Natural Language Processing Assignment-II (Task-2)

Natural Language Processing Assignment-2 report submitted to Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

by

Shatansh Patnaik

(20MA20067)

Under the supervision of

Dr. Saptarshi Ghosh

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Spring Semester, 2024-25

February 4, 2024

Contents

Chapter 1

Implementation of Task-2

1.1 Evaluations for Test Data

1.1.1 UAS Score for the Sample Data

Based on the results obtained: The UAS (Unlabeled Attachment Score) over the sample data is 25.252525252525253%.

The Unlabeled Attachment Score is a metric used to evaluate the performance of dependency parsers. It measures the percentage of words in the predicted dependency tree that have the correct head. A higher UAS score indicates better performance in capturing syntactic relationships between words.

1.1.2 Assumptions

- First of all we are assuming that the feature model which we have created using the vocabulary, POS Tags and Modifiers over the training dataset also covers the test dataset. If there is a new word or POS Tag or modifier that is seen for the first time in the training dataset then we simply represent it as a vector of all zeroes which is required bag of words vector.
- We have cleaned out the data entries with index of the form x-y where x and y are both integers. We have assumed that these labels have no impact towards the overall training and processing of data.

1.1.3 Challenges Faced and steps taken to tackle them

1.1.3.1 Taking care of illegal moves

Since we are using the feature model that we have generated over the training data to test the test data, we might have moves that are not valid for instance we might get a LEFT-ARC move even when the stack is empty.

Solution

To prevent such conditions we have used a bunch of if...else conditions to take care of illegal moves and return in legal moves depending on the configuration that is provided to the function.

1.1.3.2 Having no access to dependency relationships between tokens while testing the data

During training of the data we have access to the Gold Standard Relationship types for each arc. However, while testing we are only predicting the head-dep words without considering the relationship between them.

Solution

To take care of this, we are considering a simple majority classifier, from the training set, we collect the arcs and we find out for every pair of POS Tags, what is the most frequently occuring relationship for the given arcs. We are also using the same for getting the POS Tags of newly encountered words. We are using such estimation to take care of the situations where the bag of words is a zero vector.

1.1.3.3 Dealing with out of vocabulary words in the test set

When we are considering a new test data independent of the training data set it is quite common to encounter entirely new words in the process.

Solution

In such cases, we are simply using a zero vector as the bag of words vector. This assumption is one of the major reasons why we are getting a low UAS Score

1.1.3.4 Existence of noisy data in the training and testing dataset

When we dive in deeper through the training and the test data, we can easily observe token-id of some lines to be of the form 14-15 or 20.1

Solution

To take care of this, we perform some pre-processing and directly delete such data entries from the procured data and we are assuming that such deletion doesn't affect the overall processes of training and testing the data.