ASSIGNMNET 3 NATURAL LANGUAGE PROCESSING ASSIGNMENT

Named Entity Recognition System - Programming

Xuhong Guo 450489321

Abstract. In this assignment, we will build a named entity recognition (NER) system – a structured classifier for predicting the best sequence of named entity tags for an input sentence. The best performance in this assignment, we get accuracy: 97.11%; precision: 80.20%; recall: 82.13%; FB1: 81.18%

Contents

1 Introduce

We use the CoNLL-2003 shared task data files to achieve the our NER system, there are 16 shared task performance result, FIJZ03 have the highest performance, precision: 88.99%, recall: 88.54%,F: 88.76+-0.7. ML03 have the medium performance, precision: 84.52%, recall 83.55%, and F 84.04 +- 0.9. The last performance in this list is Ham03 whose precision:69.09%, recall: 53.26% and F: 60.15+-1.3%. While the baseline is precision is 71.91%,recall is 50.90% and f is 59.61+-1.2%. We get accuracy: 97.11%; precision: 80.20%; recall: 82.13%; FB1: 81.18%, this performance is higher than the baseline performance [1].

2 Method

In this assignment, we use the structured perceptron and viterbi algorithm to training and testing the dataset. The time is 1 epoch 1 minus, while the performance is stable after epoch larger than 10. Table 1 showed the summary of epoch.

Epoch vs Time					
Epoch	1	3	5	10	20
Time	1 minus	3 minus	5 minus	10 minus	20 minus

Epoch vs Accuracy					
Epoch	1	3	5	10	20
Accuracy	95.16%	96.59%	97.06%	97.11%	97.21%

Table 1. Summary of epoch

3 Features

In this assignment, we uses three different features to achieve our model performance. First feature, the attributes is i.e. Table 2, we use the words as feature, but the average performance is around 60%. Feature 1, it is important, because this feature have a basic information about the words. We also use this feature 1 performance as a baseline of our assignment.

Condition	Contextual predicate (Ci)
$\forall w_i$	$KLASS_{i-1} = X$

Table 2. Feature 1

Then for improve the performance, we use the second feature, the attributes is i.e. Table 3, this attributes is include the previous word, previous previous word, current word, next word, next next word and also with the previous POS tag, previous previous POS tag, current POS tag, next POS tag, next next POS tag. This average performance is around 80%. For improve the model performance we choose the second feature, because a word always show up with another words in a structure and the POS tag also in a structure, so we collected the words structure and POS tag structure as a feature, it capture a words and POS tag clusters information.

Last for improve the performance, we use the third feature, the attributes is i.e. Table 4, this attributes is include

Condition	Contextual predicate (Ci)
$\forall w_i$	$W_{i-2} = X$, $W_{i-1} = X$, $W_i = X$, $W_{i+1} = X$, $W_{i+2} = X$
$\forall w_i$	$POS_{i-2} = X$, $POS_{i-1} = X$, $POS_i = X$, $POS_{i+1} = X$, $POS_{i+2} = X$

Table 3. Feature 2

the attributes which Table 3 have, and suffix, prefix of words. Feature 3, capture more information, which is the structure of word itself. English word is base on the suffix prefix root. When we were training the data, it always have some infrequent words, but the suffix and prefix of infrequent words can have a relationship with the frequent words, so since Feature 3 captured the suffix and prefix of words information, the model will not be affected too much by the infrequent words.

Condition	Contextual predicate (Ci)
$\forall w_i$	$W_{i-2} = X$, $W_{i-1} = X$, $W_i = X$, $W_{i+1} = X$, $W_{i+2} = X$ $suffix(uptolength4) = X$ $prefix(uptolength4) = X$
$\forall w_i$	$POS_{i-2} = X$, $POS_{i-1} = X$, $POS_i = X$, $POS_{i+1} = X$, $POS_{i+2} = X$

Table 4. Feature 3

The three features we used in this assignment, feature 1 and feature 2 is the typical feature, while the feature 3 is a new feature but it is referenced by our tutorial 8. We can attribute it to this literature.

4 Experiments & Results

4.1 Features comparison

Table 5 is summary of three different features performances, which is based on the provided code with the conlleval perl script for scoring our predicted tag sequences against the gold standard.

From the Table 5, we can see the feature 3 get a highest performance. This result showed the feature 3 can get better performance than other features. Feature 3 captured sentence structure information and POS tag structure information and also the structure of word itself, it has more information than other features, so this result represent that more useful information captured in feature, higher performance the model can get.

Table 6 is the summary of three features subsets performance. In precision, we can find the PER label performance is almost same between three features, while the ORG label performance is more different between three features. In recall, PER label performance is not same, and

Feature 1	PRECISION	RECALL	FB1
LOC	74.23%	73.71%	73.97%
MISC	62.56%	56.72%	59.50%
ORG	41.05%	44.15%	42.54%
PER	82.26%	60.15%	69.49%
Overall	65.65%	60.20%	62.80%
Accuray	92.92%		

Feature 2	PRECISION	RECALL	FB1
LOC	84.35%	84.63%	84.49%
MISC	72.61%	76.42%	74.47%
ORG	70.35%	70.13%	70.24%
PER	84.97%	87.10%	86.02%
Overall	80.40%	81.66%	81.03%
Accuray	96.67%		

Feature 3	PRECISION	RECALL	FB1
LOC	86.24%	86.28%	86.26%
MISC	75.79%	80.48%	78.06%
ORG	72.22%	63.39%	67.51%
PER	81.60%	92.45%	86.69%
Overall	80.25%	81.66%	82.13%
Accuray	97.11%		

Table 5. Different features

	PRECISION			RECALL			FB1		
	Feature 1	Feature 2	Feature 3	Feature 1	Feature 2	Feature 3	Feature 1	Feature 2	Feature 3
LOC	74.23%	84.35%	86.24%	73.71%	84.35%	86.28%	73.97%	84.49%	86.26%
MISC	62.56%	72.61%	75.97%	56.72%	72.61%	80.48%	59.50%	74.47%	78.06%
ORG	41.05%	70.35%	72.22%	44.15%	70.35%	63.39%	42.54%	70.24%	67.51%
PER	82.26%	84.97%	81.60%	60.15%	84.97%	92.45%	69.49%	86.02%	86.69%

Table 6. Features subsets

the ORG label performance is still more different among tree features. So as the result, we can find the label LOC is not too much rely on the feature changing, so that means, the LOC feature is more significant even just with feature 1. But the other labels can not be identify well by the information supported by feature 1.

4.2 Viterbi modelling comparison

Table 7 is the performance of non-viterbi and viterbi. We use two same process code, one is with viterbi, other is without viterbi, From the result, we can find the overall performance is different, the viterbi algorithm improved the mode performance. In the precision we can find each label performance is be improved by viterbi, while in the recall performance, the label performance is almost same.

Non-Viterbi	PRECISION	RECALL	FB1
LOC	79.89%	85.63%	82.66%
MISC	65.22%	76.46%	70.39%
ORG	61.21%	71.89%	66.12%
PER	77.40%	82.36%	79.80%
Overall	72.05%	80.09%	75.86%

Viterbi	PRECISION	RECALL	FB1
LOC	86.24%	86.28%	86.26%
MISC	75.79%	80.48%	78.06%
ORG	72.22%	63.39%	67.51%
PER	81.60%	92.45%	86.69%
Overall	80.40%	81.66%	81.03%

Table 7. Non-verterbi vs verterbi

German

Feature 1	PRECISION	RECALL	FB1
LOC	56.97%	40.81%	47.56
MISC	45.67%	20.89%	28.67
ORG	31.67%	31.18%	31.43
PER	73.43%	29.98%	42.57
Overall	48.36%	31.04%	37.81
accuracy	90.78%		

German

Feature 2	PRECISION	RECALL	FB1
LOC	53.47%	48.26%	50.73
MISC	64.81%	18.42%	28.68
ORG	50.98%	31.43%	38.88
PER	59.90%	41.04%	48.71
Overall	55.91%	35.61%	43.51
accuracy	92.14%		

German

Feature 3	PRECISION	RECALL	FB1
LOC	52.07%	58.68%	55.18
MISC	63.28%	32.08%	42.58
ORG	49.05%	31.18%	38.13
PER	60.55%	52.03%	55.97
Overall	55.60%	44.13%	49.21
accuracy	93.05%		

Table 8. German

4.3 Baseline comparison

In this assignment, we set the feature 1 as the baseline. From the Table 5, we can find the feature 2, and feature 3, both features improved the model which base on the baseline. But feature 3 is more better.

4.4 Shared task performance comparison

In the shared task, base on our performance precision: 80.40%; recall: 81.66%; FB1: 81.03%, it is better baseline, and have a better performance than "HV03","DD03","Ham03" [1].

4.5 German file comparison

The encoding is different between the English and German. So when we use the German file , we need change the encode to 'cp1252'. When we training the German file, the training time is same with training English file. But the performance is different. As same way, the performance of German result is more poor. Table 8 showed the result of performance.
In the Table 8, we can see the feature 3 performance is still higher than other features. From the result, all three features performance decreased, but the different of feature 1 between English and German is less than other features. So as the result the feature 2, and feature 3 are language-independent. From the performance of result, those two features decreased faster, while from structure of words or word, maybe it is different way with English word.

5 Error Analysis

5.1 Typical tag confusions

The typical tag confusions our system made is the ORG tag, and MISC tag. In the three features performance result, the two tag alway have a lower performance than other tag, we can find in Table 5.

5.2 Characteristic errors

When we deal with the words which didn't show on the training, we made it as 0, this is a error, because we can not sure what weight this word should be. And when we training the attribute of previous tag and current tag, we reduced the number. Because when we use the viterbi algorithm, large number weight of previous tag and current tag will affect the weight of current feature. But the reduced number of previous tag and current tag, we set it as 0.01, it is from experience, so it will make a error.

5.3 Dataset affection

In the dataset, the punctuation and the number of O tag data affect our system's performance. Because number of 0 tag are more than the other tag, it made the dataset imbalanced.

6 Conclusion

In this assignment, we find the features with more useful information, the performance will be better, and the viterbi algorithm can useful improved the tagging model. A method can not suitable to different language.

References

[1] http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/