Lab 2: Problem Statement

Rudra Murthy V, Diptesh Kanojia
Course: Natural Language Processing (NLP)

January 27, 2022

In this week, we shall try to understand the implementation of a Part-of-Speech (PoS) tagger by utilising a rule-based approach. We have provided you with sample code which will help you understand the basics and take you a few steps ahead towards your own implementation.

You are provided with a Jupyter Notebook which consists of the following code snippets:

- Utility Function, *i.e.,* reading the data in CoNLL 2003 format.
- Data Split, *i.e.,* using the official train/test split provided by the Universal Dependency tree-bank.
- A Simple PoS Tagger Implementation, *i.e.,* tagging tokens with their most frequent PoS Tags.
- Evaluation & Analysis, *i.e.,* Evaluating the output of the most frequent PoS tagging approach and visualising a confusion matrix.

Problem 1. Should you choose to accept this 'mission', your 'mission' is to create your own rule-based PoS tagger to complete it. A rule-based PoS tagger takes specific rules into account which can be based either on linguistic intuition, or based on observations from the data. The discussion from our lectures should help you construct rules for your implementation. **You are encouraged to download the dataset in your native language for this task.**

Please see the *markdown* cell which states "Implement Rule-based System here" and start with your own code below it. You are already provided with an example rule which helps you tag the **adverbs** in your data. However, this rule tags any words ending *ly* as adverb which is a bit problematic, for example *family*. You should note that words such as '*lovely*' shall also be tagged as an **adverb** by such a tagger. Please be careful and construct rules which improve the performance of your tagger over the baseline most-frequent PoS tagger implemented already.

You need to come up with rules for other PoS tags, as many as you can create the rules for. However, it is mandatory to create rules for at least four tags. You can choose any four but you should know that creating lesser rules will impact your task performance.

Problem 2. You are also requested to create a PDF which explains the rules you have created and the motivation behind them. This PDF should contain your explanation for each rule but not the code. It should be noted that only copy-pasting the code from the notebook in this PDF will result in an award of zero marks for Problem 2. Your explanations, unlike neural network output, are as important as the implementation.

Submission You are requested to submit two files. The first file is your notebook containing the implementation for the problem described in Problem 1. The second file is the PDF containing your explanation of the rules you created for PoS tagging.

These submissions can be made at the Google Classroom portal akin to the first lab.