Using Microsoft's Translator API with Python

May 07, 2012

Before Macaronics, I experimented with automated machine translation.

Microsoft provides a Translator API which performs machine translation on any natural language text.

Unlike Google's paid Translation API, Microsoft offers a free tier in theirs, for up to 2 million characters per month.

I found the signup somewhat confusing, though, since I had to create more than one account and register for a couple of different services:

1. I had to register for a Windows Live ID
2. While logged in with my Live ID, I needed to create an account at the Azure Data Market
3. Next, I had to go to the Microsoft Translator Data Service and pick a plan (I chose the free, 2 million characters per month option)
4. Finally, I had to register an Azure Application (since I was testing, I didn't want to use a public url, and fortunately that form accepted 'localhost', though it insisted on my using 'https' in the definition)

The last form, i.e., the Azure Application registration, provides two critical fields for API access:

- Client ID — this is any old string I want to use as an identifier (i.e., I choose it)
- Client Secret — this is provided by the form and cannot be changed

With all the registrations out of the way, it was time to try a few translations.

The technical docs were well-written, but since there was nothing for Python, I've included an example for accessing the HTTP Interface.

My code is based on Doug Hellmann's article on urllib2, enhanced with Michael Foord's examples for error-handling urllib2 requests.

Here's a simple usage example from Japanese to English, in the Python REPL:

```
>>> import msmt
>>> token = msmt.get_access_token(MY_CLIENT_ID, MY_CLIENT_SECRET)
>>> msmt.translate(token, 'これはペンです', 'en', 'ja')
<string xmlns="http://schemas.microsoft.com/2003/10/Serialization/">This is a pen</string>
```

The API returns XML, so a final processing step for a real program would be to use something like lxml to parse out the translation result.

Here's a snippet for getting just the translated result out of the XML object returned by the API.

In the case of the example above, this is just the classic[1] phrase:

```
This is a pen
```

[1] It's classic in that "This is a pen" is the first English sentence Japanese students learn in school (or so I'm told)

Archived from the original at http://denis.papathanasiou.org/
Bitcoin Donate: 14TM4ADKJbaGEi8Qr8dh4KfPBQmjTshkZ2