# Assignment 6: Machine Translation

Name: Sharath Chandra Bagur Suryanarayanaprasad
ID: 800974802
Email: sbagursu@uncc.edu

## Discussion:

Foreign language – Spanish

Spanish language doesn't have the free flow structure that english has. Spanish language switches the order of the pronoun and its verb, 'took she' instead of 'she took'. The case is similar with that of Noun and its adjective, 'girl beautiful' instead of 'beautiful girl'.

Spanish is an inflected language. The verbs are potentially marked for tense, aspect, mood, person, and number (resulting in some fifty conjugated forms per verb). The nouns form a two-gender system and are marked for number. Pronouns can be inflected for person, number, gender(including a residual neuter), and case, although the Spanish pronominal system represents a simplification of the ancestral Latin system.

## Output(Problem 1):

Spanish: Algunas de estas afirmaciones pueden ser exageradas, pero otras tienen buena evidencia que las sustentan.
Direct Translation: some of these affirmations may is exaggerated, but others have good evidence what the sustain.
Custom Model Translation: some of these affirmations may be exaggerated, but others has good evidence that the sustain.
Human Translation: Some of these claims may be exaggerated, but others have good evidence to support them.

Spanish: viernes al terminar el día hospitales y servicios de emergencia operaban para dar auxilio a los heridos, según las imágenes de la televisión egipcia.
Direct Translation: friday to the end the day hospitals and services of emergency operated for give help to the wounded, according the images of the tv egyptian.
Custom Model Translation: friday at the end the day hospitals and services of emergency operated for give help to the wounded, according the images of the tv egyptian.
Human Translation: Friday at the end of the day hospitals and emergency services operated to provide relief to the wounded, according to the images of Egyptian television.

Spanish: Lo más increíble del vídeo es que en ninguna impresora hay atasco de papel.

Direct Translation: the more amazing of the video be what in any printer there jam of paper.
Custom Model Translation: the more amazing of the video is that in any printer there jam of paper.
Human Translation: The most incredible thing about the video is that there is no paper jam in any printer.

Spanish: El presidente estadounidense, Donald Trump, condenó en un tuit el horrible y cobarde ataque.
Direct Translation: the president american, donald trump, condemned in a tweet the horrible and coward attack.
Custom Model Translation: the president american, donald trump, condemned in a tweet the horrible and coward attack.
Human Translation:  The US president, Donald Trump, condemned in a tweet the horrible and cowardly attack.

Spanish: La miel de abeja se ha utilizado terapéuticamente a lo largo de la historia.
Direct Translation: the honey of bee he he has used therapeutically to the long of the history.
Custom Model Translation: the honey of bee he he has used therapeutically to the long of the history.
Human Translation: Bee honey has been used therapeutically throughout history.

Spanish: La miel tiene propiedades ideales y comprobadas para el vendaje de heridas.
Direct Translation: the honey has properties ideals and checked for the bandage of wounds.
Custom Model Translation: the honey has properties ideals and checked for the bandage of wounds.
Human Translation: Honey has ideal and proven properties for bandaging wounds.

Spanish: Un bloque que Puigdemont ha situado en el bando de la represión y el miedo.
Direct Translation: a block what puigdemont he has located in the side of the repression and the fear.
Custom Model Translation: a block that puigdemont he has located in the side of the repression and the fear.
Human Translation: A block that Puigdemont has placed on the side of repression and fear.

Spanish: Sin embargo, estas propiedades no solo están relacionadas con la miel de manuka, se darán también con otros tipos de mieles de abeja.
Direct Translation: without embargo, these properties not alone están related with the honey of manuka, he give as well with others types of honey of bee.
Custom Model Translation: without embargo, these properties not alone están related with the honey of manuka, he give as well with others types of honey of bee.
Human Translation: However, these properties are not only related to manuka honey, but also to other types of bee honeys.

Spanish: Hasta el momento ningún grupo reivindica el ataque.
Direct Translation: until the moment any group claims the attack.
Custom Model Translation: so far the moment any group claims the attack.
Human Translation: So far no group claims the attack.

Spanish: Una parte mínima de la miel de manuka tiene distintos usos, pero la mayor parte de la miel de manuka que se vende en todo el mundo se come.
Direct Translation: a part minimum of the honey of manuka has different applications, but the higher part of the honey of manuka what he sells in all the world he eat.
Custom Model Translation: a part minimum of the honey of manuka has different applications, but the higher part of the honey of manuka that he sells in all the world he eat.
Human Translation: A minimal part of manuka honey has different uses, but most of the manuka honey sold worldwide is eaten.

Spanish: Testigos indicaron que los atacantes posicionaron alrededor de la mezquitas vehículos todo terreno y luego colocaron una bomba fuera del recinto.
Direct Translation: witnesses indicated what the attackers they positioned around of the mosques vehicles all ground and then placed a bomb outside of the enclosure.
Custom Model Translation: witness indicated that the attackers they positioned around of the mosques vehicles all ground and then placed a bomb outside of the enclosure.
Human Translation: Witnesses indicated that the attackers positioned all-terrain vehicles around the mosques and then placed a bomb outside the compound.

Spanish: Una bomba estalló en la mezquita Rawda en el Sinaí egipcio, antes de que los atacantes comenzaran a disparar contra las personas que asistían a la oración semanal, indicaron responsables.
Direct Translation: a bomb burst in the mosque rawda in the sinai egyptian, before of what the attackers begin to shoot against the people what attends to the sentence weekly, indicated responsible.
Custom Model Translation: a bomb burst in the mosque rawda in the sinai egyptian, before of that the attackers begin to shoot against the people that attends to the sentence weekly, indicated responsible.
Human Translation: A bomb explodes in the Rawda mosque in the Egyptian Sinai, before the attackers began firing at people attending the weekly prayer, officials said.

Spanish: No hay evidencia científica de que comer miel de manuka ayude a quienes padecen rinitis alérgica o fiebre del heno.
Direct Translation: not there evidence scientific of what eat honey of manuka help to who suffer rhinitis allergic or fever of the hay.
Custom Model Translation: not there evidence scientific of that eat honey of manuka help to who suffer rhinitis allergic or fever of the hay.
Human Translation: There is no scientific evidence that eating manuka honey helps those suffering from allergic rhinitis or hay fever.

Spanish: Cuando estaba en la universidad estudiaba todos los días.
Direct Translation: when was in the college studies everybody the days.
Custom Model Translation: when was in the college studies everybody the days.
Human Translation: When I was in college I studied every day.

Spanish: Napoleón estudia el problema y decide no atacar.
Direct Translation: napoleon study the problem and decide not attack.
Custom Model Translation: napoleon study the problem and decide not attack.
Human Translation: Napoleon studies the problem and decides not to attack.

## Problem 1:

The data was obtained from news articles from the following sources -

1. http://cnnespanol.cnn.com/2017/11/25/la-miel-de-manuka-y-sus-beneficios-que-es-sus-mitos-y-verdades/
2. http://www.ambito.com/904503-ascienden-a-305-los-muertos-por-el-atentado-en-una-mezquita-de-egipto
3. https://www.xataka.com/musica/567-impresoras-se-transforman-en-una-asombrosa-pantalla-en-uno-de-los-videoclips-mas-alucinantes-de-este-ano

The training data and test data collected from the above links is placed in the dev.txt(training dataset) and test.txt(testing dataset) files.
Each line in the above files is of the format: <Spanish sentence>:<Actual English translation>
Further a closed working dictionary of spanish words and their translations were generated based on the data in the corpus. The words with multiple meanings have also been considered into the dictionary wherever applicable.
After acquiring the required data files, we implement the two models, Direct Translation and Custom-translation.
Custom-translation model builds on top of the direct translation model by performing parts-of-speech tagging(POS- tagging)

Pre-processing strategies -

1. Used “Latin-1” encoding and “UTF-8” encoding wherever applicable to encode and work with spanish special characters.
2. Unigram Tagger and HiddenMarkovtagger were used to tag words of Spanish and English corpus for POS-tagging.
3. Maintained flags to indicate lowercase or uppercase of characters in words. This was done to ensure the final translation state.
4. Regular expression wherever applicable. (Ex: fetching words from sentences)

Direct Translation-

1. For each Spanish sentence extracted from the corpus, we split the sentences into words.
2. Using these words we lookup the bag of words in English that directly translate to the given word in Spanish.
3. We then perform the translation by placing all the words in the order in which they occurred in the original sentence.

Custom-Translation Model(POS tagging)-

1. part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.
2. In order to perform the pos-tagging we need to generate tags, preferably, for all words in English and Spanish.
3. We use the Brown corpus provided by the nltk package. Brown corpus is a pos-tagged corpus and hence this is suitable to be used in fetching the tags for english words in our corpus.
4. In this implementation, we also use the Spanish Corpus provided by nltk package. cess_esp is the corpus that contains data in Spanish. The cess_esp corpus is a tagged corpus. To further enhance the accuracy of tagging in this foreign language, we utilise the HiddenMarkovModel tagger to tag the cess_esp corpus data.
5. After tagging the corpus, we generate a file of all words in the cess_esp corpus. This file contains all words and their corresponding POS-tags. This is later used in the Custom-translation model to better translate the relevant word in context.
6. This Spanish words tagged file name is “Espanol-Tags.txt”.
7. The Custom-translation model works as follows-
    a. We first take each word in the input foreign sentence.
    b. For each word such word a dictionary is generated which contains the word, its Spanish pos-tag.This dictionary is then utilized to fetch the translated english word based on the pos-tag.
    c.
    d. First, the bag of all words that translate to this Spanish word is pulled from the dictionary.
    e. Next task is to eliminate words from this bag and only chose the word that best fits the translation based on the POS-tag.
    f. In order to do this we consider every index position of the tag and then take further steps to eliminate words from the bag.
    g. Example: Consider a word ‘El’ which has the tag : vsn0000. The word ‘El’ translates to [he,the].
    h. Utilizing the pre-defined english tags, shown below, we compare the Spanish word tag and its best fit in the translated sentence.
    i. After series of elimination we pick with the last word that is left in the bag of

words as the most suitable word that translates to the given spanish word.

Pre-defined English Tags-

```
noun = ('NN', 'NNS', 'NNP', 'NNPS', 'NR');
noun_plural = ('NNS', 'NNPS');
verb_imperative = ('BE', 'DO', 'HV', 'VB');
verb_infinitive = ('BE', 'DO', 'HV', 'VB');
verb_gerund = ('BEG', 'HVG', 'VBG');
verb_participle = ('BEG', 'BEN', 'HVG', 'HVN', 'VBG', 'VBN');
verb_present = ('VB','VBG', 'VBP', 'VBZ', 'BEG', 'BEM', 'BER', 'BEZ', 'DO', 'DOZ', 'HV', 'HVG',
'HVZ');
verb_past = ('VBD', 'VBN', 'BED', 'BEDZ', 'BEN', 'DOD', 'HVD', 'HVN');
verb_first_person = ('BED', 'BEDZ', 'BEM', 'BER');
verb_second_person = ('BED', 'BER');
verb_third_person  = ('BED', 'BEDZ', 'BEZ', 'DOZ', 'HVZ', 'VBZ');
verb_singular  = ('BED', 'BEDZ','BER', 'BEM', 'BEZ', 'DOZ', 'HVZ', 'VBZ');
verb_plural = ('BED', 'BER');
adjective = ('JJ', 'JJ$', 'JJR', 'JJS', 'JJT');
conjunction = ('CC', 'CS', 'ABX', 'DTX');
pronoun = ('ABL', 'ABN', 'ABX', 'AP', 'AP$', 'DT', 'DT$', 'DTI', 'DTS', 'DTX', 'PN', 'PN$', 'PP$',
'PP$$', 'PPL', 'PPLS', 'PPO', 'PPS', 'PPSS', 'WP$', 'WPO', 'WPS');
```

Output after Bleu Score evaluation(Custom-Model)-

```
5 sentences evaluated.
BLEU-1 score: 53.210317
BLEU-2 score: 11.616037
```

### Problem 2 IBM model 1:

The IBM model is based on the EM (Expectation maximization) model. The EM model is as follows:

1. initialize model parameters
2. assign probabilities to missing datasets
3. estimate model parameters from completed data
4. iterate steps 2 and 3 until convergence

In this implementation of the IBM model we have implemented the following-

1. Obtain the the english and corresponding spanish sentences from the europarl corpus.
2. Then for every word in the spanish sentences, we map the word to all the corresponding english sentences.

3. Further we calculate the probabilities of the spanish word mapping to the words in the english sentences.
4. Among the bag of words along with their probabilities calculated, we pick the word that has the maximum probability.
5. To improve the system we have implemented POS-tagging as well.
6. Further, once we have the spanish word and its corresponding list of english translated words and their probabilities, we parse through the test set and translate every sentence in the foreign language, word-to-word and then replace every word in the foreign language with the english word of miximum probability.

Output after Bleu score evaluation(IBM model):
BLEU-1 score: 58.709708
BLEU-2 score: 13.919337

The translated output file can be found by the name “myTranslation.txt”.

## Error Analysis:

1. Lack of pronoun/noun subject preceding verb-
    a. This is a contrasting feature of spanish where the verbs are used in a ‘lethargic’ manner. An example was the test sentence ’Not there evidence scientific of that eat honey of manuka help to who suffer rhinitis allergic or fever of the hay.’. The start of sentence must’ve been ‘There was no evidence’ which loosely translated to ‘Not there was evidence’. We considered a post-processing strategy to add a noun or pronoun before, if one wasn’t there, but that did not seem helpful to aid the purpose because it would be impossible to account for clauses of varying lengths preceding the verb.
2. Idioms and expressions-
    a. ‘Tengo dinero’ means ‘I have money’ and it is highly impossible to decipher that through direct translation methods. Similarly, ‘dar dinero por favor’ means ‘give money please’, but a direct translation would be ‘give money by favour’. In order to resolve this google translate uses user feedback to increase the accuracies. When a user gives feedback on right translation, the accuracy is increased in the model thereby making translations updated and more relevant.
3. Unnatural phrases-
    a. Some phrases, like ‘the girl likes him to jump’, is wrong because inclusion of unnecessary words in spanish doesn’t make sense in english.

## Problem 3 Google Translate:

Spanish: Testigos indicaron que los atacantes posicionaron alrededor de la mezquitas vehículos todo terreno y luego colocaron una bomba fuera del recinto.

Google Translation: Witnesses indicated that the attackers positioned all-terrain vehicles around the mosques and then placed a bomb outside the compound.

Spanish: Una bomba estalló en la mezquita Rawda en el Sinaí egipcio, antes de que los atacantes comenzaran a disparar contra las personas que asistían a la oración semanal, indicaron responsables.
Google Translation: A bomb explodes in the Rawda mosque in the Egyptian Sinai, before the attackers began firing at people attending the weekly prayer, officials said.

Spanish: No hay evidencia científica de que comer miel de manuka ayude a quienes padecen rinitis alérgica o fiebre del heno.
Google Translation: There is no scientific evidence that eating manuka honey helps those suffering from allergic rhinitis or hay fever.

Spanish: Cuando estaba en la universidad estudiaba todos los días.
Google Translation: When I was in college I studied every day.

Spanish: Napoleón estudia el problema y decide no atacar.
Google Translation: Napoleon studies the problem and decides not to attack.

As of late 2016, machine translation used by Google Translate has seen great recent advancements enabled by Deep Learning. In September of 2016 Google announced Google Neural Machine Translation system (GNMT), a new machine translation system based on Artificial Neural Networks and Deep Learning.
In the new system, Google used Recurrent Neural Networks (RNN) which are well known to perform well on sequences (of words and phrases). By using this approach, Google has been able to continually improve quality of translations by enabling their systems to take into consideration not only source words and phrases, but also broader contexts of where they appear in sentences, and what are the other words and phrases around them.
These issues have been known for a long time to be key difference between human and simple machine translation techniques such as the ones implemented in this assignment.

## Execution Instructions:

Problem 1-

Run the python file customTranslation.py as follows-

```
python3 customTranslation.py
```

Problem 2(IBM model1) -

Run the python file ibmModel1.py as follows-

```
python3 ibmModel1.py newstest2012.en newstest2012.es
```

This above python file writes all translations to myTranslation.txt file. Use this file to compute Bleu-score.

Calculating Bleu Score-

Run the follwoing:

```
python bleu_score.py newstest2012.en myTranslation.txt
```