Saturday, November 28, 2020

Course Project Progress Report

W. Daniel Child

Project Objective

As delineated in the project proposal, my goal is to develop a suite of search functions that make it easier to search for rare Japanese patent terms and to show those terms in context. With rare terms like this, search engines often return irrelevant pages, and one has to hunt for the immediate context in question. From a translator's perspective, it would be extremely helpful to have concise reports on term contexts to understand how the terms are used. The following is a report on what I have accomplished so far.

Setting Up Search Engine Capabilities

It appears that Google has discontinued *free* options for using their search API. Another option has been developed for Python, however, and it seems to work perfectly well. So I am using that as my starting point. So I installed the search engine parser as follows:

```
pip install search-engine-parser
```

It turns out that this search-engine-parser supports different search engines, including Google, Yahoo, and (allegedly) Bing. However, Bing failed miserably, flagging searches as possibly illegal, and so I didn't use it. I did compare Yahoo and Google for the query '黒文字' ("black characters" in Japanese) and found that they had very different results.

Google Top 3

'https://www.google.com/url?q=http://verdure.tyanoyu.net/kuromoji.html&sa=U&ved=2ahUKEwjInaflt6btAhWDFVkFHeg5C2E4ChAWMAZ6BAgHEAE&usg=AOvVaw1Bg6RJJuLCKaDfYsL1sHWf'
'https://www.google.com/url?q=http://www.jugemusha.com/jumoku-zz-kuromoji.htm&sa=U&ved=2ahUKEwjInaflt6btAhWDFVkFHeg5C2E4ChAWMAd6BAgIEAE&usg=AOvVaw17dFX2w6L1A6Nug5SQGJT6'

'https://www.google.com/url?q=https://www.nakagawa-masashichi.jp/shop/g/g4547639507716/&sa=U&ved=2ahUKEwjlnaflt6btAhWDFVkFHeg5C2E4ChAWMAh6BAgJEAE&usg=AOvVaw2JVwx-Z7pwG8sEWU7Vb9Su'

Yahoo Top 3

'https://ja.wikipedia.org/wiki/%25E3%2582%25AF%25E3%2583%25AD%25E3%2583%25A2%25E3%2582%25B8'
'https://www.weblio.jp/content/%25E9%25BB%2592%25E6%2596%2587%25E5%25AD%2597'
'https://search.rakuten.co.jp/search/mall/%25E9%25BB%2592%25E6%2596%2587%25E5%25AD%2597/'

Spot-checking the different websites, I actually found some of the Yahoo websites more helpful. I have therefore decided to incorporate both search engines into the patent search system.

Parser

Once you have a list of web pages, the next step is, of course, to be able to parse these individual pages. I am currently developing the parser (based on BeautifulSoup) to look for the target terms. Since the point is to understand rare terms from context, I need to see sentences and not just headings or titles, which will not be of much help from a context standpoint. I am also going to develop a capability that is sensitive to the possibility of the term being defined, and I am also entertaining the possibility of looking for cases where the term is actually translated into English.

Text Analysis and Issues Being Faced

I have already tested Stanza (Stanford's Python wrapper for Stanford CoreNLP) and it is looks to be capable of handling Japanese so long as you install it correctly (you need to pay attention to whether you are using vanilla Python or Python via an Anaconda environment) and utilize the right module. Given its inability to properly parse some fairly straightforward terms, however, it is equally clear that the terms I will be testing against will not tend to be in the Stanza dictionaries, meaning that they will not be parsed correctly.

That may not matter. I will be able to recognize such terms as appearing in consecutive lemmas. Stanza has a function that enables you to parse sentences, so once I have identified the surrounding context, I should be to pull the context that I am looking for.

I have not yet decided whether to it will be necessary to incorporate alternative Japanese tokenizers such as Juman or Kuromoji.

What Remains to Be Done

Obviously, I need to continue working on the parser, and to make it successfully pulls out the information I need from each of the web pages returned by the search engine parser. Stanza sentence identification needs to be smoothly integrated so that it goes to work on the data extracted by the web parser.

Once this has been done, I want to demonstrate how many websites could be bypassed by using this system, and how much easier it is for a translator like myself to find contextual information about rare terms.

The final step will be to generate clean reports based on the data that I have extracted.