象，在使用的过程中会占用时间和空间，用户需要提前执行去重操作，以方便后续使用。

**7. 规范化语料库：**在执行规范化操作后，系统应该对语料库进行规范化，更新语料库。平行语料库语句的语句中可能存在意义相同，但表示形式或者编码方式略有差别的字符，例如：不同的空格、不同的逗号，他们意义相同，但在机器识别时可能会被判断为不同的内容，导致平行语料库的使用效率下降。规范化操作需要将这些内容更改为统一的标准，以方便后续操作。

**8. 浏览/查询：**在执行浏览/查询功能时，系统应该返回符合要求的语料，供用户使用和查看。

**9. 分析统计：**在执行分析统计的操作之后，系统需要对指定的语料库进行分析统计操作，并返回操作结果。分析统计主要是针对语料库中各词出现的频次进行分析，统计不同词出现的频次，以此为机器学习提供关于字典选择的参考。

**10. 翻译结果自动评价：**在获取到机器翻译的结果和参考语料库之后，系统需要将这两项作为输入，并以此分析机器翻译的效果。在这一自动评价操作中，主要是依靠BLEU自动评价体系来完成机器自动评价，在最后给出分数。

**11. 翻译结果人工评价：**在获取到机器翻译的结果和参考语料库之后，系统需要将这两项作为输入，依次显示出每一条机器翻译结果和语料库中的参考译句，并给出人工评价界面，用户对每一条语句进行评分，最后系统计算得出平均分数。

**12. 对齐：**在执行对齐操作之后，系统会自动执行单词对齐，在此之后如果用户需要，还可以再进行人工对齐评价，并对不好的对齐结果进行人工修改。

## 2.2用户需求脚本

**1、创建语料库-用户需求脚本**

**初始假设：**

* 一名用户正在创建一个全新的、空的语料库。

**正常状态：**

* 用户通过菜单选项来选择是否以已经存在的语料库为基础来创建新的语料库，并输入新语料库的名称，如果新语料库的名称已经存在，就要重新设计一个名称。
* 用户点击“创建”按钮后，系统会立即按照要求创建一个新的语料库，并提示用户创建成功，以及接下来可以进行的操作，这些操作包括：查看语料库，编辑语料库，返回（主界面）。

**有哪些会出错：**

* 新的语料库创建失败。用户应该查看系统提示的失败原因，选择菜单选项进行重新创建，或者取消创建转去解决问题。

**其他活动：**

* 当需要以已有语料库为基础创建新语料库时，该语料库不能处于“被编辑”状态，即其他用户此时只能查看但不能编辑该语料库，而正在被编辑的语料库不能用于创建新的语料库。

**完成的系统状态：**

* 用户将新的语料库添加到数据库中。行为被添加到系统日志中，系统日志显示了此语料库的创建时间、名称、是否以其他语料库为基础以及相关的语料库

**2、更新语料库-用户需求脚本**

**初始假设：**

* 用户已经使用系统创建了至少一个语料库。用户正在往某一个语料库中添加新的语料。

**正常状态：**

* 用户在语料库列表中搜索目标语料库，也可以通过输入关键字缩小搜索范围。
* 用户选中语料库后，以自由输入的方式输入想要添加的语料，之后点击“添加按钮”。
* 之后系统立即更新语料库，并提示用户更新成功，以及下一步的操作，这些操作包括：返回上一步界面，查看语料库（更新后的），继续添加。

**有哪些会出错：**

* 新语料添加失败。用户应该查看系统提示的失败原因（之后系统会自动回到上一步的界面），并选择继续添加或者取消添加。

**其他活动：**

* 当正在更新语料库时，其他用户只能查看但不能编辑该语料库。

**完成的系统状态：**

* 用户将更新好的语料库放入数据库中（这回取代原来的语料库）。行为被添加到系统日志中，系统日志显示了被更新语料库的名称，更新时间以及新添加的语料。

**3、合并语料库-用户需求脚本**

**初始假设：**

* 用户至少已经创建了两个语料库。一名用户正在合并两个语料库。

**正常状态：**

* 用户在语料库列表搜索目标语料库，或者输入关键字来缩小搜索范围。
* 用户选中两个语料库后，通关菜单选项决定是否在合并结束后删除原来的两个语料库，并输入新语料库的名称，如果新名称已存在，则需要重新输入。
* 接着用户点击“合并”按钮，系统在合并成功后会提示有关信息，如果两个语料库有重合语料，系统在合并时会自动去重。
* 用户可以选择查看语料库或者返回（主界面）。

**有哪些会出错：**

* 两个语料库至少有一个为空或者两个语料库内容相同，合并会失败。用户应该取消合并。

**其他活动：**

* 当正在合并时，两个被操作的语料库都不能被其他用户编辑或查看（因为语料库可能在合并后被删除）。

**完成的系统状态：**

* 用户将合并后的语料库放入数据库（，可选的将用于合并的语料库从数据库中删除）。行为被添加到系统日志中，系统日志显示了新语料库的生成时间，名称，用于合并的语料库的名称以及其是否已被删除。

**4、拆分语料库-用户需求脚本**

**初始假设**

* 用户已经创建了至少一个语料库，现在需要选择语料库，并且对选择的语料库进行拆分操作，拆分为不同的块。

**正常状态**

* 用户在登录之后，可以选择拆分操作，并且选择自己想要拆分的某一个语料库进行拆分。
* 在选择完需要进行拆分的语料库之后，用户还需要选择需要拆分为的份数，以及每一份所占的比例（以100为总和进行分配），最后用户需要为拆分后得到的语料库命名。
* 用户的选择都完成之后，系统会进行自动拆分并且产生新的语料库文件，原语料库文件不会被删除。

**有哪些会出错**

* 用户原来并没有创建任何一个语料库，无法进行拆分操作。这时系统应该返回错误信息，提醒用户需要先创建语料库。
* 拆分的份数=1或者<0，或者过大>10，这时的拆分操作是无效的。系统需要返回错误信息，并停止操作。
* 用户给拆分后的语料库所取的名称不合理，过长或者是和原有的语料库名重复。系统需要提醒用户更改。

**其他活动**

* 在系统执行拆分操作时用户不能进行其他操作。

**完成的系统状态**

* 在系统执行完拆分操作之后，会返回正确信息。用户可以查看到新创建的拆分后得到的语料库。

**5、过滤语料库-用户需求脚本**

**初始假设**

* 用户已经拥有了至少一个语料库，用户可以选择对该语料库进行过滤，最后该语料库得到更新。

**正常状态**

* 用户正常登录之后，选择了某一个语料库并希望对其进行过滤操作。
* 用户在选择过滤操作之后，系统会自动对语料库进行过滤，剔除其中有问题的平行语料，对已有的语料库进行更新。

**有哪些会出错**

* 用户没有语料库，无法执行过滤操作。

**其他活动**

* 在系统执行自动过滤，用户无法执行其他操作。

**完成的系统状态**

* 系统返回过滤成功信息，原来的语料库得到更新，用户可以查看更新后的语料库

**6、去重语料库-脚本**

**初始假设**

* 用户登录后，对选定的语料库选择去重操作，系统后台开始执行去重操作

**正常状态**

* 系统通过匹配字符串操作来查找重复语料，然后反馈给用户重复语料的位置以及重复次数，用户可以自己浏览情况
* 用户可以选择手动删除重复的语料，或者直接一键去重，系统将只保留所有重复语料的其中一个，并更新语料库

**有哪些会出错**

* 语料库没有重复语料。系统应该反馈给用户当前语料库无重复语料
* 用户手动删除重复语料出现误删。系统允许用户选择撤回操作，恢复到上一步的情况
* 用户选择一键去重时想要撤回。系统允许用户中止一键去重，并还原语料库至操作前的原状

**完成的系统状态**

* 系统去重了指定的语料库，语料库更新。记录被添加到系统日志中，日志显示了去重的语料库名称，以及操作开始和结束的时间

**7、规范化语料库-脚本**

**初始假设**

* 用户对指定的语料库选择规范化操作，系统后台开始执行规范化操作

**正常状态**

* 系统会遍历语料库，对于语料中出现的字符统一更换为标准美式键盘下的输入字符

**有哪些地方会出错**

* 遇到系统无法识别的字符。系统应该允许用户选择修改该字符，或者保留该字符

**完成的系统状态**

* 选定的语料库实现规范化，语料库更新。记录被添加到系统日志中，日志显示了规范化的语料库名称，以及操作开始和结束的时间

**8、浏览/查询-用户需求脚本**

**初始假设：**

一名用户正在浏览查询语料。用户已经登录，并具有浏览语料的权限。  
**正常状态：**

用户可以输入关键词或筛选条件，在查询框中执行查询，系统将返回符合要求的语料。查询结果应该包括语料库名称、创建时间、作者等基本信息，以及相关的正文内容。  
系统可以应对大规模数据查询和高并发访问等情况，保证查询效率和稳定性。  
**有哪些会出错：**

查询请求失败，例如查询请求无效、数据访问异常、网络故障等。  
查询结果不完整或错误，例如缺少有用的信息、包含无用的语料、返回给定条件不符的语料等。  
其他活动：

查询结果支持多种格式的导出，例如文本、CSV等。  
**完成的系统状态：**

用户可以通过输入关键词或条件，查询到符合要求的语料，并能够方便地查看其基本信息和正文内容。系统还能够保证查询效率和稳定性，并支持多种格式的导出。

**9、分析统计-用户需求脚本**

**初始假设：**

一名用户正在对指定的语料库进行分析统计，并具有执行该操作的权限。  
**正常状态：**

用户可以选择执行对整个语料库进行统计分析，也可以选择仅对特定标签、作者、时间段等进行分析。  
系统可以针对每个语料库中的所有词进行频次统计，并返回包括但不限于词频、词性、共现关系等统计信息的结果。  
系统可以应对大规模数据统计和高并发访问等情况，保证处理效率和稳定性。  
**有哪些会出错：**

分析统计请求失败，例如请求无效、数据访问异常、网络故障等。  
分析统计结果不准确或有误，例如漏统或多统某些词、返回信息不完整或错误等问题。  
**其他活动：**

用户可以根据分析统计结果，反推出适合的字典，用于机器学习等应用。  
**完成的系统状态：**

用户可以选择对指定的语料库进行分析统计，并收到包括但不限于词频、词性、共现关系等信息的分析结果。系统还能够应对大规模数据和高并发访问等情况，保证处理效率和稳定性，帮助用户更好地进行字典选择等应用。

**10、翻译结果自动评价-用户需求脚本**

**初始假设：**

一名用户正在分析机器翻译结果，需要依靠参考语料库和自动评价体系来完成自动评价操作。用户已经登录，并具有执行该操作的权限。  
**正常状态：**

用户可以将机器翻译结果和参考语料库上传到系统中，系统将对两者进行比较分析，并给出自动评价结果。具体来说，系统应该使用BLEU等自动评价体系，计算出机器翻译结果的相似度、含义等指标，并给出相应的分值。  
系统可以应对大规模数据评价和高并发访问等情况，保证处理效率和稳定性。  
**有哪些会出错：**

自动评价请求失败，例如请求无效、数据访问异常、网络故障等。  
自动评价结果不准确或有误，例如评分偏高或偏低、给出的评价信息不完整或错误等问题。  
**其他活动：**

用户可以在自动评价结束后，根据评价结果进一步修改和优化机器翻译系统，以提升翻译质量。  
**完成的系统状态：**

用户可以将机器翻译结果和参考语料库上传到系统中，系统将以BLEU等自动评价体系计算机器翻译结果的相似度、含义等指标，并给出相应的分值。系统还能够应对大规模数据评价和高并发访问等情况，保证处理效率和稳定性，帮助用户更好地进行翻译结果的评估和优化。

**11、翻译结果人工评价-用户需求脚本**

**初始假设**

* 系统得到机器翻译的结果和参考语料库这两项内容并作为两项输入

**正常状态**

* 系统分析两项输入的内容，识别出一共有多少个不同的句子，一句句将机器翻译的结果和参考的语料库的内容对照输出
* 系统弹出人工评价界面，每一个句子都有一个打分框
* 系统根据打分情况算出所有句子的平均分数

**有哪些会出错**

* 人工打的分过大或过小都不恰当，要有一个提示分域如0-100分，如果超出范围会报错
* 要有分数的类型限制，比如只能是整数，不能是小数

**完整的系统状态**

* 系统接受到机器翻译结果和参考语料后，逐句对照输出二者对比的结果，用户根据情况逐句打分，系统最后得到平均分数

**12、对齐-用户需求脚本**

**初始假设**

* 用户通过点击按钮等提出执行对齐操作

**正常状态**

* 系统自行进行平行语料的逐词对齐，对齐之后，用户可以查看对齐后的效果
* 效果以每一句单词对应呈现出来
* 用户分析是否对齐准确，如果准确，跳过进入下一句，如果不准确，可以用户通过拖拉方块或连线进行修改

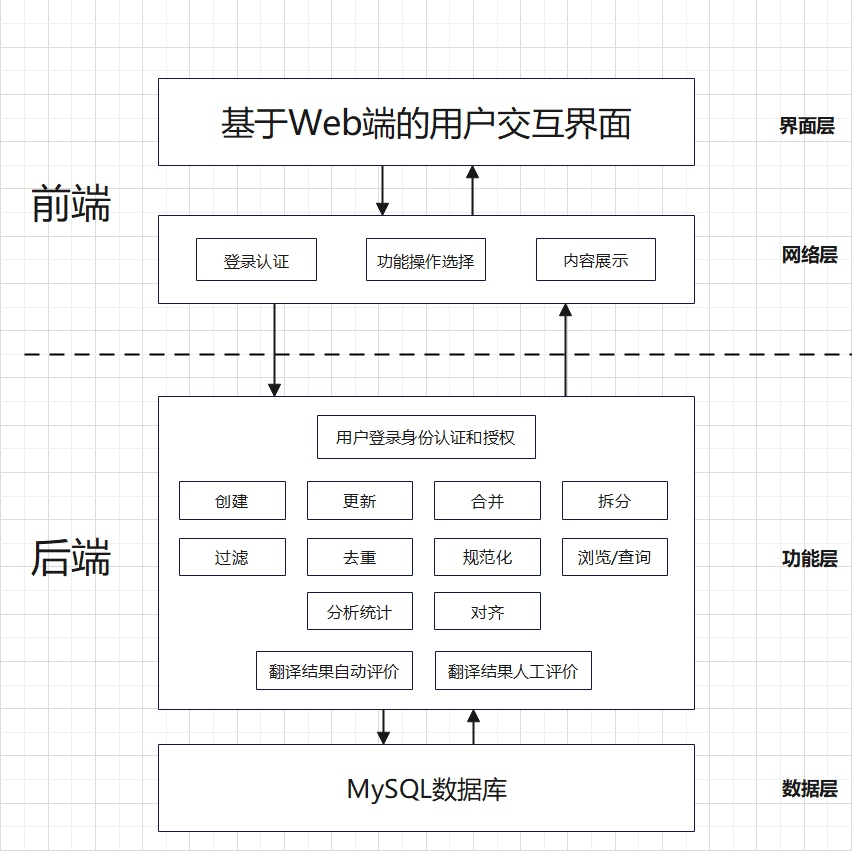
**有哪些会出错**

* 如果用户对刚刚选择的修改或者评价后悔还有返回上一步的按钮

**完成的系统状态**

* 执行对齐操作之后。系统会执行单词的对齐，用户对每一句的对齐效果进行评价，如果不满意，那么进行修改

# 3、系统架构



（系统架构图）

本项目的系统架构主要分为两大部分，前端交互界面和后端功能实现。

前端：主要是包括

1、界面层：Web端的交互界面的设计，能支持各项功能

2、网络层：和后端的接口交互、信息的收发，以确保用户需求的实现

后端：主要是包括

1、功能层：与前端接进行交互，提供各项用户需求的具体实现和接口，例如创建（在本地创建一个 新的数据库用户，或者语料库，并且返回数据库内容列表等）、分析统计（对用户指定的内容进行分析，并且给前端返回结果，用于展示）

2、数据层：利用MySQL数据库来存储用户信息，和相应的语料库，作为操作和处理的基础