Report File

Assignment – 6

Introduction to NLP

ITCS 5111

Note: Please refer to the README.txt for an understanding the folder structure and solution format.

**Solution1:**

*Language:* The choice of my foreign language for the solution 1 is **Spanish**.

*Data*: The source of my data for both spanish[es] and its corresponding conversion to english[en] is from a **user manual of a 15 inch Macbook Pro** and its link is:

<https://help.apple.com/macbookprothunderbolt3/late-2017/?lang=es>

and its corresponding english translation at:

<https://help.apple.com/macbookprothunderbolt3/late-2017/?lang=en>

*Dev-test split:*

The first 10 sentences of data from both the languages are taken as Dev data which can be found at:

/data/sentences\_es\_dev.txt for the Spanish language

The remaining 5 sentences in Spanish can be /data/sentences\_es\_test.txt

Similarly the corresponding source translated english can be found in:

/data/sentences\_en\_dev.txt - first 10 sentences translated from the source in English

/data/sentences\_en\_test.txt - nect 5 sentences translated from the source in English

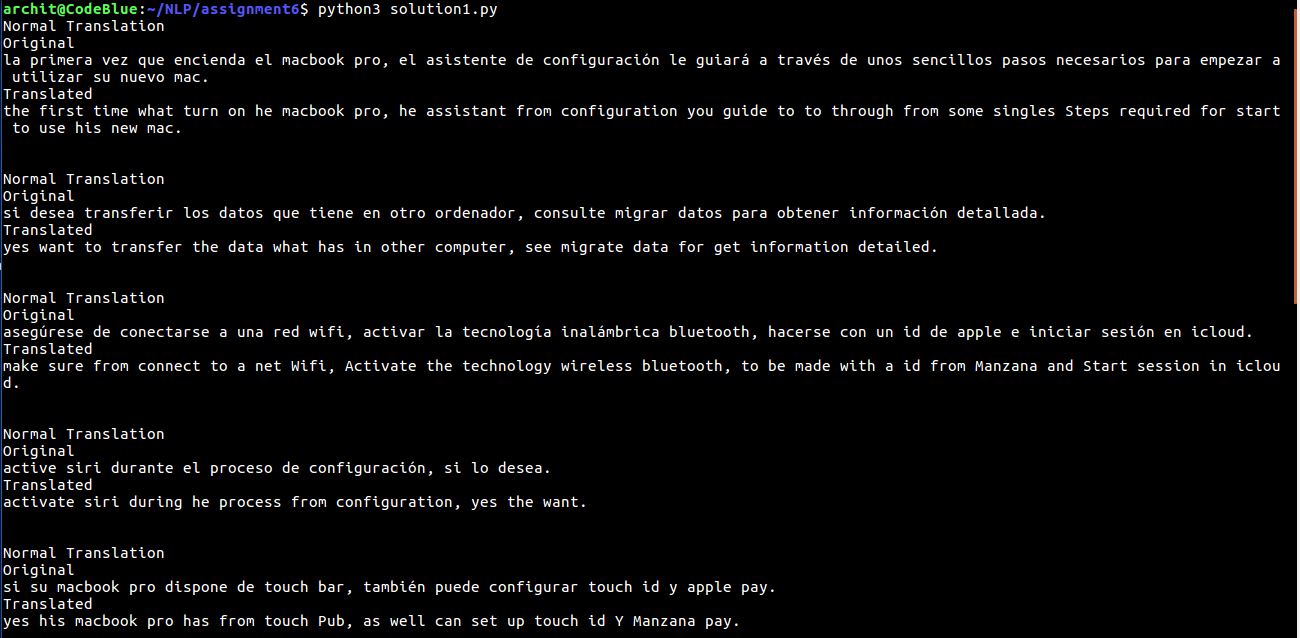
*Dictionary:*

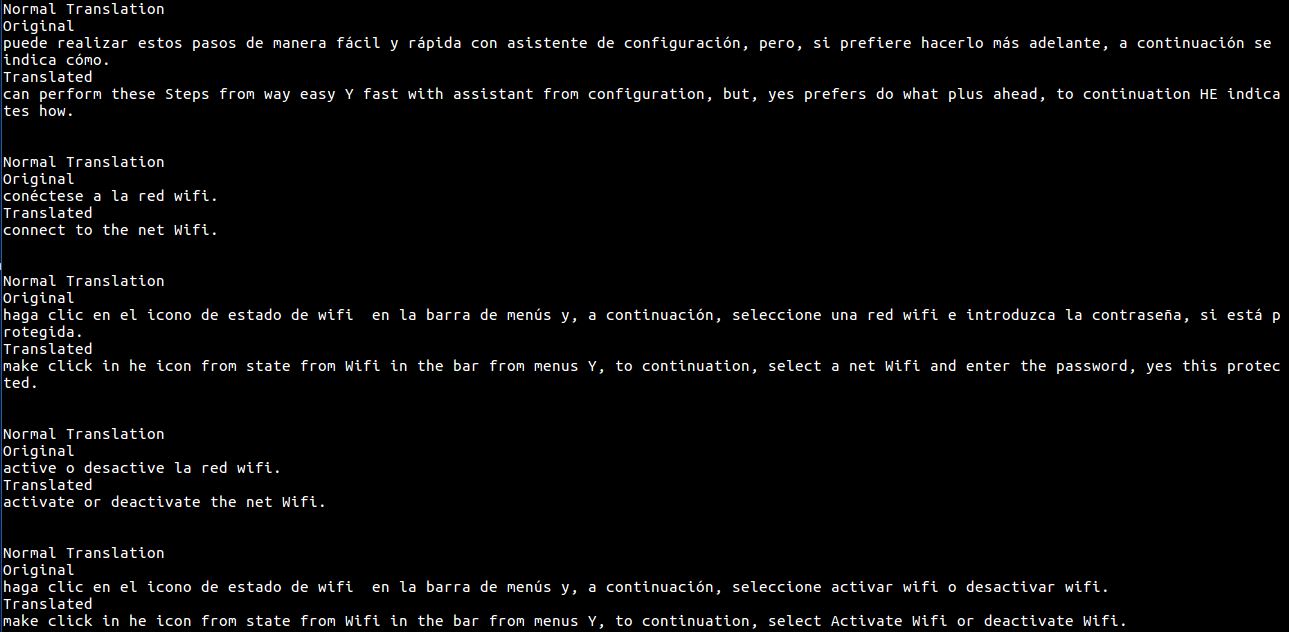
The corresponding bilingual dictionary of the distinct words in the sentences selected using an online service in JSON can be found at /data/dictionary.json

*Translation System:*

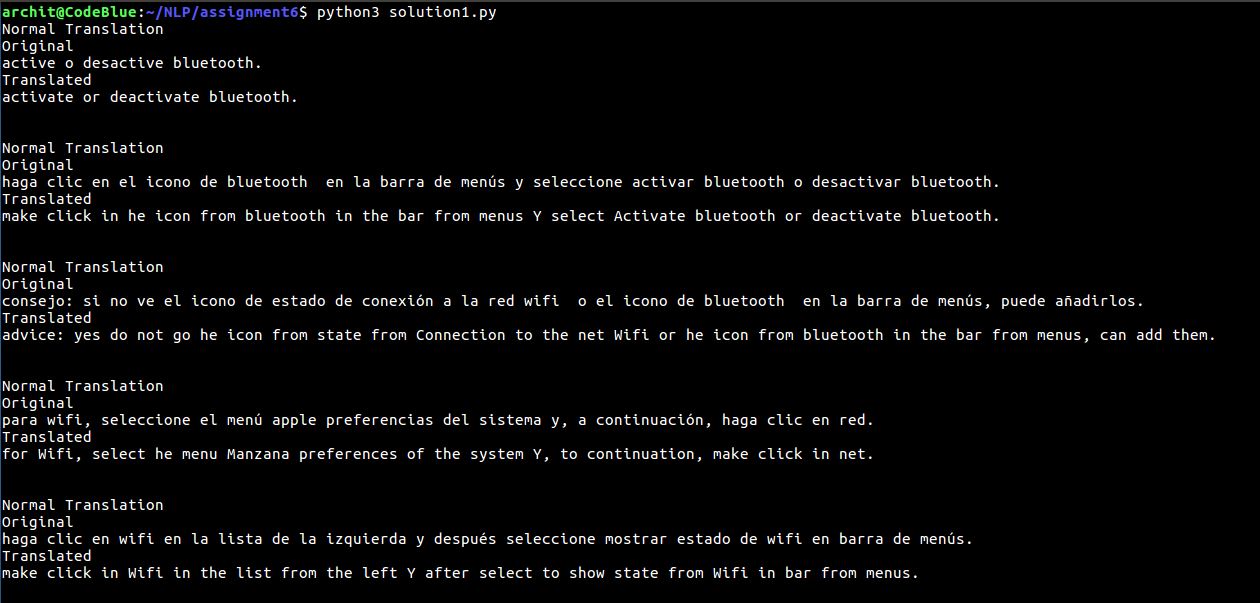
For the Direct MT, we have used our bilingual dictionary to translate the Spanish[es] (foreign language) into English[en] using **word tokenization annotation** using the NLTK toolkit.

*Dev data*, i.e. 10 lines selected in Spanish with normal MT (i.e. word to word translation) gives the output as:





*Test data*, i.e. 5 lines selected in Spanish with normal MT (i.e. word to word translation) gives the output as:



\*Note: A consolidated output with all the improvements over the Normal Machone Translation can be found in the Output.txt file for the solution 1.

Now as far as the Pre/Post – Processing (improvements) strategies are considered. Following are the improvements used by me:

# Improvement 1: Swap the nearest verb with the word after noun

# Improvement 2: Swap the nearest adjective with the word after noun

# Improvement 3: Bigram Language Model

# Improvement 4: Trigram Language Model

# Improvement 5: Bigram POS Language Model

# Improvement 6: Rearrangement of POS

**Solution 2**

The selected corpus for this part of the assignment is es-en i.e. spanish to english and the translated files generated after running the solution2.py on dev and test data sets can be found in:

/es-en/dev/newstest2012.translated

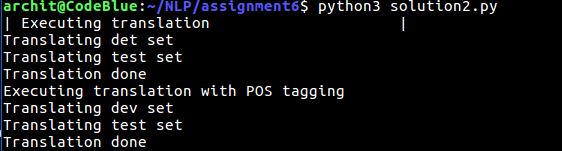
/es-en/test/newstest2013.translated

solution2.py also contains the improvement implementation i.e. pos tagging as a solution to the second part of the question2 and its translated files can be found at

/es-en/dev/newstest2012\_pos.translated

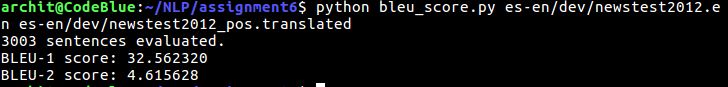
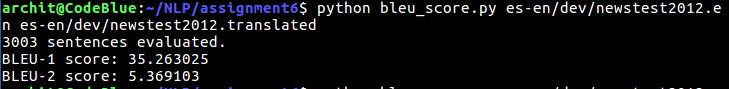
/es-en/test/newstest2013\_pos.translated

Excecution of solution2.py looks like this (takes about 30 mins to execute because of the size of the dataset):

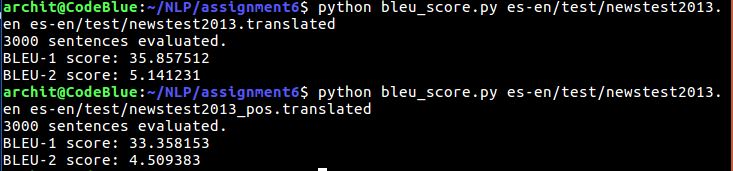


bleu\_score.py is run on these files and gives the following output:

Bleu Score on Dev test (IBM Model 1 and IBM Model 1 with POS tagging)



Bleu Score on Test test (IBM Model 1 and IBM Model 1 with POS tagging):



IBM Model 1 only uses Lexical translation for example lexical translation of the word huas is house, building, household and shell.

EM Algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values.There are two main applications of the EM algorithm. The first occurs when the data has missing values, due to problems with or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing/hidden parameters. The Expectation-Maximization (EM) algorithm is an iterative method to maximize the log-likelihood function for parameter estimation. Hence, convergence analysis of the EM algorithm will have final results as 0 and 1 where 1 is the probability of the word f in French language translated to e in English language. We can also observe that initially the convergence of all the words are same and as we do iterations the probability of the word which is similar to the word in foreign language increases and the probability of the word with less similarity decreases. We will iterate until the final values converge to either 0 or 1.

In this problem, the model is trained on Spanish/English corpus. The code implements expectancy maximization methodology for iterative computation of the probabilities of target language words given the source language word. For checking the performance of the translations

**Solution3**

**For the direct MT Problem:**

* Line1 in test set:

Active o desactive Bluetooth.

my best translation:

activate or deactivate bluetooth.

google translation:

Enable or disable Bluetooth.

* Line2 in test set:

Haga clic en el icono de Bluetooth en la barra de menús y seleccione Activar Bluetooth o Desactivar Bluetooth.

my best translation:

click or bluetooth Y from Activate in bluetooth icon he bluetooth from in the enus select make bar deactivate.

google translation:

Click on the Bluetooth icon in the menu bar and select Activate Bluetooth or Deactivate Bluetooth.

* Line3 in test set:

Consejo: Si no ve el icono de estado de conexión a la red WiFi o el icono de Bluetooth en la barra de menús, puede añadirlos.

my best translation:

advice: do not go see state from Connection he the net: or he icon from bluetooth in the bar from menus, can add them. to Wifi icon

google translation:

Tip: If you do not see the Wi-Fi connection status icon or the Bluetooth icon in the menu bar, you can add them.

* Line4 in test set:

Para wifi, seleccione el menú Apple Preferencias del Sistema y, a continuación, haga clic en Red.

my best translation:

for Wifi, select he menu Apple of the preferences system Y, to continuation, make click in net

google translation:

For Wi-Fi, select the Apple System Preferences menu, and then click Network.

* Line5 in test set:

Haga clic en wifi en la lista de la izquierda y después seleccione Mostrar estado de wifi en barra de menús.

my best translation:

make click in Wifi in the list from the left Y after select to show state from Wifi in bar from menus.

google translation:  
 Click on wifi in the list on the left and then select Show wifi status in the menu bar.

**For the statistical MT Problem:**

* Line1 in test set:

Una estrategia republicana para obstaculizar la reelección de Obama

my translation:

a strategy non-majestic for hinder the re-election of obama

google translation:

A Republican strategy to hinder Obama's re-election

* Line2 in test set:

Los dirigentes republicanos justificaron su política por la necesidad de luchar contra el fraude electoral .

my translation:

the leaders republicans justify their policy for the need of fight against the fraud electoral.

google translation:

The Republican leaders justified their policy because of the need to fight against electoral fraud

* Line3 in test set:

Ahora bien , el Centro Brennan considera esto último un mito y afirma que el fraude electoral es menos frecuente en los Estados Unidos que el

my translation:

now well, the centre as this last a myth and says that the fraud electoral is least often in the states united that the number of people that

google translation:  
 However, the Brennan Center considers the latter a myth and asserts that electoral fraud is less frequent in the United States than in the United States.

* Line4 in test set:

número de personas que mueren a causa de la caída de un rayo .

my translation:

die to cause of the fall of a like.

google translation:

number of people who die because of lightning.

* Line5 in test set:

De hecho , los abogados republicanos no han encontrado más que 300 casos de fraude electoral en los Estados Unidos en diez años .

my translation:

of fact, the lawyers republicans not have found more that 300 cases of fraud electoral in the states united in ten years.

google translation:  
 In fact, Republican lawyers have not found more than 300 cases of electoral fraud in the United States in ten years.

**Observation on compairing my translation with google translation on Statstical MT and Direct MT:**

* I believe my choice of data of already translated material (15 lines) for the solution1 was not great as it was manual and there weren’t a lot of things that I could have experimented on.
* Google does a very good job with changing the liternal meaning of the word with something that is more likely to be used with that senetnce in that particular usage
* A lot of my senetnces (translated by my model)in the Direct MT were almost similar to the Google’s translation.
* My translation for direct MT performed almost at par with Google’s translator for small sentencees
* Google’s translation is very good when it comes to long sentences and makes mpore sense when there is a lot of heavy context involved in the sentences.
* As far the statstical MT is considered the results were a little different. Here, my translation performed at par with Google’s translation even when there were big sentences. I guess this has to do with the fact that statastical MT works on the arrangement of the words and when it comes to a language like spanish, that I have selected, it is really needed and completely makes sense.
* Google tends to change literral meanings of the words to give a more contextual meaning to the translations which was a hit and miss in some of the cases. I feel you can’t go wrong in placing the literal meaning in the translations and that is where our model perfoemed perfectly.