Original Research Article

Design and Implementation of an English-Chinese Translation System Based on Crawler

Yanfei Sun

Zibo Vocational Institute, ZiBo, China

Abstract: This system uses crawler technology to obtain relevant data through simulation translation software, and it saves these data to the local. It can finish accurate and real-time translation of Chinese and English. The system is very practical and it provides convenience for the scenes that need to be translated in daily life.

Keywords: English-Chinese translation; Crawler; Python

Translation between English and Chinese is a particularly important issue in daily life. For example, when we are studying or browsing an English website, we often encounter unfamiliar English words. Sometimes a word can make it difficult for us to understand the meaning of a sentence. At this time, we have to look up the English-Chinese dictionary, but it is time-consuming and laborious. We can design and implement a software for English-Chinese translation to accurately translate English in real time.

1. Introduction to reptiles

With the rapid development of the Internet era, the network has become the carrier of big data information. How to effectively extract and utilize this information from the network is a challenging problem. The common method for users to retrieve information from the network is search engine, but it has certain limitations. For example, users with different expectations often have different retrieval purposes and needs. The content returned by the common search engine often contains a lot of things that users don't need. If you want to monitor the development and changes of online news in real time, you must use relevant tools. The speed of manual work is too slow, so the network crawler came into being.

2. The design and implementation of the system

2.1. Overall system structure design

Reptile system software structure diagram:

Input the URL of the webpage to the crawler system, the crawler will open the webpage, parse and process it, extract the webpage text, and then output the webpage text, as shown in Figure 1[1].

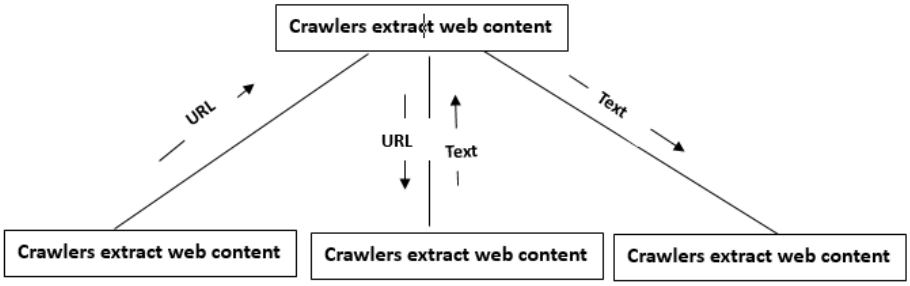

Figure 1. Crawler system structure diagram.

This system is designed to be composed of three subsystems, namely, a network crawler system, a news analysis system, and a final result display system, as shown in Figure 2.

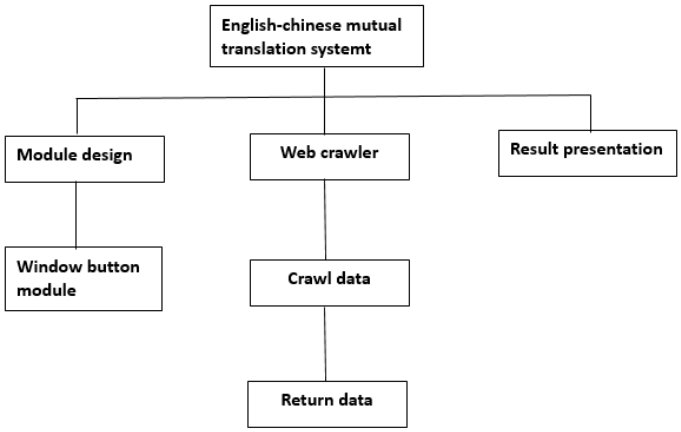

Figure 2. System hierarchy diagram.

2.2. System function design

The system structure logically consists of four parts: the first part is the window module, responsible for building windows, input box labels, and buttons; The second part is the button function implementation module, which implements the specific functions of the buttons in the window module; The third part is responsible for processing and analyzing web page data after crawling; The fourth layer is the data display module, which is responsible for displaying the analyzed and organized data in the form of text in the translated text box.

The core part is the third part which is the crawler analysis module.

The design idea of crawler analysis: crawl the website address, obtain the corresponding page, extract useful data, and display the data in the text box we want to display it in[2].

(1) The first URL address we need to crawl is http://fanyi.youdao.com/ .We can open the Youdao translation address in a browser for web analysis. As shown in Figure 3.

(2) When we want to translate a content, we must send some data to the server which is send in the form of form parameters. We can find it in the browser.

(3) When the network is smooth and there are no other obstacles, the server receives the data sent to it and also returns a response. In fact, it is a json data type. This includes the content we input and the content returned by the server after analysis. The data type is as follows:

{"translateResult":[[{"tgt":"hello","src":"你好"}]],"errorCode":0,"type":"zh-CHS2en","smartResult":{"entries":["","hello\r\n","hi\r\n","how do you do\r\n"],"type":1}}

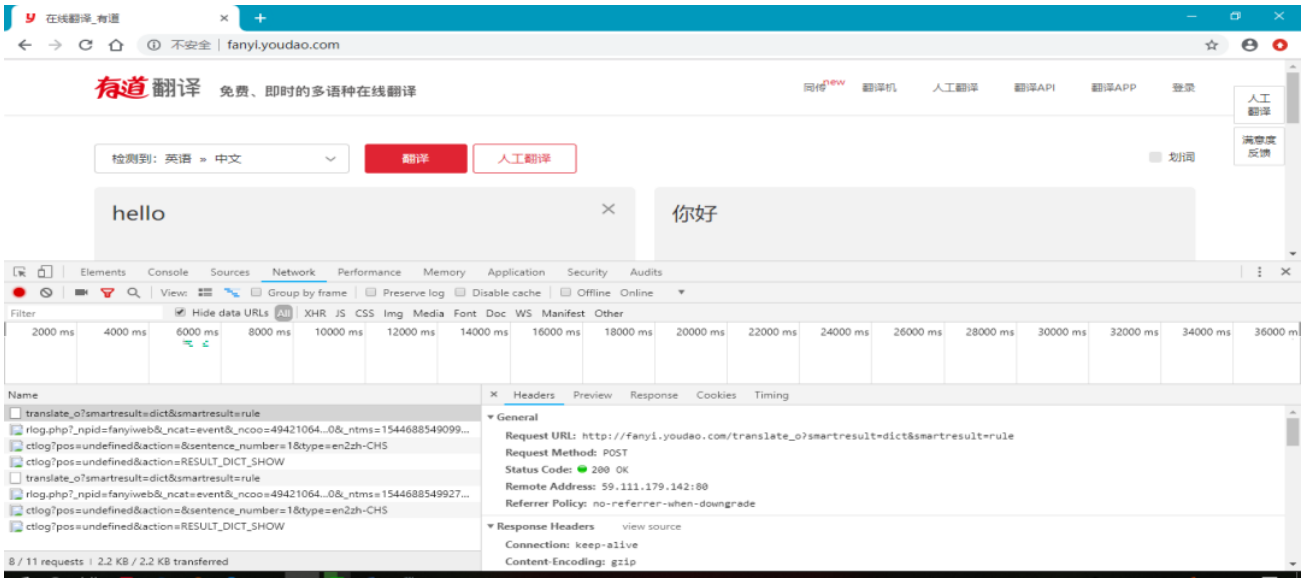

Figure 3. Page analysis.

Next, we can use Python to simulate this process to achieve the effect of English-Chinese translation. We can use Python to send the same content as the browser sends to the server. When we simulate the browser with Python, the server will also return a parameter to us.

Since we are simulating this process with Python, we need to go to the POST request to find the URL address of the same path, which is http://fanyi.youdao.com/translate_o?smartresult=dict&smartresult=rule

Then we copy all the parameters from the browser to Python. We need to modify some data, such as data ['i '], which means we need to input the content and send it to the server of Youdao Dictionary. Here, we can directly make data ['i']=content. And data ['salt '] means timestamp, which means a complete and verifiable string of fifteen digits that exists at a specific time. Youdao Translator cannot accept timestamps, we can directly delete them.Another option is the encryption of Youdao Cloud, which is data ['sign ']. Of course, it can also be cracked, but it is very troublesome. Here, we can directly delete the data ['sign']. The subsequent code is as follows:

```
data = {}
data['action'] = 'FY_BY_REALTIME'
data['client'] = 'fanyideskweb'
data['doctype'] = 'json'
data['from'] = 'AUTO'
data['i'] = content
data['keyfrom'] = 'fanyi.web'
#data['salt'] = '1568883546838'
#data['sign'] = '69bb6794183edace5bc5c2ddb25ac936'
data['smartresult'] = 'dict'
data['to'] = 'AUTO'
data['typoResult'] = 'false'
data['version'] = '2.1'
```

After debugging, we can run it. Of course, it will return a data to us like the browser.

{"translateResult":[[{"tgt":"hello","src":"你好"}]],"errorCode":0,"type":"zh-CHS2en","smartResult":{"entries":["","hello\r\n","hi\r\n","how do you do\r\n"],"type":1}}

The next step we need to do is to extract this part of the data and display it in the text box. Here we need to extract the value from a list. The code is as follows:

```
s = result.json()
transResult = s['translateResult'][0][0]['tgt']
# print(transResult)
res.set(transResult)


res = StringVar()
entry1 = Entry(root,font = ('微软雅黑',10),textvariable = res)
entry1.grid(row = 1,column = 1)
```

2.3. Exception handling

In the process of running the system because of the complex operating environment, there may be various abnormal problems. And these abnormal problems need us to deal with slowly.

The general situation of crawler anomaly is to enable the maintenance personnel of crawler to understand the overall operating state of crawler in time.

Robustness of the crawler. By observing the webpage document data crawled by the crawler, we can find out various abnormal problems when the crawler is in the webpage document data through analysis, which is convenient for the optimization of the crawler system code.

The crawl efficiency of a crawler. The log system records the log printed when the crawler takes the web page, and after some statistical integration, it is possible to obtain the data collection ability of the crawler. It is possible to obtain the data collection ability of the crawler.

The influence of crawler on the object website. If it is found that after the crawler runs for a period of time, it can no longer collect the page document data from a website, even if you consider whether the crawler crawls the page document data too frequently and is blocked by the website administrator.

When crawlers crawl a large number of websites, they can seriously occupy the resources of the website. Therefore, many websites have implemented anti-crawler mechanisms. When crawling a large amount of website data, errors such as "Access Denied" may occur, and the web server directly denies access. At this time, crawlers need to be able to disguise themselves as a real browser. There are the following methods.

(1) Disguised User-Agent

User-Agent specifies the type of browser so that the Web server can recognize different types of browsers. At present, in order to obtain website data in time, most crawlers usually set up a User-Agent of some browser to "deceive" the website.It will tell the Web server that they are a certain browser, and then the Web server will return the real web page data.

(2) You need to log in to access web data

Although most websites can access the content of each page without landing. With the increasing demand of various research institutions for Internet data, a variety of crawlers are born. But some crawlers will uncontrollably crawl the website, greatly consuming the bandwidth of the website, resulting in normal users can not access the website. So now the vast majority of websites have correspondingly launched some protection against "malicious crawlers" strategy. The solution is to log in to the host with the IP address using a browser. After the login is successful, if the browser is Chrome or firefox, you can directly view the cookie of the browser,

and set the cookie with HttpClient.

(3) Use proxy IP addresses for access

If the web server maintenance personnel use the number of visits to a certain IP in a certain period of time to determine the crawler, and then block the IP of these crawlers, the above camouflage will fail. First of all, we need to simulate a decentralized and independent user, which requires random and dispersed IP agents all over the country, and when the crawler runs, some IP is randomly selected from these IP agents to use as a proxy, so that we can perfectly solve the problem of single IP high frequency and high traffic access to the website.

Here you can do something similar to the database connection pool.Of course, here is not the database connection, but one by one proxy IP, and then specify the relevant proxy IP allocation policy. After the proxy IP pool is done, a load balancing should be done, and the IP that can be used normally in the proxy IP pool should be used for cyclic access each time, so that a single IP will rapidly decline to the web server, which is very obvious.

3. System testing

After the system design is completed, a series of tests need to be conducted. Changes in the internal environment and external factors during the debugging process will affect the operation and operation of the system. When the system adapts to these changes, it gradually becomes perfect and achieves the desired results. According to the functional situation of this system, black box testing is the main method, supplemented by white box testing. The task of black box testing is to detect whether each function of the system can operate normally and whether the operation result is correct. White box testing treats the project as a transparent white box, requiring the operator to know the project process and project code, and detect whether the function meets the requirements according to the specifications. It requires a high level of operator skills[4].

4. Conclusion

The system uses the GUI tkinter library and integration, uses the requests library to carry out crawler project, and completes the realization of each function of the whole website.The system has good performance and operability, high response speed and efficiency, and it conforms to the functional requirements of real-world English-Chinese translation systems, providing convenience for everyday translation scenarios.

References

1. SUN Woyu. Design and Implementation of Sina Weibo Crawler Program Based on Python [J]. SCIENCE & TECHNOLOGY INFORMATION, 2022, (12.)
2. Zhao Wenjie,Gu Ronglong.The web crawler technology based on Python[J].Hebei Agricultural Machinery,2020(8).
3. Guo Lirong. Design of web crawler program based on Python[J]. Electronic Technology and Software Engineering,2017,12:248-249.
4. Xu Lihua.Software Test[M].Beijing: Higher Education Press,2013.05.