**Assignment 6**

**Problem 1**

In this problem, we need to perform direct machine translation. Direct machine translation is the oldest approach to MT. The translation is based on large dictionaries and word-by-word translation with some simple grammatical adjustments. A direct translation system is designed for a specific source and target language pair.

I have created my corpus from the book Angels and Demons written by Dan Brown.

1. Link for book written in English language is

[http://romenplus.com/wp-content/uploads/2014/02/Dan-Brown-Angels-Demons-orig.pdf](http://romenplus.com/wp-content/uploads/2014/02/Dan-Brown-Angels-Demons-orig.pdf%20)

1. Link for book written in French language is <http://hullofromlabry.free.fr/Ebooks/Brown%20Dan%20%5BRobert%20Langdon%5D%20(2000)%20Anges%20et%20d%C3%A9mons%20(Angels%20&%20demons).pdf>

I have picked up 15 sentences from both French and English versions of the book. Further, I have left out 5 sentence pairs of French-English for testing. So, I have a set of 5 sentences which is be my test set and a set of 10 sentences which is my dev set. The dev set will be used for both translation problem and to produce evaluation during the development process. Test set will be used to see if the system I have developed generalizes the problem well.

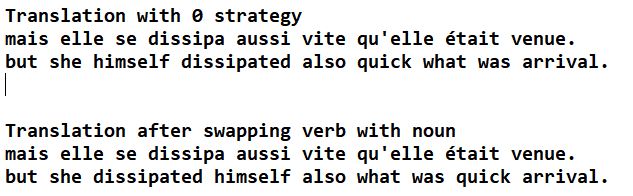
After I have created the dev and test set, I have created a bilingual French-English dictionary for each word in the corpus. For this, I have used Google Cloud Translation API which provides a simple programmatic interface for translating an arbitrary string into any supported language using state-of-the-art Neural Machine Translation.

The problem with normal translation there is only with word to word translation which uses the bilingual dictionary to translate each word from French to English. Hence, for word to word translation it will translate a sentence without taking in account the context of the sentence. So we won’t be able to make out any sentence because of the order in which the words are and grammatical patterns. For normal translation from French to English I have used word tokenization to translate the sentences that is given a character of sequence, tokenization is the task of chopping sequences into pieces called tokens.

The below output displays sentences with normal translation and with various strategies applied to the translation such as bigram model, trigram model, swapping the nearest verb with noun, bigram model with pos tagging, pos rearrangement and swapping parts of speech to improve translation.

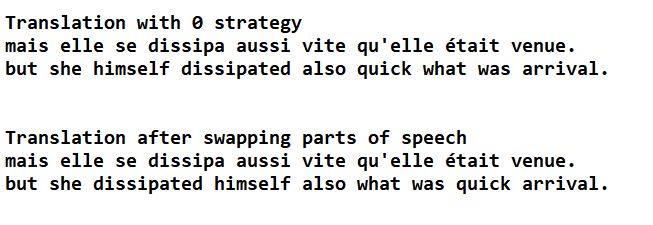
Observations made on dev set

Translation after swapping verb with nearest noun



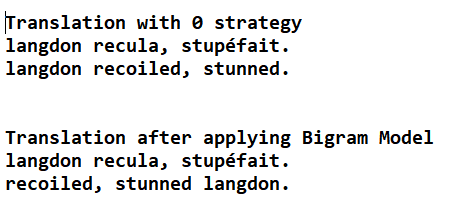
The problem here with normal translation is it doesn’t take into account parts of speech of the sentences while translating the sentence. Here we have [('but', 'CC'), ('she', 'PRP'), ('himself', 'PRP'), ('dissipated', 'VBD'), ('also', 'RB'), ('quick', 'JJ'), ('what', 'WP'), ('was', 'VBD'), ('arrival', 'NN'), ('.', '.')]. Both the personal pronouns come together which makes it difficult to read/make sentence of the sentence so to remove this disambiguation if two nouns come together I have swapped the nearest noun before the verb.

Translation after swapping parts of speech



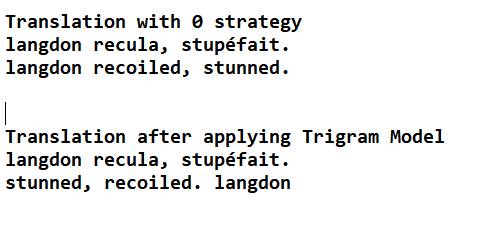
The problem with swapping noun with verbs was that it doesn’t take other parts of speech into account so to improve that in this strategy I have considered other parts of speeches as well. Normal translation the result is “but she himself dissipated also quick what was arrival.” [('but', 'CC'), ('she', 'PRP'), ('himself', 'PRP'), ('dissipated', 'VBD'), ('also', 'RB'), ('quick', 'JJ'), ('what', 'WP'), ('was', 'VBD'), ('arrival', 'NN'), ('.', '.')] Personal pronouns ‘himself’ and ‘she’ are coming together and also adjective is coming before the pronoun. I have swapped adjective, personal pronouns or possessive pronoun with the nearest verb or pronoun on the same translation and the output is “but she dissipated himself also what was quick arrival.” which is much better in comparison to normal translation and swapping verbs with nearest noun.

Translation after applying Bigram Model



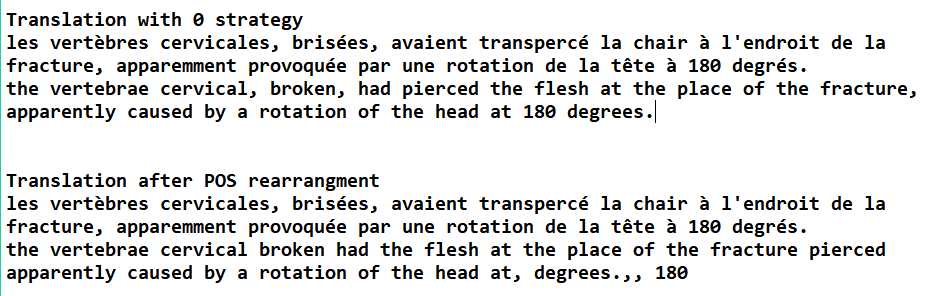
When we apply bigram model to above translation the meaning of the sentence changes. A stunned Langdon, recoils whereas in normal translation Langdon recoils and then he is stunned. The emphasizes should be given to the sentence thinking and what can be inferred from the entire sentence and not individual words. Here, both the sentences are formed well in their own context.

Translation after applying Trigram Model



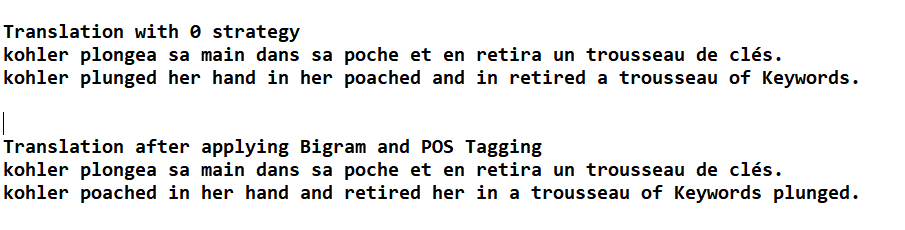
Another way to improve the translation is to apply trigram model to the translation.

Translation after POS re-arrangement



In this strategy I have rearranged pos tags. This feature is implemented using bigram model of parts of speech. For rearrangement, the probabilities of all permutations of four words is calculated and probability of the highest word is picked. The results in this case give better structure overall than the previous bigram model and trigram model.

Translation after applying Bigram and POS Tagging

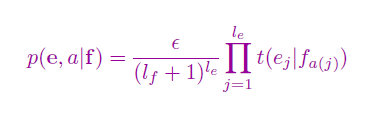


This strategy will set pos tags before training the model on English dev set using NTLK POS. The model is then trained on these tagged words and word with the highest probability is picked. With pos tag the normal bigram model is further improved.

Note: All of the strategies mentioned above have been implemented on the test set and the output of which can be found in output\_test.txt

**Problem 2**

IBM Model 1 only uses Lexical translation for example lexical translation of the word huas is house, building, household and shell. To calculate the translation probability we use the following formula.



Where f=(f1….,f(l\_f) is the foreign sentence of length l\_f

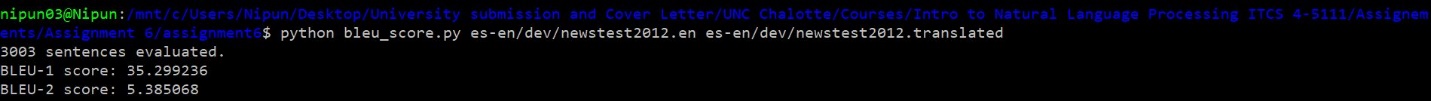
e=(e1…,e(l\_e)) is the English sentence of length l\_e

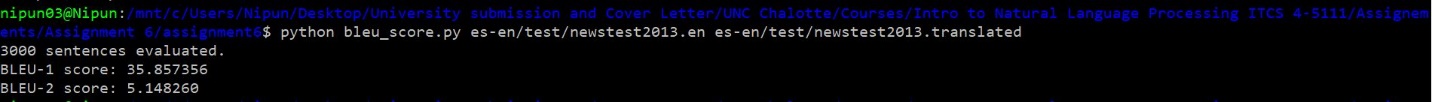
a is the alignment of each English word e\_j to a foreign word f\_i according to the alignment function a: j🡪i

€ is a normalization constant

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, Bleu Score is used the screenshot of which can be found below:

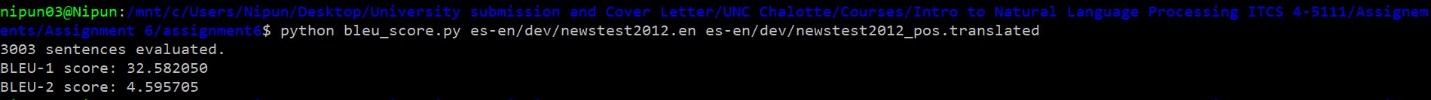
Bleu Score of Dev Set

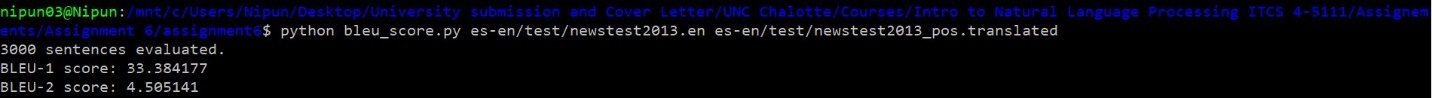
Bleu Score of Test

The problem with my model is that it is not considering out of vocabulary words that is the words appearing in the test set. I have left those words intact.

Improve the system

The IBM model needs to be enhanced with one improved feature, I will be using POS tagging. For POS tagging, every word in the training set is tagged with part of speech before it is trained. The source words are also tagged with parts of speech for translations. The following is the bleu score with the improvement system:

Bleu Score of Dev Set with POS tagging

Bleu Score of Test Set with POS tagging

The bleu score with POS tagging is comparatively low when compared without the POS tagging but it performs better with short sentences.

Original Sentence: I cannot understand how one can survive in New York on this money .

Normal Translation: how new york money is possible survive in with, for me is something that incomprehensible.

Translation with POS tagging: how is possible survive in new york with that money, for me is something incomprehensible.

The problem with IBM model is that there is no direct relation between the position of a word in the target language and position of a word in the source language. So when translating a phrase we often get incorrect arrangement of words. Another problem is that IBM model doesn’t have a big dictionary to select proper words.

To improve the efficiency of IBM Model one thing we can do is to train it on large corpus, another way is to remove the uniform probability and instead initialize it with ratio of the word and translation pair.

Comparing Machine Translation and IBM Model

I have used Spanish to English translation but these languages are very similar to one another. So when the semantics of two languages are similar that is they have similar character sets, direct machine translation works better because we are translating word by word from the dictionary and we already have the required structure.

IBM model is statistics based that is it does not refer to any dictionary for determine meaning of words which can be considered as a drawback.

**Problem 3**

The following are translations of the test set used in Problem 1 with Google Translate:

* Original Sentence: Ce symbole paraissait encore moins vraisemblable maintenant qu'il l'avait sous les yeux.

Google Translation: This symbol seemed even less likely now that he had it in front of him.

My Model Translation (with swapping partings of speech): this symbol appeared again less likely now had it under the eyes.

Original Translation: The symbol seemed even less conceivable now that he was not hear.

* Original Sentence: Le bourrelet de chair grillée était parfaitement dessiné et le symbole se détachait avec une absolue netteté.

Google Translation: The roast of grilled flesh was perfectly drawn and the symbol stood out with absolute clarity.

My Model Translation (with swapping partings of speech): the bead of flesh grilled was perfectly designed and the symbol detached himself with a absolute clearness

Original Translation: The raised broiled flesh was perfectly delineated the symbol flawlessly formed.

* Original Sentence: Il avait beau avoir détaillé cette blessure très attentivement sur la télécopie, la brûlure était beaucoup plus impressionnante dans la réalité.

Google Translation: Although he had detailed this injury very carefully on the fax, the burn was much more impressive in reality.

My Model Translation (with swapping partings of speech): had he beautiful to have detailed this injury very closely sure the fax, the burn was a lot more impressive in the reality.

Original Translation: Although Langdon had stared at the symmetrical wound a dozen times on the fax the burn was infinitely more commanding in real life.

* Original Sentence: Langdon se demanda si le frisson qui le parcourait était dû à l'air glacial ou à sa stupéfaction devant le spectacle qu'il venait de découvrir.

Google Translation: Langdon wondered if the thrill that ran through him was due to the cold air or his amazement at the sight he had just discovered.

My Model Translation (with swapping partings of speech): langdon asked himself if the thrill who the went about was of at the air glacial or at her stupefaction in front of the show had it of discover

Original Translation: Langdon wondered if the intense chill now raking through his body was the air conditioning or his utter amazement with the significance of what he was now staring at.

* Original Sentence: Son coeur cognait à grands coups tandis qu'il faisait le tour du cadavre pour lire le même mot, répété identiquement à l'endroit et à l'envers comme pour proclamer le génie de la symétrie.

Google Translation: His heart was beating fast as he circled the corpse to read the same word, repeated identically upside down as if to proclaim the genius of symmetry.

My Model Translation (with swapping partings of speech): his heart knocking at great shots while was it the tower of corpse for read the even word, say again identically at the place and at upside as for proclaim the genius of the symmetry.

Original Translation: His heart pounded as he circled the body, reading the word upside down, reaffirming the genius of the symmetry.

The problem with Google Translation is that the model has no knowledge about the syntax or semantics of natural languages, it can't make reasonable decisions about meaning when there are multiple ambiguous options. Where the sentences are longer Google Translate is not able to predict the sentence correctly. The reason for this that in French language, for example “tu” is used in a singular, informal situation, mostly for speaking to friends. “Vous” on the other hand, is used for formal situations. A problem lies when we are speaking with someone of higher authority, if we put just “you are funny” in Google translate, with no context, it might give you the “tu” version which the other person can consider as rude. Also, word like “à” have several meaning such as at, to, in, with, upon, by which Google doesn’t know how to use correct. Also, it depends upto the language from which translating English. For example, it's good with Chinese, as it has similar grammar to English. But with, Japanese it jumbles up the sentences this is because English has Subject-Verb-Object grammar and Japanese has Subject-Object-Verb so translating longer sentences from Japanese language to English is nothing more than gibberish.

Although Google Translate is quite good when it comes to short sentences or single word translation. But my model performs better when we have long sentences because when we translate word to word it is in a way better than statistical model used by google.