DEVELOPMENT OF WORD BASED MACHINE TRANSLATION SYSTEM

A

Mini Project

Report

Submitted

To

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BY

SUVAM BASAK

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JIS COLLEGE OF ENGINEERING
BLOCK 'A' PHASE 'III' KALYNAI NADIA-741235
December, 2014

CERTIFICATE

We hereby notify that the work which is being presented (Report) entitled **"project name"** and submitted to the Department **Computer Science and Engineering** of JIS college of Engineering is an authentic record of our own work carried out during a period from **"duration"** under the supervision of **Name & Designation of supervisor(s),** CSE Department.

Signature of Candidates

R.No.

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Signature of Supervisor(s)

Date: 20th November

Name & Designation of Supervisor(s)

Head
CSE Department

JIS College of Engineering

(An Autonomous Institution)

Department of Computer Science and Engineering

UG Project Proposal'2015-2016

Title: DEVELOPMENT OF WORD BASED MACHINE TRANSLATION SYSTEM USING "**NATURAL LANGUAGE PROCESSING IN PYTHON**".

Abstract:

"Natural language processing" here refers to the use and ability of system to process sentences in a natural language such as English. NLP is highly essential in artificial intelligence. We used Python for it, because Python is a simple yet powerful programming language which excellent functionality for processing data.

Software Requirement : Some several free software packages is required for our processing purpose. Python version 2.7, Natural language processing toolkit (NLTK), NLTK-Data, NumPy etc.

Language translation : At first we used a English-Spanish Parallel Corpus [The Europarl parallel courpus is extracted from the proceedings of the Europran Parliament]. Then we extracted words from the corpus to another file using a python program which was written by us. After that we needed to map English words with Spanish words. For the mapping purpose we used Anymalign.py and Treetagger was used for finding parts of speech. After Anymalign processing we extracted first and second column (English and Spanish) from the output file. Then we made a Program to translate English text to Spanish text. The program takes input only in English text and split it word by word and then searches for corresponding Spanish word then displays.

MEMBER DETAIL :

1.Suvam Basak
2.Shaif Aslam
3.Sayantan Pal
4.Sandip Pore

Mentor's Name : Sainik Kr. Mahata

CONTENT

1 Introduction

2 Types of Machine Translation

3 Things required

4 Assignments

5 Advantage of machine translation

6 Disadvantage of machine translation

7 Resource

8 Reference

9 Conclusion

Introduction:

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) language. As such, NLP is related to the area of human-computer interaction. Many challenges in NLP involve Natural language understanding, that is, enabling computer to derive meaning from human or natural language input, and others involve natural language generation. Now here our job is Machine translation, so Machine translation some thing referred to by the abbreviation MT is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

On a basic level, MT performs simple substitution of words in one language for words in another, but that alone usually cannot produce a good translation of a test because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problems in linguistic typology, translation of idioms, and the isolation of anomalies.

Current machine translation software often allows for customization by domain or profession , improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domain where formal or formulaic language is used . It follows that machine translation of government and legal documents more readily produces output that conversation or less standardised text.

Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously indentified which words in the text are proper names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators and, in a very limited number of cases, can even produce output that can be used as is.

The progress and potential of machine translation have been debated much through its history. Since the 1950s, a number of scholars have questioned the possibility of achieving fully automatic machine translation of high quality. Some critics claim that there are in-principle obstacles to automatizing the translation process.

Types of machine translation :

Total four types of machine translation are available.

Those are :

1. Word based machine translation. [WBMT]
2. Example based machine translation. [EBMT]
3. Phrase based machine translation. [PBMT]
4. Statistical machine translation. [SMT]

From four types of machine translation our topic is Word based machine translation [WBMT].

Things required :

1.Python programming language : Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale.

Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library.

Python interpreters are available for installation on many operating systems, allowing Python code execution on a wide variety of systems. Using third-party tools, such as Py2exe or Pyinstaller, Python code can be packaged into stand-alone executable programs for some of the most popular operating systems, allowing the distribution of Python-

based software for use on those environments without requiring the installation of a Python interpreter.

CPython, the reference implementation of Python, is free and open-source software and has a community-based development model, as do nearly all of its alternative implementations. CPython is managed by the non-profit Python Software Foundation.

2.Natural Language Toolkit : NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resource such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

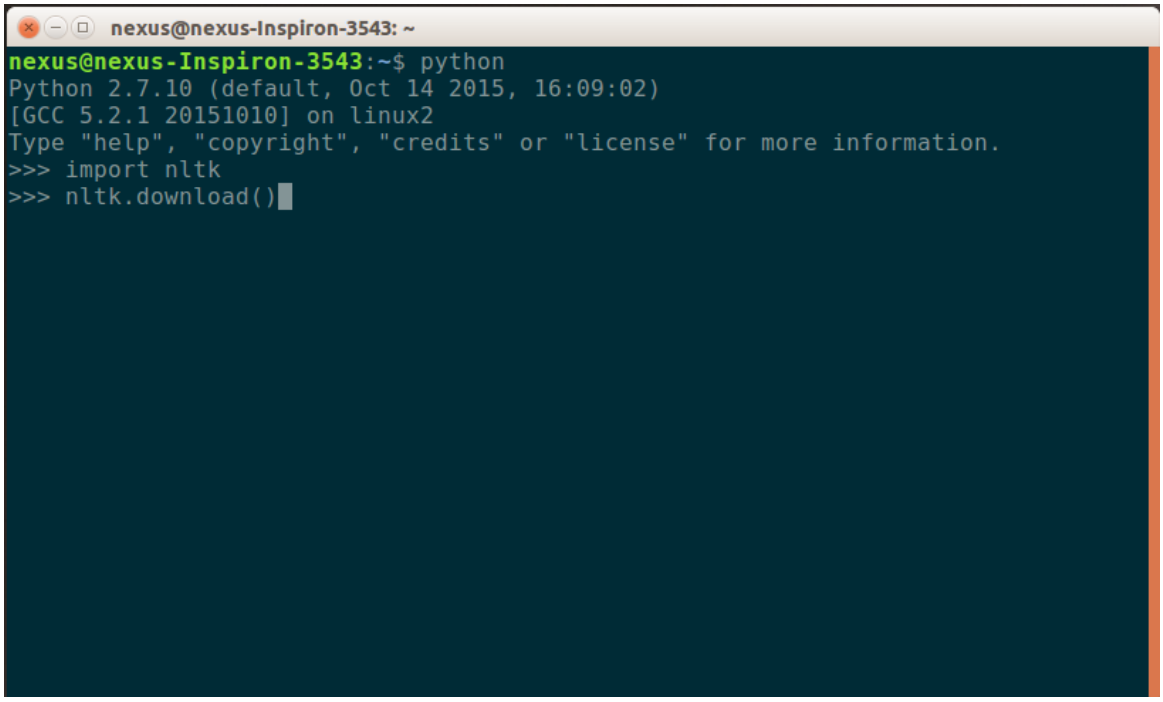
NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Precessing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed. And in case of our project we used Python 2.7 version)

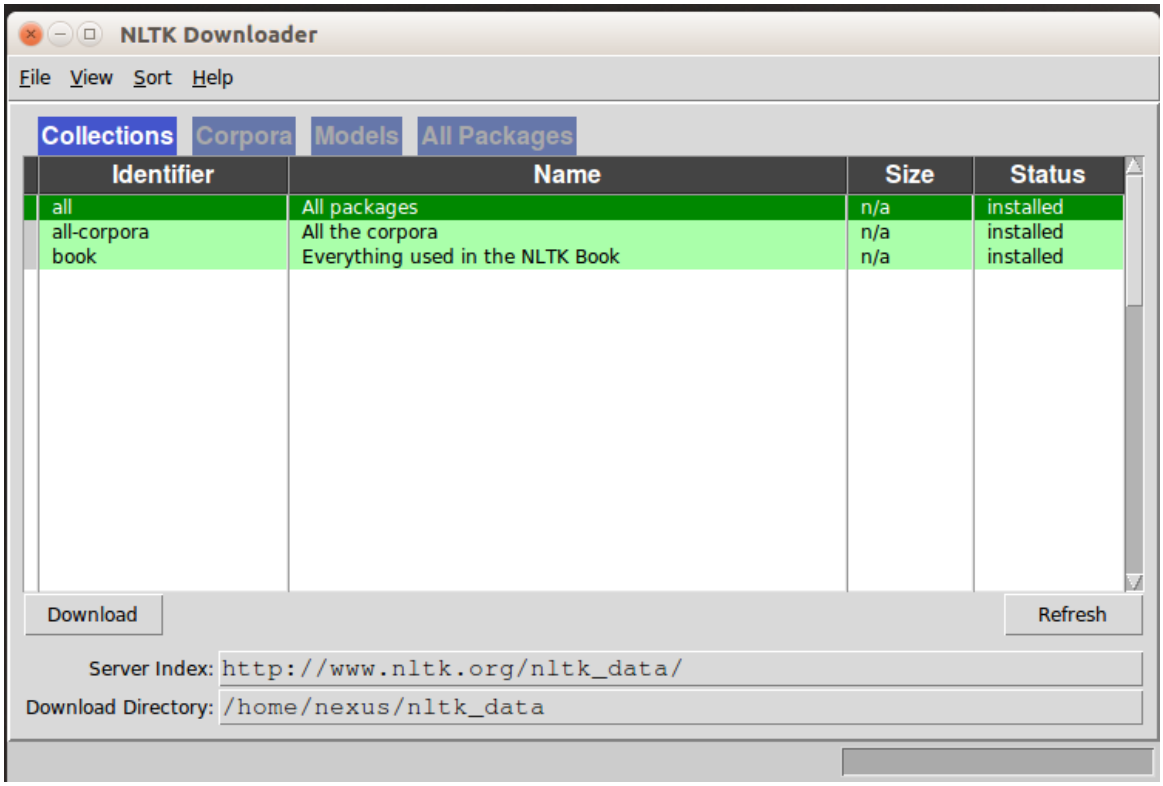
NLTK installation command :

1. sudo apt-get install python-pip
2. sudo pip install -U numpy
3. sudo pip install -U nltk
4. For testing installation : Run python then import nltk

3.NLTK_data : NLTK comes with many corpora, toy grammars, trained models, etc. A complete list is posted at:http://nltk.org/nltk_data/. After installing NLTK then used NLTK's data downloader to download nltk data.



Installation command



NLTK Downloader

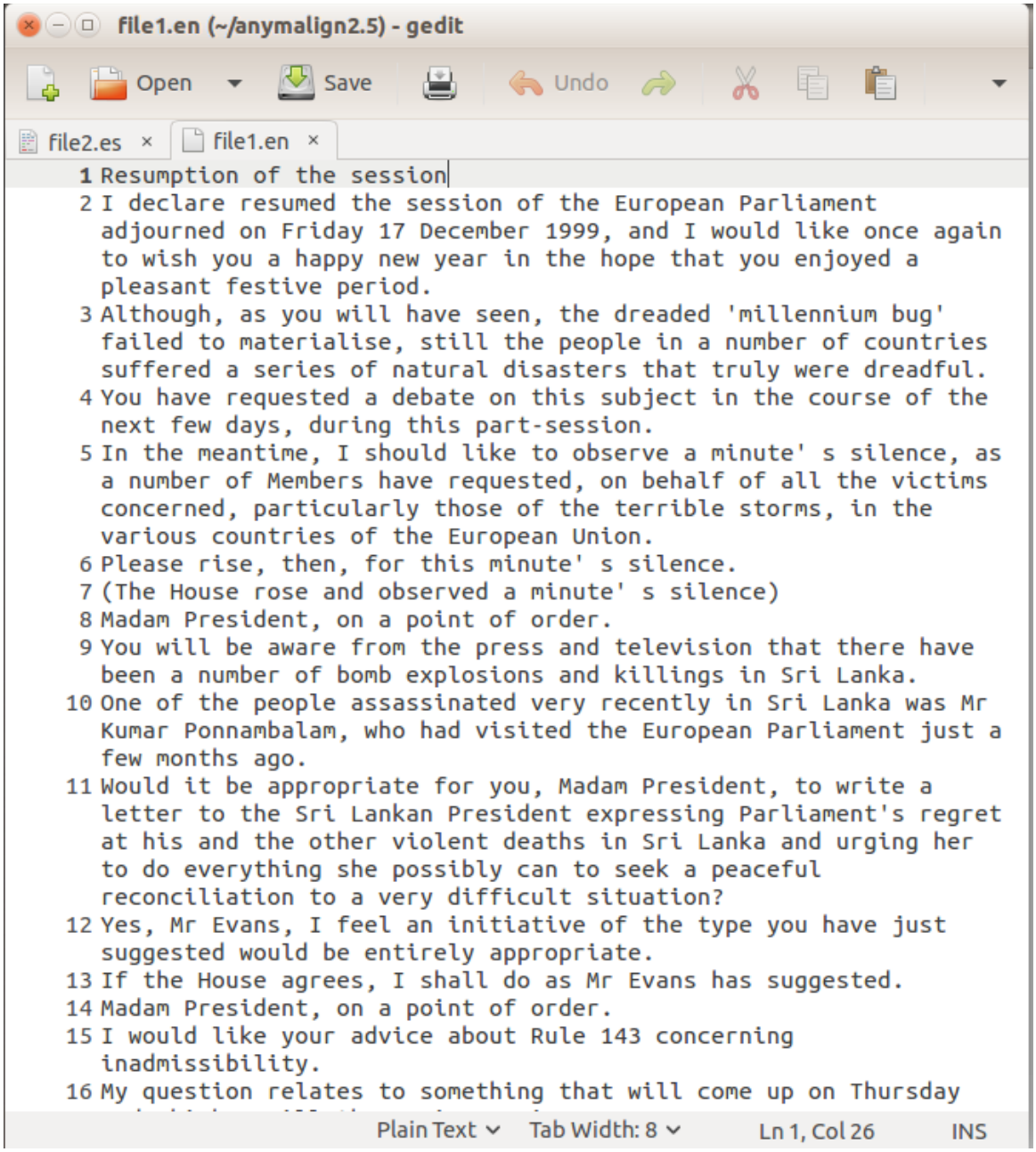
4.Parallel corpus : A parallel corpus is a corpus that contains a collection of original texts in language L_1 and their translations into a set of languages L_2 ... L_n. In most cases, parallel corpora contain data from only two languages. Closely related to parallel corpora are 'comparable corpora', which consists of texts from two or more languages which are similar in genre, topic, register etc. without, however, containing the same content. http://www.statmt.org/europarl/ this link used to download the English to Spanish parallel porpus.

5.Word alignment : There are three types of word alignment are available. Those are :

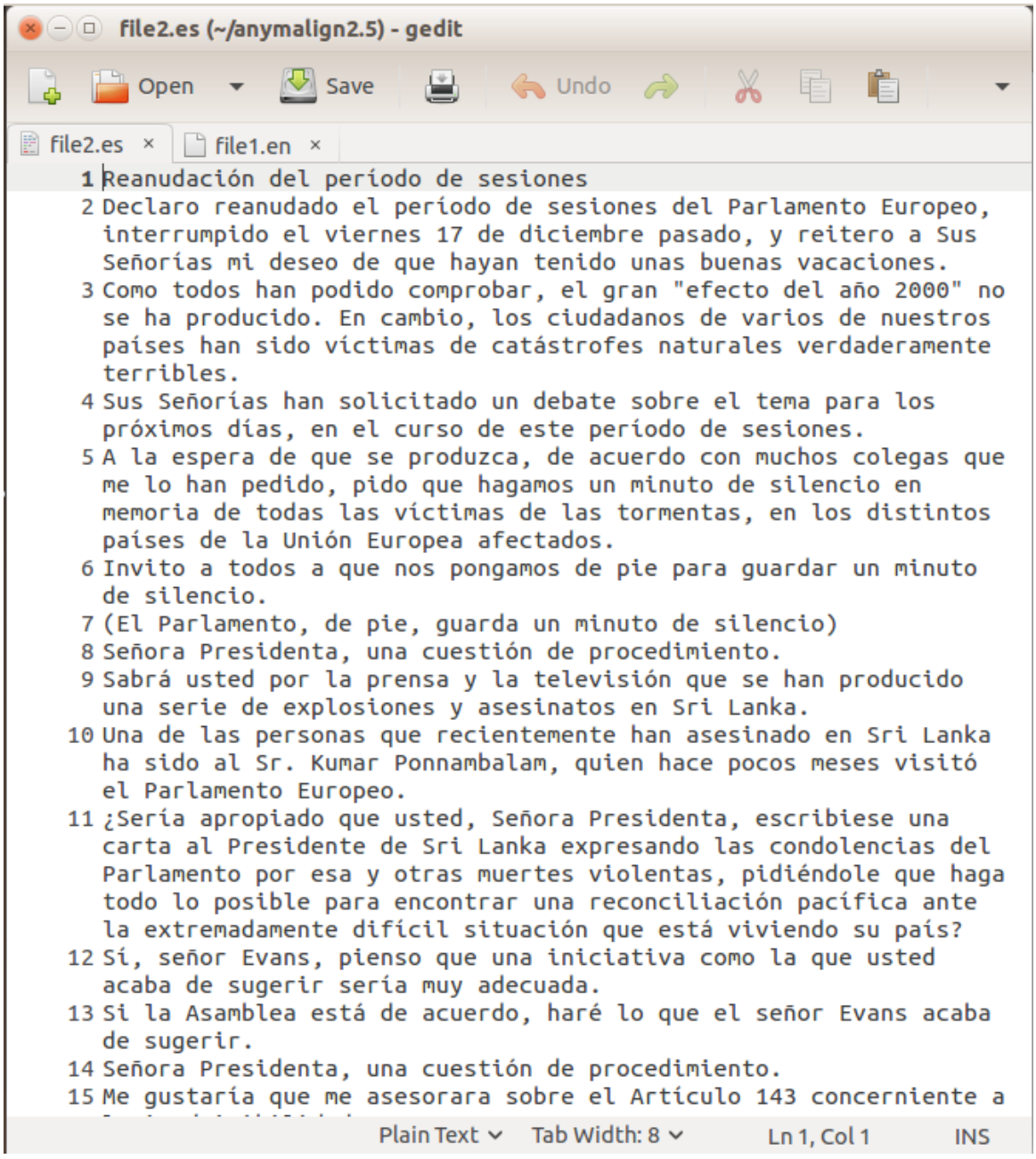
1. GIZA++
2. MGIZA++
3. Anymalign

In case of out project, we used anymalign2.5 . Anymalign is a multilingual sub-sentential aligner. It can extract lexical equivalences from sentence-aligned parallel corpora. Its main advantage over other similar tools is that it can align any number of languages simultaneously. This package is downloaded from https://anymalign.limsi.fr/. Other info like how to use it, input file format, out put file format also available in this link.

Input to the system : anymalign.py can read input data in separate files, where each file may contain one or more languages. Typically, multilingual texts are available in separate files, one file per language, one sentence per line, the corresponding lines being translations (all files have the same number of lines). For instance, you may have a tiny trilingual corpus in English, French, and German, where each file is made of 2 lines:
Here out input is English and Spanish text file.

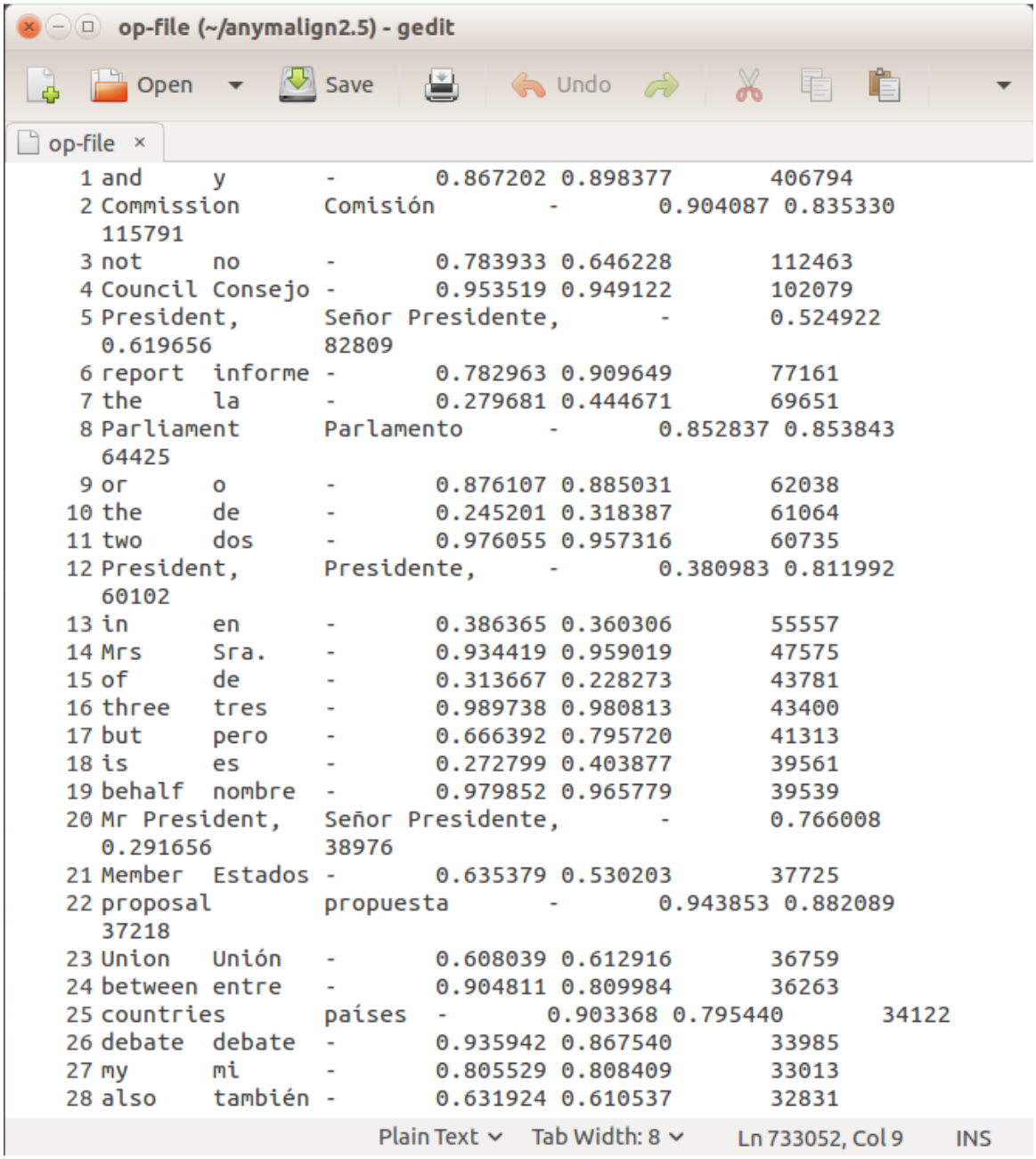


Frist input for anymalign.py



second input for anymalign.py Output of the system :

Output files have basically the same format as the all-languages-in-one input file. After running this process almost 1 hour we get this kind of output.



Output of anymalign

Assignments :

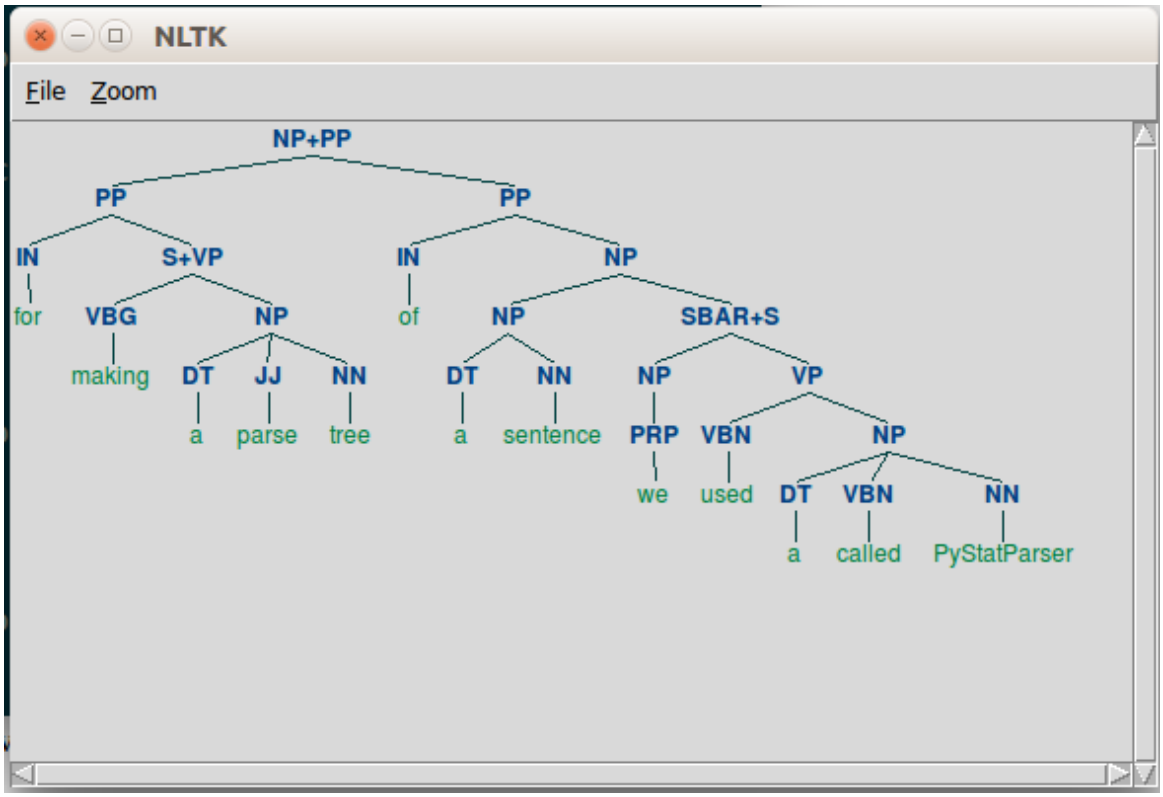
POS tagger : Our first job was POS tagger. POS tagger is build-in feature of the nltk module. For determine parts of speech of a sentence at first we need to split the

sentence word by word. So at first we take input the sentence as a string. Then split it word by word using word tokenize (built it function). Then using pos_tagger we got a list of words with parts of speech tag.

```
[('In', 'IN'), ('lexical', 'JJ'), ('analysis', 'NN'), (',', ','), ('tokenization
', 'NN'), ('is', 'VBZ'), ('the', 'DT'), ('process', 'NN'), ('of', 'IN'), ('break
ing', 'VBG'), ('a', 'DT'), ('stream', 'NN'), ('of', 'IN'), ('text', 'NN'), ('up'
, 'RB'), ('into', 'IN'), ('words', 'NNS'), (',', ','), ('phrases', 'NNS'), (',',
 ','), ('symbols', 'NNS'), (',', ','), ('or', 'CC'), ('other', 'JJ'), ('meaningf
ul', 'JJ'), ('elements', 'NNS'), ('called', 'VBN'), ('tokens', 'NNS'), ('.', '.'
), ('The', 'DT'), ('list', 'NN'), ('of', 'IN'), ('tokens', 'NNS'), ('becomes', '
NNS'), ('input', 'VBP'), ('for', 'IN'), ('further', 'JJ'), ('processing', 'NN'),
 ('such', 'JJ'), ('as', 'IN'), ('parsing', 'NN'), ('or', 'CC'), ('text', 'NN'),
('mining', 'NN'), ('.', '.')]
>>>
```

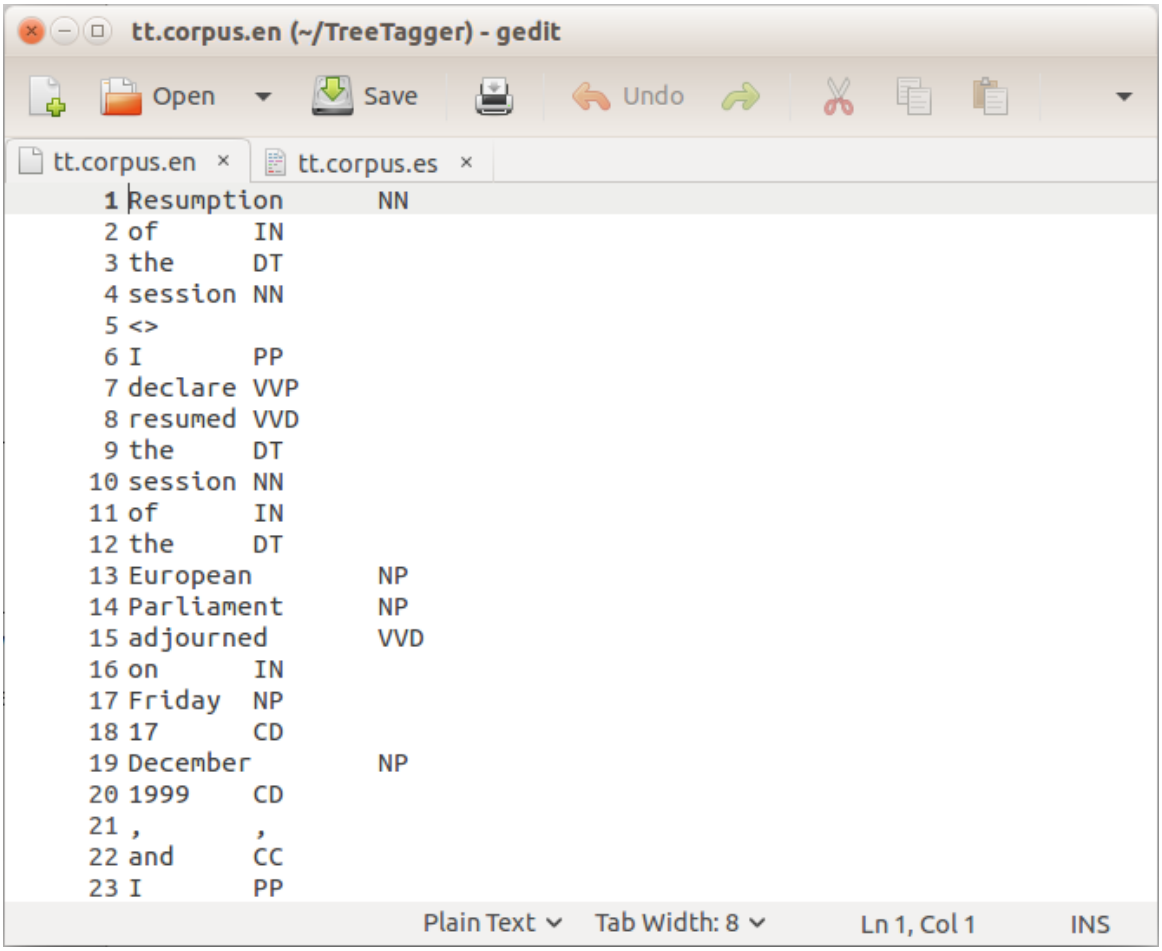
Output of pos tagger

Parse tree : For making a parse tree of a sentence we used a called PyStatParser. Which is available in www.github.com. Using this module we made a parse tree of a sentence.

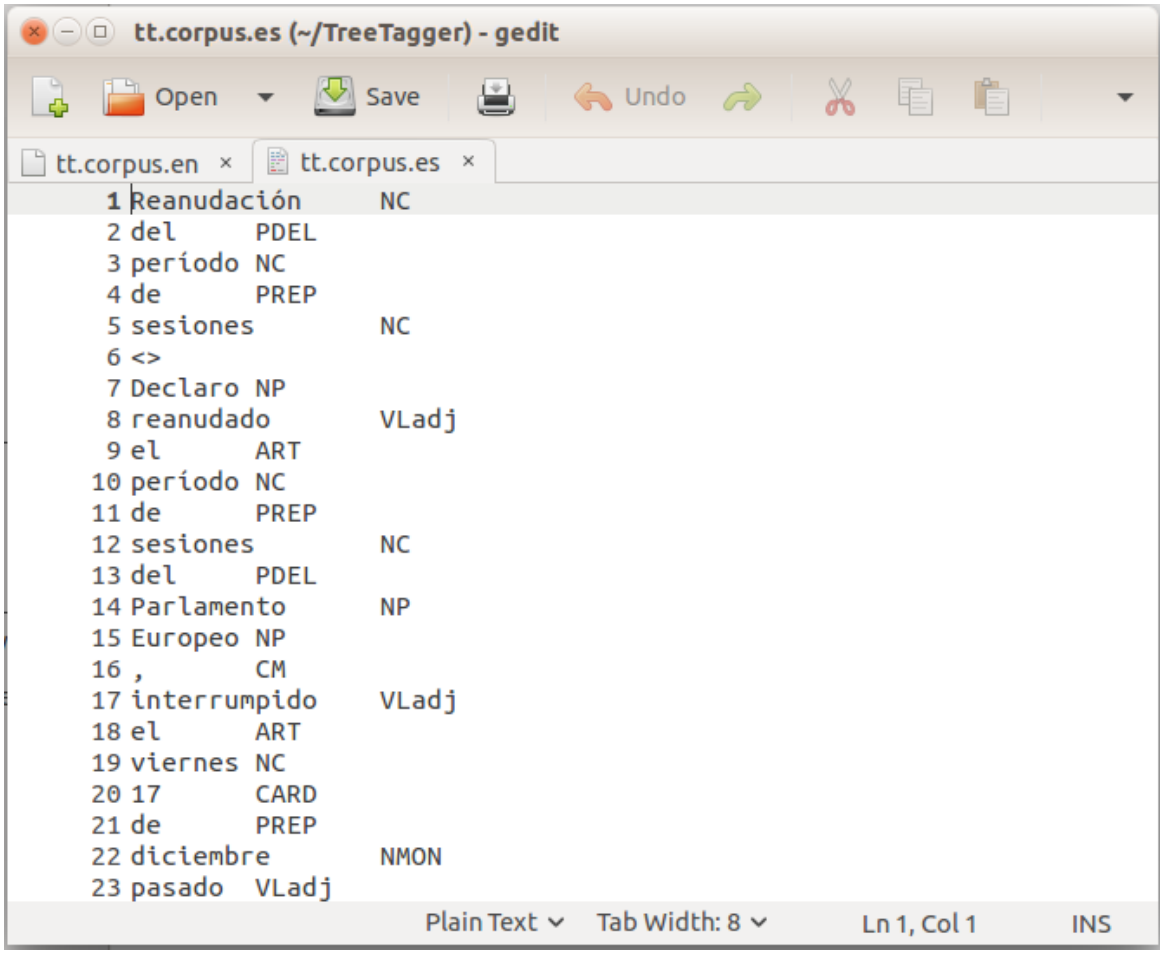


Parse tree

Tree tagger : Before starting word based translation we need to find out the pars of speech of English and Spanish words. Using TreeTagger we did this job.



Tree tagging of English

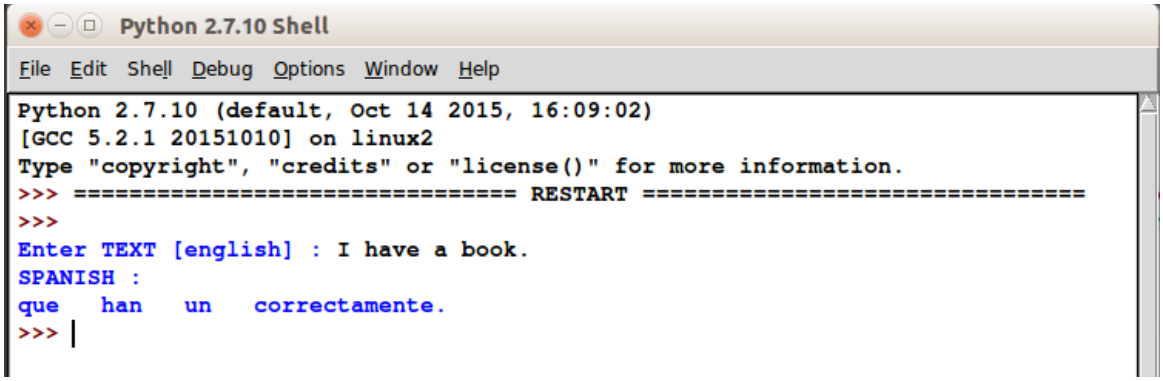


Tree tagging of Spanish

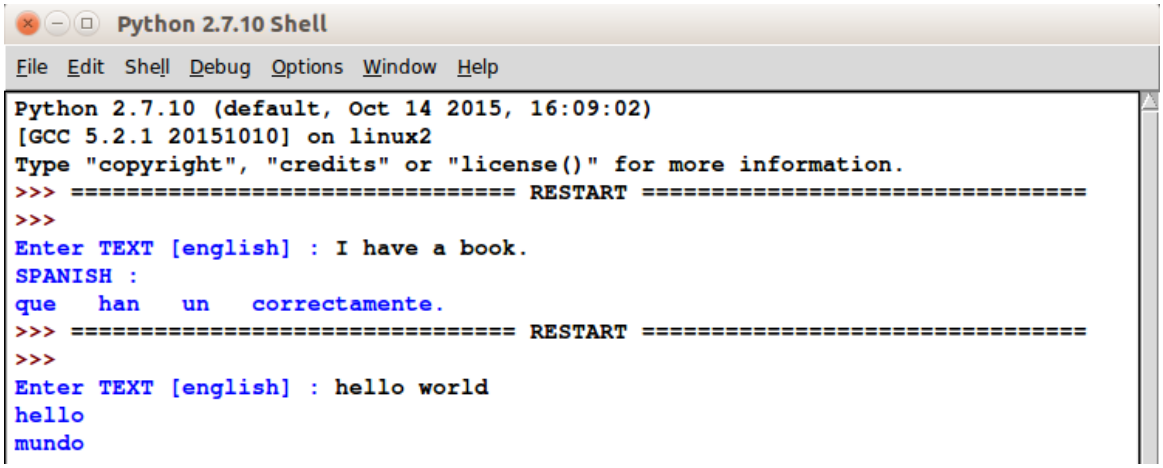
English to Spanish translator : In case of translator first job is making a file which contains two columns. The first column contains English words and second column contains Spanish words which is same meaning with left column's English word. Then our second job is a program to translate English word in Spanish word.

The program takes input as a string then split that string in word by word. The program search in the first column for each word. If the word is available in that list then picked up the Spanish word which is present right side of that English word. And then displays.

If the English word is not present in the list then it prints the English word (which is input) as output.



Translation of English words into Spanish



When English word not found

Advantage of Machine Translation :

The rate of machine translation is exponentially faster than that of human translation. The average human translator can translate around 2,000 words a day. Multiple translators can be assigned to a given project to increase that output, but it pales in comparison to machine translation. Machine translation can generate thousands of words each minute. One should note that the output of machine translation is not in its final useable form right away, but in certain scenarios it can be quite useful. Even when adding a post-editing step, machine translation takes a fraction of the time that human translation takes.

Cost : The cost of machine translation is also a mere fraction of the cost of human translation. While the major cost for standard translation projects is the cost of the

translators, the biggest cost for machine translation projects is in the post-editing process, which ends up saving the client a great deal.

Adaptation : Machine translation can memorize key terms and phrases that are used within a given industry. This leads to translations that are very consistent across the entire file, something that is more difficult to achieve when using multiple human translators.

Disadvantages of Machine Translation :

1. Accuracy is not offered by the machine translation on a consistent basis. You can get the gist of the draft or documents but machine translation only does word to word translation without comprehending the information which might have to be corrected manually later on.

2. Systematic and formal rules are followed by machine translation so it cannot concentrate on a context and solve ambiguity and neither makes use of experience or mental outlook like a human translator can.

Resource :

Reference :

1. http://www.tutorialspoint.com/

2. http://www.nltk.org/

3. http://www.nltk.org/book/

4. https://github.com/

5. http://www.statmt.org/europarl/

6. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

7. http://stackoverflow.com/

Conclusion :

So I would like to conclude that benefits of machine translation allover the world. It is too difficult for a human to learn lots of languages. But a machine translation program can translate any language to another language. This technology makes our life familiar with unknown languages. A popular translator Google Translate is a free multilingual Statistical Machine Translation service provided by Google to translate text, speech, images, or real-time videos from one language into another. It offers a web interface, mobile interface(for Android & iOS) it can be use as a browser extensions. It supports 90 language at various level.