CS5012: Practical 1 – Part-of-Speech Tagging and Unknown Words

Student ID: 220029176

Date of Submission: 18/03/2023

Word Count: 1437

1. Introduction

The objective of this assignment was to produce a natural language processing model to determine parts of speech tags, based on a given corpus using the Viterbi algorithm. Three corpora were provided to implement and evaluate: English, French and Ukrainian. Additionally, the model aimed to make informed predictions on unknown words based on UNK tags. The UNK tags are based on common patterns seen within the language Corpus. The report will examine the accuracy of the tagging before and after the introduction of UNK tags, explain how the UNK tags were chosen, and discuss the factors that may have contributed to the different accuracies obtained for each corpus.

1.1 Compilation and Execution Instructions

1. Download the zip file "CS5012-Assignment1-220029176" and unzip it to a chosen folder location.
2. Open a terminal prompt from within the src folder location.
3. Enter the following command into the terminal "/usr/local/python/bin/python3 P1.py"

Parameter Estimation

First, the corpora were broken down into a list of sentences of words and a list of sentences of tags using the conllu module. Then, the lists were converted into bigrams and processed into a frequency distribution to form the basis of the Viterbi algorithm. The Viterbi algorithm was utilised to create a state trellis which is backtracked to find the most likely tags for a given sentence. The final accuracy of the part-of-speech tagging for each corpus was then calculated and printed.

The UNK tags are introduced to improve the accuracy of the Viterbi algorithm. Any hapax legomena within the training data were replaced with the UNK tags. The frequency distribution of the words was then built on these new tags. Within the test data, any unknown words located were also replaced with the UNK tags.

2. Evaluation

Basic Model

The basic model does not manipulate any data to increase the accuracy of the model. To recreate the results seen in Appendix 1, comment lines 247 to 251 and rerun the program. The basic implementation provides a foundation for all future work towards an effective model. The EN accuracy was 90.1% which, given the specification, is an accurate model for the English corpus, as seen in Appendix 1. The French and Ukrainian corpora were less accurate than English, with the lowest accuracy being Ukrainian at 84.6%. There are several factors that contribute to this result. From a data perspective, English has the most data available as seen is Appendix 2, this could influence the results of the accuracy by creating a selection bias. However, this is unlikely to be the case within this model as Ukrainian has a much larger availability of data than French yet a lower accuracy. This indicates a linguistical issue with the model rather than a data discrepancy.

From a linguistic perspective, the Ukrainian language has a complex morphology with a rich system of inflection and declension (Melymuka, 2017). This complexity can make it challenging to identify the correct part of speech for a given word. Similarly, French also has a complex morphology and syntax, with many irregularities and exceptions to grammatical rules (Franck, 2016). The French language also includes liaison, elision, and many other phenomena that can make it challenging to identify the correct part of speech for a given word. In contrast, English has a relatively simple morphology and syntax compared to Ukrainian and French. Furthermore, many of the irregularities in English have been standardised or reduced over time, making it easier to identify the correct part of speech for a given word (Lieberman, 2007).

With UNK Tags

For the model with UNK tags, the UNK tags were chosen based on morphology research on each language. However, as the developer for this project can only speak English, there is linguistic bias within the decision of each tag. This may impact the effectiveness of the tags on the accuracy of the model as well as the potential analysis of each individual tag. A list of relevant tags was created, and each tag was tested against a base UNK model, where every unknown tag was allocated UNK. The best tags were chosen for the final model. As the tags were based on the accuracy of the model, confirmation bias may have occurred. This may lead to overfitting of the model, leading to poor generalisation to new or unseen data.

For English, -ly, in-, capitalised, -ed, -ness, -ing, -ful, dis-, mis- and, re- were chosen for the UNK tags based on the information within Appendix 4. Comparing the accuracy shown in Appendix 3, with the model without UNK tags in Appendix 1, the accuracy improved from 90.15% to 91.04%. This increase of 0.89% indicates the chosen UNK tags were effective at providing the correct parts of speech tags. Interestingly, by just adding the capitalisation tag, the accuracy is greater than the basic model. This can either indicate a strong correlation between capital letters and a part of speech tag or a large quantity of unknown words appear at the start of the sentence. Linguistically, English is a relatively consistent language when it comes to prefix and suffix patterns for determining parts of speech which is likely to be a determining factor toward the increase in accuracy (St. Clair, 2009).

For French, -al, -ion, -er, -eur, -ment, and -e were chosen based on the information seen within Appendix 5. Comparing the accuracy shown in Appendix 3, with the model without UNK tags in Appendix 1, the accuracy improved from 91.7% to 91.8%. This 0.1% increase is significantly smaller than that seen in English. This may be due to linguistic bias, as the developer was not able to utilise their own knowledge when determining the French tags and was reliant on alternate sources. Linguistically, many prefixes and suffixes in French can have multiple meanings depending on the context, making it difficult to determine the correct part of speech to use. Additionally, French has many irregular verbs and noun forms, including masculine and feminine nouns, which can make it challenging to use prefix and suffix patterns to determine parts of speech (Franck, 2016).

For Ukrainian, -ння, -тва, -ати, -ити, -ти, -ють, -ий, -о, -and -и were chosen based on the data seen within Appendix 6. Comparing the accuracy shown in Appendix 3, with the model without UNK tags in Appendix 1, the accuracy decreased from 84.6% to 83.6%. There are many factors that could have led to this decrease of 1%. Firstly, like French, many prefixes and suffixes in Ukrainian can have multiple meanings depending on the context, making it difficult to determine the correct part of speech (Melymuka, 2017). Additionally, Ukrainian has many irregular words that do not follow the typical prefix and suffix patterns for determining parts of speech. For example, the noun "людина" (person) is irregular and does not follow

the typical feminine noun ending "-ка." (Melymuka, 2017). Furthermore, the developer has very limited understanding of Ukrainian which may have resulted in inaccuracies, finding it difficult to interpret the alternate writing system and Cyrillic alphabet.

In summary, there were many determining factors within each corpus which caused a variance in the UNK tags accuracy, the largest of which was the linguistic influence. In English and French, using prefix and suffix patterns to determine parts of speech can be an efficient and accurate way to process and analyze large amounts of text quickly and accurately. This could be due to the semantic change of the language over time causing a narrowing in lexical ambiguity. In Ukrainian, using prefix and suffix patterns to determine parts of speech could potentially be an efficient and accurate way to process and analyze large amounts of text quickly. However, we have shown that this is not the case for the provided corpus. This may be because Ukrainian has many irregular words that do not follow the typical prefix and suffix patterns for determining parts of speech. As well as the developers limited understanding and the possibility that the prefix and suffix patterns are too broad which can lead to overgeneralisation. Overall, we have demonstrated the potential advantages and disadvantages of utilizing prefixes, suffixes, and capitalisation to influence the ability to accurately determine parts of speech within a given language.

To enhance the reliability of the results, it would be necessary to address potential sources of linguistic bias. Collaborating with native speakers to refine the model can help mitigate any existing bias. Additionally, to avoid confirmation bias, it may be beneficial to use unseen test data or refine the method of determining unknown UNK tags. Once these biases are mitigated, the model can effectively perform in the target language, as well as potentially be applied to new languages.

References

Franck, Floricic. (2016). “French morphology”.
https://www.researchgate.net/publication/312330064_French_morphology

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. (2007). "Quantifying the evolutionary dynamics of language." Nature 449, no. 7163: 713-716.

Melymuka, M., Lapesa, G., Kisselew, M., & Padó, S. (2017). Modeling derivational morphology in Ukrainian. In IWCS 2017—12th International Conference on Computational Semantics—Short papers.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. Cognitive Science, 33(7), 1317-1329.

Appendices

```
fr:
0.9178082191780822
en:
0.9014820396885205
uk:
0.8463727710983331
```

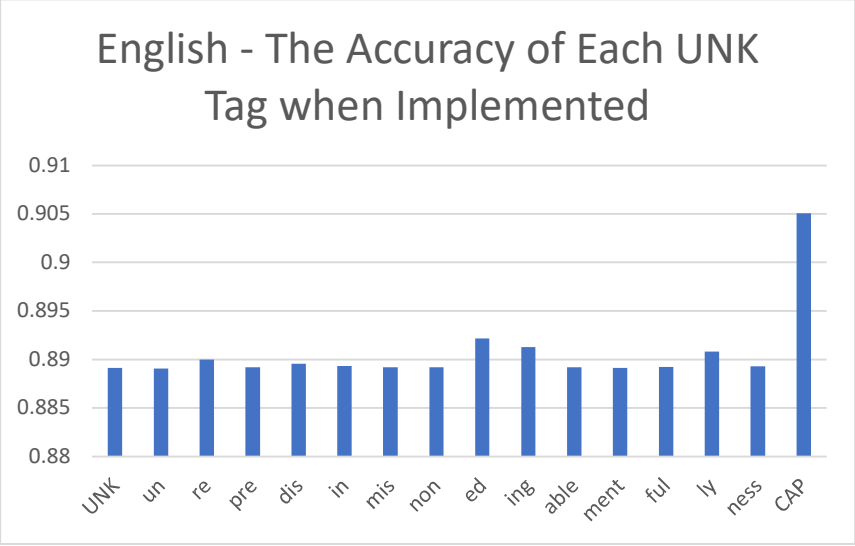
Appendix 1: The accuracy score of the basic model

```
fr:
1288 training sentences
840 test sentences
en:
6911 training sentences
1096 test sentences
uk:
5521 training sentences
898 test sentences
```

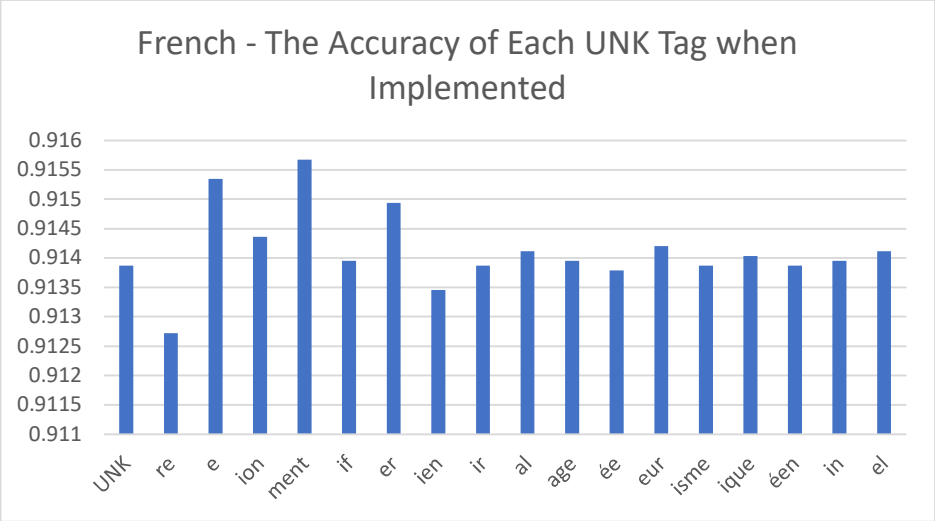
Appendix 2: Availability of data from each Corpus

```
fr:
0.9183003855303092
en:
0.9103240391861341
uk:
0.8358018237788233
```

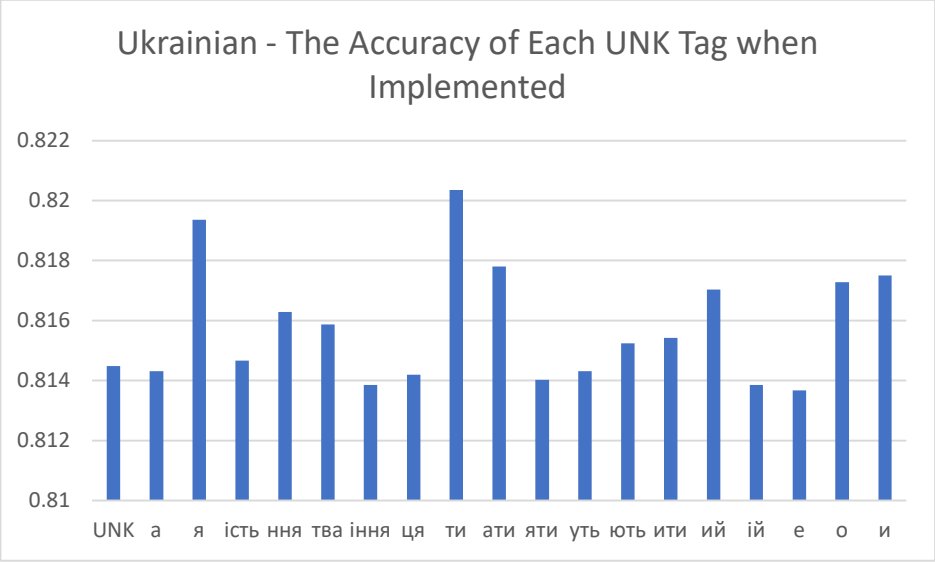
Appendix 3: Accuracies with UNK Tags



Appendix 4: A Bar Chart of each English Tag against their accuracy with UNK as the Base Line



Appendix 5: A Bar Chart of each French Tag against their accuracy with UNK as the Base Line



Appendix 6: A Bar Chart of each Ukrainian Tag against their accuracy with UNK as the Base Line