Report

Name: Krishna Sreenivas

Student ID: 800984436

Email ID: [ksreeniv@uncc.edu](mailto:ksreeniv@uncc.edu)

**Description:**

*Problem 1:*

Direct Translation:

In this method, Spanish sentences were chosen from the source and were split into words. A unique English translation dictionary was created with multiple meanings in English. One to one mapping of words is performed and appended to a list and the list is outputted.

Improved Direct Translation:

Here, Parts of speech tagging concept was used. Parts of speech tagger, tags the words based on pats of speech such as Nouns, pronouns, verbs etc. All the sentences in Spanish corpus which was provided by *cess\_esp* (which is a package built in nltk) was used to were tagged based on parts of speech. The corpus used for English text is Brown corpus which had inbuilt methods for POS tagging. Additionally, Unigram Tagger and Hidden Markov Tagger was used to tag these corpuses.

The tags generated were stored in a file which was then used to translate the words. Tags of Spanish text were compared to that of English. Words were picked which matched the corresponding tags. Now, the task was to remove words whose tags were not matching with the English ones. Only those tags were considered which had the correct positioning of indices between them. The words corresponding to the tags were appended to a list and then outputted.

Pre/Post Processing Strategies:

Following are the strategies that were implemented in the model.

1. Tokenizing the individual sentences into chunks of words.
2. Preserving the Spanish encodings by specifying the format UTF-8.
3. Stripping the new line characters.
4. Removal of special symbols such as punctuations etc., from the Spanish dev and test set prior to execution of the program.
5. Lowering the sentences of both Spanish and English corpuses.

Sources:

The Spanish dev and test sentences were obtained from various articles from a news outlet called El Comercio. Link for which is below.

Link- http://www.elcomercio.es/

Bleu score:

Direct MT

5 sentences evaluated.

BLEU-1 score: 50.450761

BLEU-2 score: 11.046417

Improved Direct MT

BLEU-1 score: 44.288249

BLEU-2 score: 8.576257

Output:

###############Spanish Text###############

TEST case

10 : Ustedes, que probablemente recorran con cierta costumbre el trayecto hasta Asturias, habrán comprobado cuánta mejora ha supuesto en tiempo, en seguridad y en comodidad, ha detallado.

11 : Una vez terminada esta vía, existe otra ambición relacionada con las comunicaciones del noroeste, que los consejeros de Infraestructuras de Galicia, Castilla y León y Asturias han plasmado en una declaración institucional que expresa una demanda común.

12 : Ha respaldado así la petición de un corredor ferroviario atlántico que dé prioridad al enlace con los puertos de Vigo, A Coruña, Avilés y Gijón y se complete posteriormente con una línea de Costa entre A Coruña, Avilés y Gijón, similar al existente en el Corredor Mediterráneo, que más tarde conecte con Francia.

13 : La única pega de la A-8 es su demora, la tardanza acumulada en su finalización, por lo que espera que la petición conjunta de las tres comunidades merezca la atención del Ministerio de Fomento para que la pereza administrativa y la desidia presupuestaria, siempre acechantes, no posterguen estas reivindicaciones.

14 : Ha defendido asimismo el planteamiento porque, en su opinión, redundaría en beneficio de todo el noroeste peninsular, aún necesitado de un fuerte impulso a sus infraestructuras.

###############Direct Translation###############

TEST case

10 : you, what probably travel with certain habit the journey until asturias, they will have checked how much improvement he has supposed in weather, in security and in comfort, he has detailed.

11 : a time completed this one vía, exists other ambition related with the communications of the northwest, what the counselors from infrastructure from galicia, castile and lion and asturias have captured in a declaration institutional what expresses a demand common.

12 : he has backed up a yes the petition from a runner railway atlantic what from priority to the link with the ports from vigo, to coruña, avilés and gijón and he complete later with a line from coast between to coruña, avilés and gijón, similary to the existing in the runner mediterranean, what plus late connect with france.

13 : the only hit from the to-8 is his delay, the delay cumulative in his ending, by the what wait what the petition joint from the three communities deserve the attention of the ministry from promotion for what the sloth administrative and the lazy budgetary, always stalkers, do not postpone these claims.

14 : he has defended likewise the approach why, in his opinión, would be redundant in benefit from all the northwest peninsular, yet needy from a strong impulse to their infrastructure.

###############Improved Direct Translation###############

TEST case

10 : you, what probably walk in spite of certain habit the journey till asturias, they will have checked how much improvement have so-called at weather, at security and at comfort, have elaborate.

11 : a time finished this one vía, exists other ambition allied in spite of the communications from the northwest, what the counselors of infrastructure of galicia, castile and lion and asturias have captured at a declaration institutional what expresses a demand common.

12 : have protected a yes the petition of a runner railway atlantic what from priority to the link in spite of the ports of vigo, at coruña, avilés and gijón and he complete later in spite of a line of coast between at coruña, avilés and gijón, similary to the existing at the runner mediterranean, what plus late connect in spite of france.

13 : the only hit of the at-8 is his delay, the delay cumulative at his ending, by the what wait what the petition joint of the three communities deserve the attention from the ministry of promotion to what the sloth administrative and the lazy budgetary, always stalkers, do not postpone these claims.

14 : have defended likewise the approach why, at his opinión, would be redundant at benefit of all the northwest peninsular, yet needy of a strong impulse at their infrastructure.

###############Manual Translation###############

TEST case

10 : You, who probably travel in a certain way to Asturias, have verified how much improvement has been in time, security and comfort, has detailed.

11 : Once this route is over, there is another ambition related to the Northwest communications, which the infrastructure advisors of Galicia, Castilla y León and Asturias have expressed in an institutional declaration that expresses a common demand.

12 : It has thus supported the request of a Atlantic railway corridor that prioritises the link with the ports of Vigo, A Coruña, Avilés and Gijón and is subsequently completed with a coastline between A Coruña, Avilés and Gijón, similar to the existing one in the Mediterranean corridor. , which later connects with France.

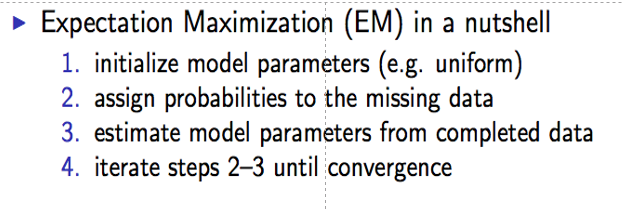
13 : The only drawback of the A-8 is its delay, the delay accumulated in its completion, so it hopes that the joint request of the three communities deserves the attention of the Ministry of Development so that the administrative laziness and the budgetary apathy, always stalking , do not defer these claims.

14 : He also defended the approach because, in his opinion, it would benefit the whole of the peninsular Northwest, still in need of a strong impulse to its infrastructures

*Problem 2:*

IBM Model:

The Algorithm of the IBM model is defined below:



Firstly, the sentences from English and Spanish Europarl corpuses are split into individual words and one to one mapping is performed. Then, probabilities and uniform probabilities of those words are calculated. Further, the concept of gradient descent is used to fine tune the probability. The corresponding words are then appended into a list. Next, based on the Spanish words present in the corpus, the equivalent English words are chosen. These words are chosen based on the maximum probability values. These words are then appended to a list in order to form a sentence, and the list is then written to a file. The number of iterations performed i.e. the gradient descent value is set to 10.

Error Analysis:

* The order of subject-verb-predicates differed in Spanish. So, when direct mapping or improved direct mapping was performed the gist of the meaning or intention behind the utterance could be obtained. But the order of subjects, verbs and predicates in English were jumbled. This problem could be mitigated by using advanced tagging techniques which takes multiple parts of speech into consideration.
* Another error was that the certain words in Spanish language had multiple meanings. The context mattered a lot. This was a real problem in texts such as idioms. Even the IBM model couldn’t obtain the correct and meaningful translation. Hence, I believe Google has an upper hand in obtaining correct translations since they would have applied advanced methodologies.
* Also, found some words in Spanish which didn’t make much sense when translated to English, the liberal use of few words was translated as is, without taking the context into consideration.
* And the major problem in Spanish is that there’s no one-to-one mapping in usage of tenses. Hence translating a word based on the tense requires additional POS tagging techniques which takes these into consideration.

Bleu score:

The Bleu score of translated text using IBM model is as below:

3000 sentences evaluated.

BLEU-1 score: 53.532324

BLEU-2 score: 12.318258

Problem 3:

Below are the test sentences of Direct MT translated using Google translate.

1. Ustedes, que probablemente recorran con cierta costumbre el trayecto hasta Asturias, habrán comprobado cuánta mejora ha supuesto en tiempo, en seguridad y en comodidad, ha detallado.

Translated using Google: You, who probably travel with a certain habit the journey to Asturias, will have seen how much improvement has been made in time, safety and comfort, has detailed.

1. Una vez terminada esta vía, existe otra ambición relacionada con las comunicaciones del noroeste, que los consejeros de Infraestructuras de Galicia, Castilla y León y Asturias han plasmado en una declaración institucional que expresa una demanda común.

Translated using Google: Once this route is finished, there is another ambition related to communications from the northwest, which the Infrastructure Councilors of Galicia, Castilla y León and Asturias have expressed in an institutional declaration that expresses a common demand.

1. Ha respaldado así la petición de un corredor ferroviario atlántico que dé prioridad al enlace con los puertos de Vigo, A Coruña, Avilés y Gijón y se complete posteriormente con una línea de Costa entre A Coruña, Avilés y Gijón, similar al existente en el Corredor Mediterráneo, que más tarde conecte con Francia.

Translated using Google: It has thus backed the request of an Atlantic rail corridor that gives priority to the link with the ports of Vigo, A Coruña, Avilés and Gijón and is subsequently completed with a coastline between A Coruña, Avilés and Gijón, similar to that existing in the Corridor. Mediterranean, which later connects with France.

1. La única pega de la A-8 es su demora, la tardanza acumulada en su finalización, por lo que espera que la petición conjunta de las tres comunidades merezca la atención del Ministerio de Fomento para que la pereza administrativa y la desidia presupuestaria, siempre acechantes, no posterguen estas reivindicaciones.

Translated using Google: The only downside of the A-8 is its delay, the delay accumulated in its completion, so it hopes that the joint request of the three communities deserves the attention of the Ministry of Public Works so that the administrative laziness and budget deficit, always stalking , do not postpone these claims.

1. He has also defended the approach because, in his opinion, it would be to the benefit of the entire peninsular northwest, still in need of a strong boost to its infrastructure.

Translated using Google: He has also defended the approach because, in his opinion, it would be to the benefit of the entire peninsular northwest, still in need of a strong boost to its infrastructure.

Statistical MT

1. Las nuevas leyes electorales exigen que los electores presenten un documento de identidad con una fotografía , además de una prueba de ciudadanía americana .

Translated using Google: The new electoral laws require that voters present an identity document with a photograph, in addition to proof of American citizenship.

1. Antes de las elecciones de 2006 , ningún estado americano exigía que los electores presentaran un documento de identidad con fotografía .

Translated using Google: Prior to the 2006 elections, no American state required that voters present a photo ID.

1. El Estado de Indiana fue el primero en exigirlo .

Translated using Google: The State of Indiana was the first to demand it.

1. El Tribunal Supremo de los Estados Unidos confirmó en 2008 la constitucionalidad de la ley de Indiana .

Translated using Google: The Supreme Court of the United States confirmed in 2008 the constitutionality of the Indiana law.

1. Antes de 2004 ningún estado exigía la prueba de ciudadanía para votar .

Translated using Google: Before 2004, no state required proof of citizenship to vote.

Explanation:

The translations provided by Google translate were incredibly accurate and the translation error was very much negligible, this is because of the algorithm google had implemented which goes by the name of Google Neural Machine Translation System (GNMT). This system not only known for translation on sequences of words and phrases but also can take context into consideration.

The words of the sentences translated by Google were in same order as compared in Spanish. But however, the tenses of the words especially verbs and gerunds were handled excellently while translating the sentence from Spanish to English. This is partly because, the Google translate employs user feedback experience which allows the user to rate the translation provided by the system and allows to suggest improvements.

The main drawback of the translator created in this assignment is that it performs translation based on words instead of phrases. Google translate performs well in this regard, since it has large set of well defined and annotated expressions for phrases.