NLP Assignment 3 : Named Entity Recognition with the Structured Perceptron

180128022

March 2019

1 Introduction

A Named Entity Recogniser (NER) with Structured Perceptron has been implemented. It is the process of labeling named-entities in the text. Named entities are real-world objects such as persons, locations, organisations etc, that can be denoted by a proper name.

2 Implementation

The following functions have been implemented keeping **5** iterations in mind to train the data :
1. **load_dataset_sents** : This is used to obtain word and tag sequences for each sentence.
2. **merge_dictionaries** : This is used to merge dictionaries.
3. **ngrams_generation** : This is used to generate n-grams.
List of ϕ_1 functions -
4. **word_label_phi_1** : This is used to return the counts of current word-current label.
5. **sentence_label** : This is used to break the training data into lists of sentences and labels.
6. **phi_1_func** : This is used to return the dictionary with counts of 'cw_cl_counts' keys in the given sentence.
7. **train** : This is used to train and return the weights.
8. **predict** : This is used return a predicted tag sequence for.
9. **test** : This is used to get the f1 measure.
10. **top_10** : This is used to get the top 10 for each tag.
Similar functions have been used for ϕ_2, namely -
word_label_phi_2, **phi_2_func**, **train_phi_2**, **predict_phi_2**, **test_phi_2** and **top_10_phi_2**.

3 Answers to the Questions

- F1 score Table:

Seed Value	Φ_1	$\Phi\_1 + \Phi\_2$
180128022	75.71%	73.83%

- These values make sense as the most of the named-entities are labelled correctly with an average accuracy of close to 75% for both the features.

- Yes, the differences among the feature sets in micro-F1 score are expected due to the difference between current word-current label and previous label-current label.
  No, taking Bigram into account didn't improve the accuracy.
  The accuracy is not increasing because ϕ_2 has more information about the feature sets and because of this high dimension, it is sparse, and as a result, the accuracy gets a little lower.

Tag_Value	1	2	3	4	5	6	7	8	9	10
O	1996-08-22_O	._O	BORROWER_O	LAST_O	AA+_O	REOFFER_O	=_O	NOTES_O	S_O	SHORT_O
PER	Peter_PER	Colleen_PER	Siegel_PER	Hassan_PER	Hafidh_PER	Hilary_PER	Gush_PER	Steve_PER	Stricker_PER	O'Meara_PER
LOC	BRUSSELS_LOC	LONDON_LOC	BEIJING_LOC	FRANKFURT_LOC	ATHENS_LOC	TUNIS_LOC	BAGHDAD_LOC	MANAMA_LOC	DUBAI_LOC	IRAQ_LOC
ORG	BAYERISCHE_ORG	VEREINSBANK_ORG	S&P_ORG	THAWRA_ORG	AN-NAHAR_ORG	AS-SAFIR_ORG	AL-ANWAR_ORG	AD-DIYAR_ORG	NIDA'A_ORG	AL-WATAN_ORG
MISC	C$_MISC	Canadian_MISC	Open_MISC	Malaysian_MISC	League_MISC	Baseball_MISC	AMERICAN_MISC	LEAGUE_MISC	EASTERN_MISC	DIVISION_MISC

Figure 1: Top 10 Features for Feature Set, Φ_1

Tag_Value	1	2	3	4	5	6	7	8	9	10
O	,_O	from_O	AT_O	out_O	0-0_O	Friday_O	1-0_O	Sunday_O	:_O	)_O
PER	Slight_PER	Kocinski_PER	Jim_PER	Corser_PER	Armstrong_PER	McEwen_PER	Fogarty_PER	Paul_PER	R._PER	Capiot_PER
LOC	England_LOC	Japan_LOC	YORK_LOC	Finland_LOC	Pakistan_LOC	Spain_LOC	Calif._LOC	PARIS_LOC	Russia_LOC	BONN_LOC
ORG	Newsroom_ORG	Cincinnati_ORG	Western_ORG	Atletico_ORG	Haitai_ORG	St_ORG	PITTSBURGH_ORG	BALTIMORE_ORG	CALIFORNIA_ORG	Milwaukee_ORG
MISC	Dutch_MISC	English_MISC	Scottish_MISC	German_MISC	C$_MISC	League_MISC	European_MISC	Yugoslav_MISC	French_MISC	LEAGUE_MISC

Figure 2: Top 10 Features for Feature Set, Φ_1 + Φ_2