[COM4513-6513] Lab 3: Named Entity Recognition with the Structured Perceptron

Registration Number : **180128033**

1 Introduction

We are implementing code to learn a named entity recogniser (NER) using the structured perceptron.

2 Function Phi1

Function `Phi_1` is function that takes sentence, count as the input and returns a dictionary with counts of the `cw_cl_counts` keys in the given sentence.Function `load_dataset_sents` gives us the training data.Function `extraction_of_feature` gives us word and tag count by taking training data as input.Function `Word_Tag_Separator` gives the list of tag, which we use in the training function for the comparison.Dictionary and list returned from these function are used in the training function to find weight using structured binary perceptron. Further these weights returned from the training function are used as input in testing function along with `cw_cl_counts` and `test_file` to find the f1score and top 10 features.

2.1 Top 10 Features

f1 score for the Phi1 is 76.08 percent.Yes, features make sense.For ex- In "Per", we are receiving all the names of a person(Peter,Blackburn).In "LOC", we are receiving name of location(BRUSSELS,LONDON).In "ORG", we are receiving name of organisation(THWARA, VERINSBANK).In "MISC", we are receiving miscellaneous things(Open,Canadian).In "O" , we are receiving other words such as (REOFFER,NOTES).So, we according to the f1 score and the output received, we say that the implementation of Phi1 makes sense.

```
Top 10 features of O are
 [('1996-08-22_O', 0), ('._O', 0), ('BORROWER_O', 0), ('LAST_O', 0), ('AA+_O', 0), ('REOFFER_O', 0), ('=_O', 0), ('NOTES_O', 0), ('S_O', 0), ('SHORT_O', 0)]
Top 10 features of PER are
 [('Peter_PER', 1), ('Blackburn_PER', 1), ('Colleen_PER', 1), ('Siegel_PER', 1), ('Hassan_PER', 1), ('Hafidh_PER', 1), ('Hilary_PER', 1), ('Gush_PER', 1), ('Steve_PER', 1),
  ('Stricker_PER', 1)]
Top 10 features of LOC are
 [('BRUSSELS_LOC', 1), ('LONDON_LOC', 1), ('BEIJING_LOC', 1), ('FRANKFURT_LOC', 1), ('ATHENS_LOC', 1), ('JERUSALEM_LOC', 1), ('TUNIS_LOC', 1), ('BAGHDAD_LOC', 1), ('MANAMA_LOC',
  1), ('DUBAI_LOC', 1)]
Top 10 features of ORG are
 [('BAYERISCHE_ORG', 1), ('VEREINSBANK_ORG', 1), ('S&P_ORG', 1), ('THAWRA_ORG', 1), ('AN-NAHAR_ORG', 1), ('AS-SAFIR_ORG', 1), ('AL-ANWAR_ORG', 1), ('AD-DIYAR_ORG', 1),
  ('NIDA'A_ORG', 1), ('AL-WATAN_ORG', 1)]
Top 10 features of MISC are
 [('CS_MISC', 2), ('Canadian_MISC', 1), ('Open_MISC', 1), ('Malaysian_MISC', 1), ('Major_MISC', 1), ('League_MISC', 1), ('Baseball_MISC', 1), ('AMERICAN_MISC', 1), ('LEAGUE_MISC',
  1), ('EASTERN_MISC', 1)]
```

3 Function for Combination of Phi1 and Phi2

Function `Phi_2` is function that takes sentence, count and merged dictionary as the input and returns a dictionary with word-tag and tag-tag count.Function `combined_dictionary` returns the merger of dictionary of combination of return from phi1 and phi2.Function `load_dataset_sents` gives us the training data.Function `extraction_of_feature_2` gives us tag and tag count by taking training data as input.Function `n_grams_generation` gives the pair of tag, which are used in the `Phi_2`. Dictionary and list returned from these function are used in the training function to find weight using structured binary perceptron. Further these weights returned from the training function are used as input in testing function along with `cw_cl_counts` from both phi1 and phi2 along with `test_file` to find the f1score and top 10 features.

3.1 Top 10 Features

f1 score for the Phi1 is 76.56 percent.Yes, features make sense.For ex- In "Per", we are receiving all the names of a person(Kocinski,Yoshikawa).In "LOC", we are receiving name of location(England,Russia).In "ORG", we are receiving name of organisation(Newsroom, Oakland).In "MISC", we are receiving miscellaneous things(German,Dutch).In "O" , we are receiving other words such as (Out,2).So, we according to the f1 score and the output received, we say that the implementation of Phi1 makes sense.

```
Top 10 features of O are
 [('out_O', 6.3), ('2_O', 6.1), ('AT_O', 6.0), (':_O', 5.8), ('1_O', 5.7), ('-_O', 5.7), ('66_O', 5.7), ('of_O', 5.6), ('3_O', 5.6), ('Attendance_O', 5.5)]
Top 10 features of PER are
 [('Kocinski_PER', 6.5), ('Yoshikawa_PER', 5.5), ('Slight_PER', 5.2), ('Koerts_PER', 5.1), ('Vialle_PER', 4.8), ('Fogarty_PER', 4.8), ('Corser_PER', 4.7), ('McEwen_PER', 4.6),
 ('PAULO_PER', 4.6), ('Paul_PER', 4.1)]
Top 10 features of LOC are
 [('England_LOC', 8.7), ('Russia_LOC', 6.6), ('Netherlands_LOC', 6.4), ('YORK_LOC', 5.9), ('Germany_LOC', 5.9), ('Finland_LOC', 5.6), ('Pakistan_LOC', 5.6), ('CITY_LOC', 5.5),
 ('NEW_LOC', 5.4), ('Belgium_LOC', 5.4)]
Top 10 features of ORG are
 [('Newsroom_ORG', 6.7), ('Oakland_ORG', 5.2), ('Norwich_ORG', 5.1), ('Hanwha_ORG', 5.0), ('LG_ORG', 4.9), ('HOUSTON_ORG', 4.8), ('Cincinnati_ORG', 4.8), ('Rennes_ORG', 4.8),
 ('OAKLAND_ORG', 4.7), ('St_ORG', 4.7)]
Top 10 features of MISC are
 [('German_MISC', 5.9), ('Dutch_MISC', 5.8), ('GMT_MISC', 5.6), ('C$_MISC', 5.6), ('French_MISC', 5.5), ('English_MISC', 5.4), ('Polish_MISC', 5.3), ('League_MISC', 5.0),
 ('South_MISC', 5.0), ('Scottish_MISC', 4.6)]
```

4 Discussion

No, accuracy decreased in the combination of Phi1 and Phi2 (68.38) as compared to the Phi1 due to more number of features available.

Score	**Tabular**
f1 score for Phi1	76.08
f1 score for Phi1 + Phi2	68.38