Named entity recognition with the structured perceptron report

Ziyang Li

1. Feature choice

In this lab, selected features are current word with current label, current word with bigram labels and bigram words with trigram labels respectively.

2. Top 10 most positively-weighted features of each feature

In the process, the iteration runs 5 times on each feather. Furthermore, the value of seed in random function can affect accuracies as lab4.

top 10 for each class on current word with current label

O		ORG		LOC		PER		MISC	
‘,’	0.0	'Sydney'	1.0	'HAARLEM'	1.0	'N.'	1.0	'C$'	2.0
'Results'	0.0	'Newsroom'	1.0	'Netherlands'	1.0	'Eichmann'	1.0	'German'	1.0
'of'	0.0	Santos'	1.0	'DUBAI'	1.0	'Ruch'	1.0	'Canadian'	1.0
'first'	0.0	'Morelia'	1.0	'MOSCOW'	1.0	'Davis'	1.0	'Open'	1.0
'division'	0.0	‘Warwickshire’	1.0	'Namibia'	1.0	'Inzamam-ul-Haq'	1.0	'National'	1.0
'--'	0.0	'Casablanca'	1.0	Col'	1.0	'Anton'	1.0	'League'	1.0
'61-2'	0.0	'Tetouan'	1.0	'BRUSSELS'	1.0	'Ferreira'	1.0	'Major'	1.0
'race'	0.0	'Raja'	1.0	'LONDON'	1.0	'Adrian'	1.0	'Baseball'	1.0
'2'	0.0	'Newcastle'	1.0	'BAGHDAD'	1.0	'Warner'	1.0	'Dutch'	1.0
'236'	0.0	'Sheffield'	1.0	'ROME'	1.0	'REUTER'	1.0	'Slovak'	1.0

Micro-F1 score of current word with current label is 0.770341207349.

top 10 for each class on current word with bigram labels

O		ORG		LOC		PER		MISC	
'1996-08-21' LOC_O	3.0	'Commodities' ORG_ORG	2.8	'ANGELES' LOC_LOC	2.8	'Teutenberg' O_PER	2.6	'League' MISC_MISC	2.0
'4' ORG_O	2.8	'Newsroom' ORG_ORG	2.0	'Botswana' O_LOC	2.6	'PAULO' PER_PER	1.8	'Korean', MISC_MISC	2.0
'1996-08-23' LOC_O	2.8	'Newsdesk' ORG_ORG	2.0	'Uganda' O_LOC	2.4	'Armstrong' O_PER	1.8	'C$' O_MISC_	2.0
'through' PER_O	2.8	'FRANCISCO' ORG_ORG	2.0	'Mhow' None_LOC	2.4	'Wessels' PER_PER	1.8	'German' O_MISC	1.0
'1996-08-22' LOC_O	2.8	'LOUIS' ORG_ORG	2.0	'Ujjain' None_LOC	2.2	'Jayasuriya' PER_PER	1.8	'Canadian' O_MISC	1.0
'bln' O_O	2.8	'York' ORG_ORG	2.0	'Namibia' None_LOC	2.0	'Lombardi' O_PER	1.8	'Open' MISC_MISC	1.0
'1996-08-27' LOC_O	2.8	'NATIONS' ORG_ORG	2.0	'DELHI' LOC_LOC	2.0	'Lietti' O_PER	1.6	'Baseball' MISC_MISC	1.0

'2' ORG_O	2.6	'YORK' ORG_ORG	1.8	'Africa' LOC_LOC	2.0	'Croft' O_PER	1.6	'McLaren' None_MISC	1.0
'1996-08-25' LOC_O	2.6	'Vienna' ORG_ORG		'CITY' LOC_LOC	2.0	'Lotte' None_PER	1.6	'FI' MISC_MISC	1.0
'3' ORG_O	2.6	'HOUSTON' None_ORG	1.8	'Congo' O_LOC	2.0	'Capiot' O_PER	1.4	'GTR' MISC_MISC	1.0

Micro-F1 score of current word with bigram labels is 0.817065287654.

top 10 for each class on bigram words with trigram labels

O		ORG		LOC	
'DIEGO'_'1996-08-26' LOC_LOC_O	2.4	'6'_'Chicago' ORG_O_ORG	2.0	'NEW'_'DELHI' None_LOC_LOC	2.0
'ANGELES'_'1996-08-22' LOC_LOC_O	2.4	'JERUSALEM'_'POST' None_ORG_ORG	1.8	AT'_'LOS' ORG_O_LOC	2.0
'CHICAGO'_'63' None_ORG_O	2.0	'2'_'New' ORG_O_ORG	1.8	'AT'_'CHICAGO' ORG_O_LOC	2.0
'TORONTO', '61' None_ORG_O	2.0	'8'_'Philadelphia' ORG_O_ORG	1.8	'AT'_'ATLANTA' ORG_O_LOC	2.0
'CHICAGO', '1996-08-29' None_LOC_O	2.0	'3'_'SAN' ORG_O_ORG	1.8	'AT'_'NEW' ORG_O_LOC	1.8
'TORONTO'_'1996-08-21' None_LOC_O	2.0	'5'_'Minnesota' ORG_O_ORG	1.8	'AT'_'SAN' ORG_O_LOC	1.8
'BALTIMORE'_'67' None_ORG_O	2.0	'6'_'NEW' ORG_O_ORG	1.6	'AT'_'DETROIT' ORG_O_LOC	1.8
'Chicago'_'newsdesk' O_LOC_O	2.0	'9'_'Texas' ORG_O_ORG	1.4	'AT'_'TORONTO' ORG_O_LOC	1.6
'MONTREAL'_'70' None_ORG_O	2.0	'SAN'_'DIEGO' None_ORG_ORG	1.4	'AT'_'MONTREAL' ORG_O_LOC	1.6
'BALTIMORE'_'1996-08-22' None_LOC_O	2.0	'1'_'Philadelphia' ORG_O_ORG	1.2	'NEW'_'YORK' O_LOC_LOC	1.2

PER		MISC	
'None'_'N.' None_None_PER	1.0	'National'_'League' None_ MISC_MISC	2.0
'N.'_'Williams' None_PER_PER	1.0	'of'_'German' O_O_MISC	1.0
'2.'_'Eichmann' None_O_PER	1.0	'the'_'Canadian' O_O_MISC	1.0
','_'Ruch' PER_O_PER	1.0	Canadian'_'Open' O_MISC_MISC	1.0
'None'_'Davis' None_None_PER	1.0	'Major'_'League' None_MISC_MISC	1.0
'None'_'Inzamam-ul-Haq' None_None_PER	1.0	'League'_'Baseball' MISC_MISC_MISC	1.0
'None'_'Anton' None_None_PER	1.0	'of'_'Slovak' O_O_MISC	1.0
'Anton'_'Ferreira' None_PER_PER	1.0	'None'_'McLaren' None_None_MISC	1.0
'None'_'Adrian' None_None_PER	1.0	'McLaren'_'FI' None_MISC_MISC	1.0
'Adrian'_'Warner' None_PER_PER	1.0	'FI'_'GTR' MISC_MISC_MISC'	1.0

Micro-F1 score of bigram words with trigram labels is 0.736383442266.

3. Discussion

There is a difference between the micro-F1 score of bigram words with trigram labels and my expectation. Version with more features should have better performance like that current word with bigram labels is better than current word with its label. However, the last version does not improve the result and the result seems worst. The reason may be the training file we used. The training file have 1 word or 2 words on some lines. Another reason is bigram words feature which may need more data to be better. I tried to use the feature bigram words but it does not play well too and some features on the test file was not appeared before which affects the performance. For more features, I think the length of words and frequency of words could be helpful in Named entity recognition.