Final Project - L-555
Dr. Francis Tyers
Hanan Aloraini

A Part of Speech (POS) Tagger Program for Moroccan Arabic Corpus

Developing a Part of Speech (POS) tagger can be challenging, especially for languages with complex linguistic structures like Arabic, can be extremely challenging, given the programming skills required for this task coupled with the complexity of the language. Building a POS tagger is crucial for enhancing the processing and understanding of the Arabic text to be able to integrate it for various Natural Language Processing (NLP) applications.

Arabic, with its dialectical diversity and complex syntax and morphology, requires careful considerations before creating a POS tagger for the language. Even as a native Arabic speaker, I found it challenging to tag certain words, especially since I worked on Moroccan Arabic, which varies significantly from Modern Standard Arabic or Saudi Arabic. I believe this represents the value of computational linguistics as these computational applications require a linguist and would not be possible to be completed entirely by a specialized programmer.
As for the tagging decisions, I relied on guidelines from the Universal POS tags project.

A POS tagger can be developed using different approaches depending on the purpose of the tagger and the type of data available. Some of the major methods used for POS tagging include rule-based tagging and probabilistic (statistical) taggers. The tagger developed for this project falls under the statistical approach, which works by calculating the frequency of each tag for every word. This method uses a pre-tagged corpus for training, which allows for the estimation for various parameters. Compared to the rule-based approach, the probabilistic approach is generally less labor-intensive and more cost-effective (Abumalloh et al., 2016).

A significant challenge that I faced was working with the CONLL-U file format, which stores annotated corpus. This format is helpful as it facilitated the annotation of morphological analysis, part of speech, syntactic, and semantic information. Moreover, for the purpose of this project, I tagged 10 sentences from the Arabic corpus. However, when I ran the code, I realized

that the sample size was insufficient for producing more accurate output. This made me realize the necessity for tagging a larger and more varied dataset for training purposes. In the future, I plan to expand on this project and work on more training set that will enhance the accuracy of the tagger.

The program used for this project can be found in the practical section under the name "tagger.py". The program begins by initializing a dictionary, reading, and processing the model file, identifying the most frequent tag, and tagging the new input. This program utilizes two "for loops", the first loop is designed to read, split and extract the word and its corresponding tag, while the second loop processes lines from the standard input and tags new sentences.

Moreover, the process also involved using a program to segment "segmenter.py" to break down the texts into segments and a tokenizer program "tokeniser.py" to split the text into individual words. These pre-processing steps are crucial for effective for effective functioning of the tagger, which then works on the language model developed for this project.
Below is a screen shot from the output:

```
# sent_id = 43329
# text = السفة طبق مغربي .
1       السفة   _       NOUN    _       _       _       _       _       _
2       طبق     _       NOUN    _       _       _       _       _       _
3       مغربي   _       NOUN    _       _       _       _       _       _
4       .       _       PUNCT   _       _       _       _       _       _

# sent_id = 43330
```

Also, when I looked at the output, I saw this quote from the Holy Quran that seems to be typed in a different font than the rest of the text which made it appear in this way:

```
# sent_id = 43430
# text = وَيَا قَوْمِ لَا يَجْرِمَنَّكُمْ شِقَاقِي أَنْ يُصِيبَكُمْ مِثْلُ مَا أَصَابَ قَوْمَ نُوحٍ أَوْ قَوْمَ هُودٍ أَوْ قَوْمَ صَالِحٍ وَمَا قَوْمُ لُوطٍ مِنْكُمْ بِبَعِيدٍ .
1       وَيَا    _       NOUN    _       _       _       _       _       _
2       قَوْمِ   _       NOUN    _       _       _       _       _       _
3       لَا      _       NOUN    _       _       _       _       _       _
4       يَجْرِمَنَّكُمْ _       NOUN    _       _       _       _       _       _
5       شِقَاقِي _       NOUN    _       _       _       _       _       _
6       أَنْ     _       NOUN    _       _       _       _       _       _
7       يُصِيبَكُمْ _       NOUN    _       _       _       _       _       _
8       مِثْلُ   _       NOUN    _       _       _       _       _       _
9       مَا      _       NOUN    _       _       _       _       _       _
10      أَصَابَ  _       NOUN    _       _       _       _       _       _
```

This was interesting to witness and made me realize the possible challenges that I might face when working on a POS tagger in the future.

References

Abumalloh, R. A., Al-Sarhan, H. M., Ibrahim, O., & Abu-Ulbeh, W. (2016). Arabic part-of-speech tagging. *Journal of Soft Computing and Decision Support Systems, 3*(2), 45-52.