

**Assignment – 3**  
**Speech & NLP**  
**Name : Prerit Gupta Roll No: 14EC35001**

**Approach:**

The features in the 10<sup>th</sup> column were taken together with the 5<sup>th</sup> column and concatenated to form the 5<sup>th</sup> column of the .conllu file with Morphological features. The corresponding files “train\_sent\_m.conllu” & “test\_sent\_m.conllu” were stored.

On the other hand, the features in the 5<sup>th</sup> column of the .conllu file were replaced by ‘\_’ to form the .conllu file without Morphological features. The corresponding files “train\_sent\_nm.conllu” & “test\_sent\_nm.conllu” were stored.

The hyper-parameters of the classifier were tested with different sets of values. The parameter values which gave the least overfitting difference between training scores and testing scores, as well as higher testing accuracy were selected. Then Arc-Standard and Arc-Eager methods were run.

**Results:**

No. of training examples : 500

No. of valid projective examples : 477

No. of test examples : 100

When trained with Morphological Features:

(1) By Arc-Standard:

| Training Algorithm  | Training Scores (%) |        | Testing Scores (%) |        |
|---------------------|---------------------|--------|--------------------|--------|
|                     | LAS                 | UAS    | LAS                | UAS    |
| SVM                 | 98.486              | 93.322 | 91.308             | 83.296 |
| Logistic Regression | 95.529              | 89.052 | 86.697             | 76.720 |
| MLP                 | 98.586              | 93.538 | 86.621             | 77.022 |

(2) By Arc-Eager:

| Training Algorithm  | Training Scores (%) |        | Testing Scores (%) |        |
|---------------------|---------------------|--------|--------------------|--------|
|                     | LAS                 | UAS    | LAS                | UAS    |
| SVM                 | 98.558              | 93.480 | 90.854             | 82.540 |
| Logistic Regression | 96.106              | 90.004 | 90.249             | 80.272 |
| MLP                 | 98.687              | 93.581 | 85.034             | 75.888 |

When trained without Morphological Features:

(1) By Arc-Standard:

| Training Algorithm  | Training Scores (%) |        | Testing Scores (%) |        |
|---------------------|---------------------|--------|--------------------|--------|
|                     | LAS                 | UAS    | LAS                | UAS    |
| SVM                 | 97.937              | 92.629 | 84.732             | 76.266 |
| Logistic Regression | 91.086              | 81.898 | 79.289             | 68.178 |
| MLP                 | 98.183              | 92.889 | 79.063             | 68.707 |

(1) By Arc-Eager:

| Training Algorithm  | Training Scores (%) |        | Testing Scores (%) |        |
|---------------------|---------------------|--------|--------------------|--------|
|                     | LAS                 | UAS    | LAS                | UAS    |
| SVM                 | 98.096              | 92.788 | 87.075             | 77.324 |
| Logistic Regression | 92.471              | 83.874 | 84.354             | 72.789 |
| MLP                 | 98.413              | 93.091 | 83.296             | 71.353 |

## Analysis:

The SVM classifier has following optimized parameters:

- kernel = 'poly'
- degree = 2
- coef0 = 0
- gamma = 0.2
- C=0.5

The Logistic Regression classifier has following optimized parameters:

- C = 0.5
- solver = 'lbfgs'

The MLP classifier has the following optimized parameters:

- hidden\_layer\_sizes=(100,50,)
- learning\_rate = 'adaptive',
- max\_iter=1000

- The training was extremely slow for MLP and SVM, especially in the case where morphological features were included. Though, accuracy for both labelled and unlabelled, was comparatively better. Arc-Eager gave better accuracy compared to Arc-Standard transitions.
- The SVM is giving the best score followed by the MLP & then Logistic Regression.
- The large difference in training and testing scores in some cases indicate overfitting of the model. Training parameters need to be specifically optimized for such cases.

- A high overfitting occurs when models are trained with data without morphological features. Training scores are very high and testing scores are low. So, including morphological features give better result.

**Conclusion:**

The Arc Eager method should be used with SVM classifier with Morphological features to obtain the best result for Dependency Parsing.