英文阅读器 使用说明

潍坊上海新纪元学校 高二 邱江坤

摘要

英文阅读器是一个 Web 应用程序，设计给英语学习者使用，可以起到辅助阅读的功能，让他们感受到阅读的快乐。本文阐述了英文阅读器内部的使用方法和工作原理。

目录

摘要 **1**

1 使用方法 **3**

2 Python 环境 **3**

3 Python Web 环境 **3**

4 全文翻译 **3**

5 前端 UI 库 **4**

6 MDX 文件格式 **4**

7 词典文件 **5**

8 获取例句 **5**

1 使用方法

本软件需要的环境几乎全都在程序包中。需要的环境如下：

- python3 环境
- chrome 等现代浏览器
- Windows 操作系统
- 需要阅读的文章
- 本程序的完整程序包

首先解压本软件的压缩包到一个合适的目录，然后双击目录中的 start.bat，即可启动程序。随后会出现浏览器界面，访问的正是本软件的界面。

将想要阅读的文章粘贴到文本框内，然后点击阅读按钮，会发现屏幕被分为三部分，左上部分是阅读区域，这里有之前输入的文本，双击这里的文本，右侧会出现双击的单词的释义和例句。

本程序提供了全文翻译功能，具体实现会在接下来的文章里详解。

在下方，会有一个单词表，里面包含了文章中出现的所有有难度的词汇和汉语意思，按照字母升序排列。此举会帮助使用者更好地记忆词汇，消除词汇障碍，并在在接下来的阅读中学习和巩固所学的单词，也可以更好地理解阅读。Python 是一种解释型、面向对象、动态数据类型的高级程序设计语言。

2 Python 环境

Python 由 Guido van Rossum 于 1989 年底发明，第一个公开发行版发行于 1991 年。

像 Perl 语言一样, Python 源代码同样遵循 GPL(GNU General Public License) 协议。

3 Python Web 环境

Flask 是一个使用 Python 编写的轻量级 Web 应用框架。其 WSGI 工具箱采用 Werkzeug ，模板引擎则使用 Jinja2 。Flask 使用 BSD 授权。

Flask 也被称为 “microframework”，因为它使用简单的核心，用 extension 增加其他功能。Flask 没有默认使用的数据库、窗体验证工具。

Flask 本是作者 Armin Ronacher 的一个愚人节玩笑，不过后来大受欢迎，进而成为一个正式项目。”It came out of an April Fool’s joke but proved popular enough to make into a serious application in its own right.” Flask 受到了基于 Ruby 语言的 Sinatra 项目的影响。

由于程序本身需要的只是一个简单的 web 环境，并没有选择 Django 等重量级 web 框架，而是选择了轻巧的 Flask 框架。

4 全文翻译

本程序的全文翻译功能采用的是百度翻译。

机器翻译一直以来是一个值得研究的方向，从古至今，出现了规则翻译、统计翻译、实例翻译、神经网络等方式，而且发展的时机有所不同。

规则翻译是按照人工制定语法规则进行翻译的翻译方式，是最早的翻译方式之一。但是因为人工考虑的情况不够全面，容错率不高，成本太大，效果不好，渐渐被后续出现的方式取代。

统计翻译是后来出现的翻译方式，这种方法需要调用大量语料数据，对翻译模型进行训练，这种方式效果好，对配置要求相对较低，无需人工干预，成为几十年来一直使用的方式。

实例翻译类似于查字典，将别人翻译过的经典的整句直接在数据库中查找，然后呈现出翻译后的整句。这种方式的效果最好，但 chen

神经网络模型在很早之前就被提出。神经网络的精髓是，模仿大脑中的神经元，训练模型进行翻译。无奈早期计算机运算能力不足，无法支撑如此大规模的模拟运算。随着计算机计算能力的提升，神经网络重新进入人们的视野中，在近几年内取得了很大的成果。神经网络获得结果相对较快，效果相对较好，受到了从厂家到个人的喜爱。

如百度百科所言：

百度翻译目前支持 28 种语言的互译。百度翻译在海量翻译知识获取、翻译模型、多语种翻译技术等方面取得重大突破，实时准确地响应互联网海量、复杂多样的翻译请求。所研发的深度学习与多种主流翻译模型相融合的在线翻译系统以及“枢轴语言”翻译等技术，处于业内领先水平。

1. 对神经网络翻译（NMT）方法进行了系统而深入的研究，针对 NMT 存在的问题提出了系列创新方法，发表于领域顶级会议 ACL、EMNLP、AAAI、IJCAI 等。其中『Multi-Task Learning for Multiple Language Translation』被纽约时报评价为『开创性的工作』。

2. 创新性地提出了将深度学习模型和多种主流翻译模型相融合，包括传统的基于规则、基于实例、基于统计等翻译策略，做到发挥多种方法各自优势，从而从整体上提升翻译效果。

3. 系统提出了基于“枢轴语言（pivot language）”的机器翻译模型，攻克了机器翻译中小语种覆盖和语言快速迁移的难题。

4. 将百度先进的搜索技术与翻译技术相结合，基于网页检索、网站权威性计算、大数据挖掘、新词侦测等技术，从海量的互联网网页中获取高质量翻译知识。

5 前端 UI 库

Semantic 是一个设计良好的 UI 库. 关键特点：

- 50 多个 UI 元素
- 3000 多个 CSS 变量
- 交互式设计

Semantic 允许开发人员快速构建漂亮的网站，简洁的 HTML，直观的 javascript 和简化的调试，有助于使前端开发成为一种愉快的体验。语义是响应性设计，允许您的网站在多个设备上扩展。Semantic 是生产就绪的，并且与 React，Angular，Meteor 和 Ember 等框架相结合，这意味着您可以将它与任何这些框架集成，以便与应用程序逻辑一起组织 UI 层。

6 MDX 文件格式

什么是 MDX?

MDX 是一种 dBASE IV 使用的文件格式，是用来存储数据库的多重索引文件。通俗来讲，MDX 是多数电子词典应用程序的文件格式的事实标准，目前支持的词典应用程序有

- Golden Dict

- Mdict 欧路词典
- Stardict

此外，还有大量的读取 MDX 的程序库可供使用。本程序选用了 readmdict 库。

7 词典文件

词典文件采用了知乎用户 韦易笑 的《简明英汉必应版》，这个词典从《简明英汉增强版》入手，补充更多短语、谚语、新词、俚语和专业术语，并对前 20 万基础词汇使用必应释义进行了校对，最终发布这个收录 432 万词条的《简明英汉必应版》。

收词情况如下：

- 牛津高阶第八版：7.2 万词条
- 朗文当代第五版：6.2 万词条
- Merriam-Webster's Collegiate Dictionary：11.9 万
- 柯林斯 Cobuild 5：3.4 万
- 21 世纪英汉词典：37.7 万
- 有道本地增强版离线词库：40 万
- 欧陆离线词库：40 万

《简明英汉必应版》除了整合了市面上各类免费和开源资料，还利用 BNC/COCA 语料库进行词频矫正，并使用 NodeBox, WordNet 等自然语言处理工具包对各类时态语态，派生词等进行补充和标注。再根据考试大纲和柯林斯星级还有牛津 3000 核心词进行标注，容易看出某个单词的重要性。

由于以上的特点，加上阅读文章时不能中断的思维，《简明英汉必应版》很适合作为本程序的配套词典。

每个词语都附有考纲要求，如“中高研四六托雅宝”，分别对应中考、高考、研究生考试、大学四级、大学六级、托福、雅思、GRE 红宝书的要求。

8 获取例句

本程序的例句是从 bing 词典获取的，原理是向 bing 词典发送 get 请求，然后通过程序处理，获得例句部分。同时根据适配，配合调整好的 CSS，放置到程序的合理位置，方便用户使用。