Cameo字典翻译

项目背景

将“冲突与调解观察(Conflict and Mediation Event Observations, CAMEO)”事件分类表由英文形式翻译为中文形式，以对中文新闻报道事件进行分类。

项目需求

项目聚焦于动词字典的翻译，暂未对同义词集部分进行处理。预期目标为实现尽可能全面的机器翻译，覆盖尽可能多的英文所对应的中文翻译结果。而后加以人工处理对翻译结果进行审核校对。

项目设计

介词

由于项目本身不需要涉及到介词的提取，故所有涉及到介词的部分全部略过。

动词

1. 首先将单词或词组提取出来
2. 采用 Wordnet *（由Princeton 大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典）* 获取英文同义词集。
3. 而后同样通过 Wordnet 由英文同义词集获取每个同义词的中文同义词集。
4. 基于 Wordnet 将翻译结果与块类别的英文单词的语义相似度进行对比，筛除相似度过低的结果，以降低人工负担。
5. 在 Wordnet 中未收录的同义词翻译，通过调用翻译API获取翻译结果。
6. 将翻译结果与之前的结果进行对比，筛除已收录或已被人工否定的部分。
7. 使用翻译后的词语替换英文，字典其余部分不作改动，保存全部结果。

名词

1. 首先将单词或词组提取出来
2. 采用 Wordnet 获取英文同义词集。
3. 而后同样通过 Wordnet 由英文同义词集获取每个同义词的中文同义词集。
4. 在 Wordnet 中未收录的同义词翻译，通过调用翻译API获取翻译结果。

5. 使用翻译后的词语替换英文，字典其余部分不作改动，保存全部结果。

项目实现

项目依赖

项目采用 Python3 和 Vue 进行开发。其中 Python3 负责处理字典，Vue 实现 Web 页面上的用户交互。

第三方库NLTK

NLTK *(Natural Language Toolkit)* 是 Steven Bird 和 Edward Loper 在宾夕法尼亚大学计算机和信息科学系开发的 Python 开源库，适合 NLP *(Natural Language Process, 自然语言处理)* 的研究和开发。

项目中的 Wordnet 有关部分均借助该库进行开发。

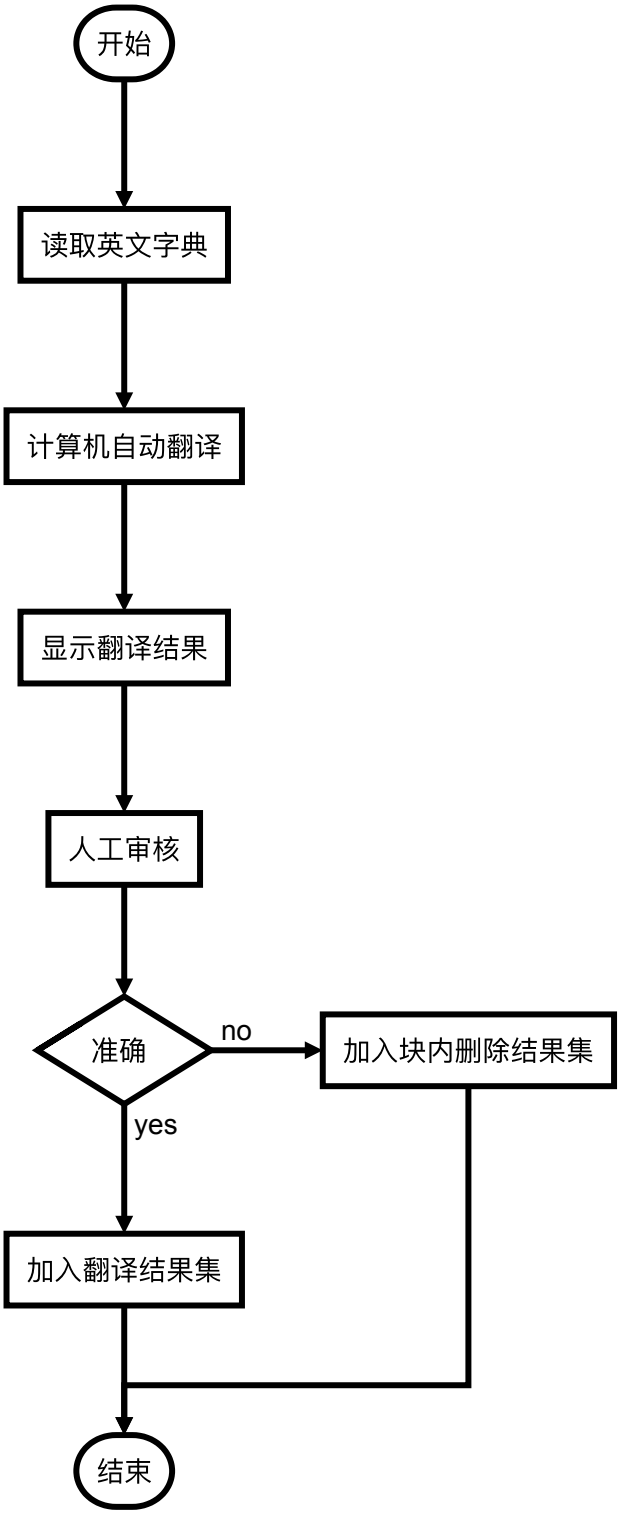
翻译API

项目采用网易有道翻译API和百度翻译API相结合的方式进行结果请求。

网易有道翻译的优势在于返回的结果多，劣势在于有时并不能获取准确的翻译结果；百度翻译的优势在于结果准确率极高，且必定能获取翻译结果，劣势在于仅返回一个结果。

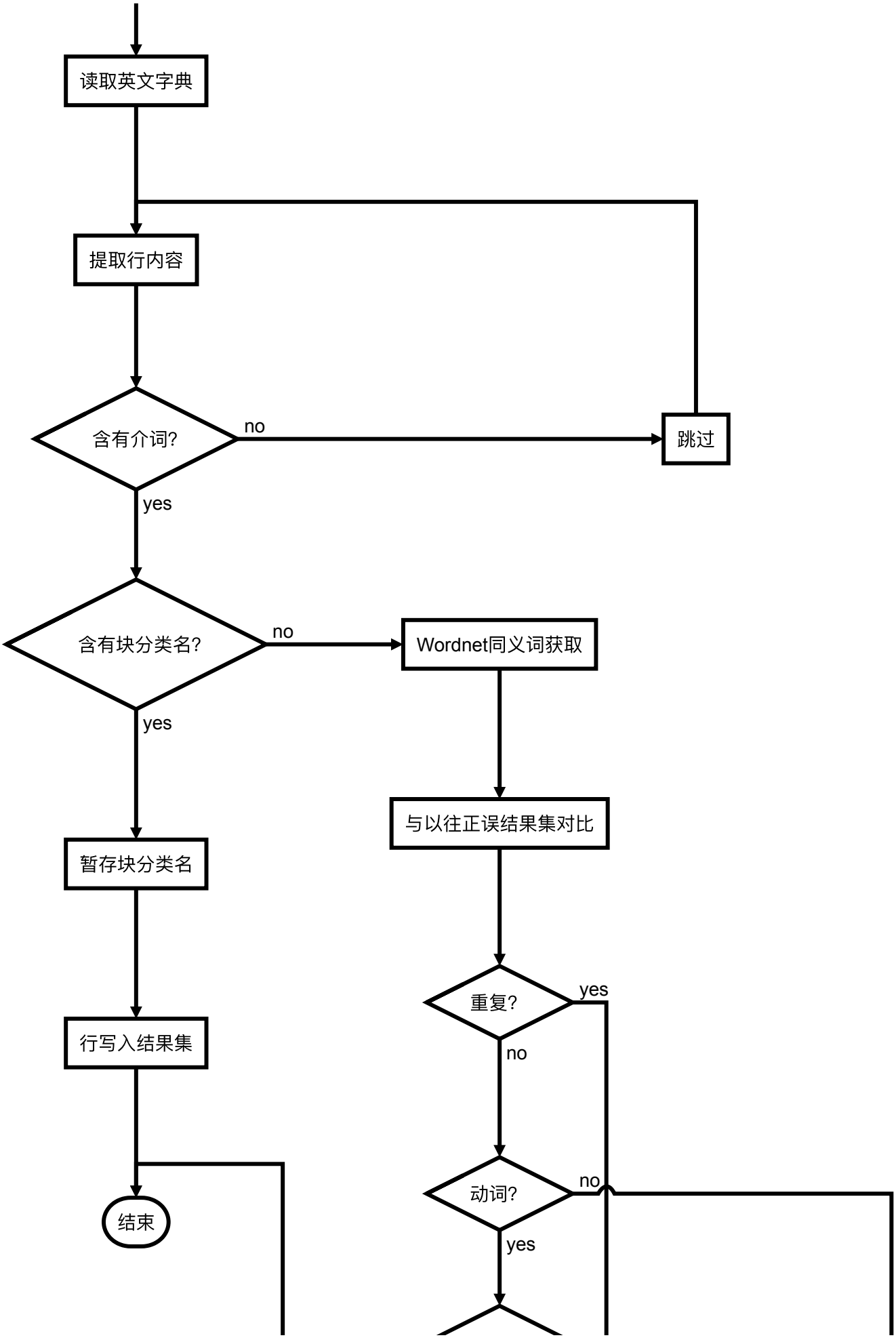
流程

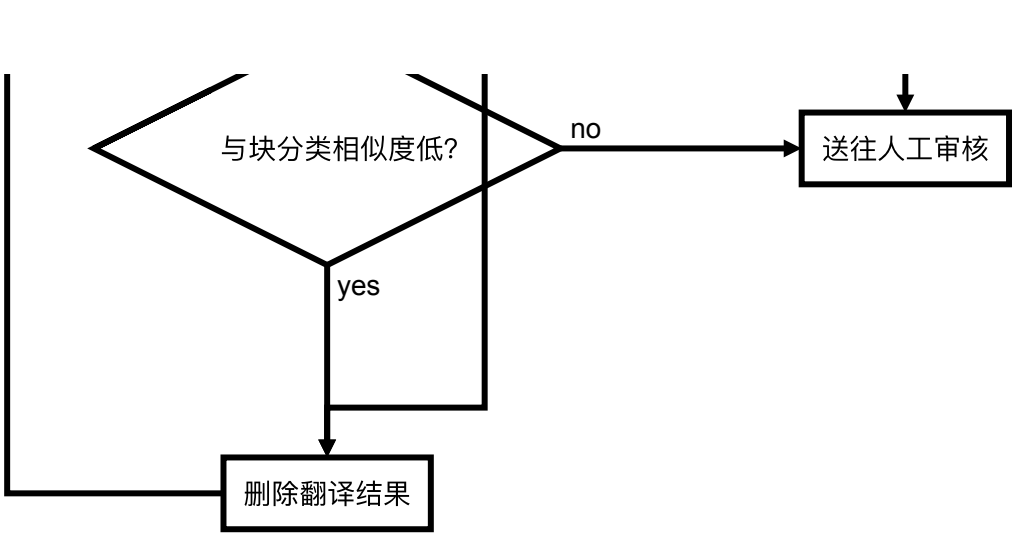
整体流程


开始
读取英文字典
计算机自动翻译
显示翻译结果
人工审核
准确
no
加入块内删除结果集
yes
加入翻译结果集
结束

计算机翻译流程


开始


读取英文字典
提取行内容
含有介词?
no
跳过
yes
含有块分类名?
no
Wordnet同义词获取
yes
暂存块分类名
与以往正误结果集对比
行写入结果集
重复?
yes
no
结束
动词?
no
yes


与块分类相似度低?
no
送往人工审核
yes
删除翻译结果