Comp6490 Document Analysis IE Assignment

Guyue HU
u5608260

Q2

(a) The performance**:**

Use 1/4 of the training data (esp.train1):

esp.testa:

processed 46167 tokens with 3076 phrases; found: 2927 phrases; correct: 1923.
accuracy: 95.51%; precision: 65.70%; recall: 62.52%; FB1: 64.07
LOC: precision: 70.52%; recall: 64.79%; FB1: 67.54 882
MISC: precision: 50.00%; recall: 18.39%; FB1: 26.89 96
ORG: precision: 61.34%; recall: 70.43%; FB1: 65.57 1402
PER: precision: 71.85%; recall: 61.99%; FB1: 66.55 547

esp.testb:

processed 47696 tokens with 3877 phrases; found: 3597 phrases; correct: 2155.
accuracy: 93.39%; precision: 59.91%; recall: 55.58%; FB1: 57.67
LOC: precision: 54.58%; recall: 65.07%; FB1: 59.37 1048
MISC: precision: 36.88%; recall: 15.53%; FB1: 21.85 160
ORG: precision: 59.13%; recall: 63.09%; FB1: 61.05 1610
PER: precision: 73.43%; recall: 51.58%; FB1: 60.59 779

Use 2/4 of the training data (esp.train2)**:**

esp.testa:

processed 46167 tokens with 3076 phrases; found: 2924 phrases; correct: 2031.
accuracy: 96.02%; precision: 69.46%; recall: 66.03%; FB1: 67.70
LOC: precision: 71.44%; recall: 70.10%; FB1: 70.77 942
MISC: precision: 52.80%; recall: 25.29%; FB1: 34.20 125
ORG: precision: 67.76%; recall: 69.37%; FB1: 68.56 1250
PER: precision: 73.31%; recall: 70.19%; FB1: 71.72 607

esp.testb:

processed 47696 tokens with 3877 phrases; found: 3574 phrases; correct: 2276.
accuracy: 93.93%; precision: 63.68%; recall: 58.71%; FB1: 61.09

LOC: precision: 54.21%; recall: 70.99%; FB1: 61.48 1151
MISC: precision: 41.35%; recall: 22.63%; FB1: 29.25 208
ORG: precision: 67.80%; recall: 59.44%; FB1: 63.35 1323
PER: precision: 75.00%; recall: 60.32%; FB1: 66.87 892

Use 3/4 of the training data (esp.train3):

esp.testa:

processed 46167 tokens with 3076 phrases; found: 2927 phrases; correct: 2133.
accuracy: 96.43%; precision: 72.87%; recall: 69.34%; FB1: 71.06
LOC: precision: 74.49%; recall: 72.71%; FB1: 73.59 937
MISC: precision: 60.40%; recall: 34.48%; FB1: 43.90 149
ORG: precision: 71.61%; recall: 72.32%; FB1: 71.96 1233
PER: precision: 75.99%; recall: 72.87%; FB1: 74.40 608

esp.testb:

processed 47696 tokens with 3877 phrases; found: 3592 phrases; correct: 2348.
accuracy: 94.34%; precision: 65.37%; recall: 60.56%; FB1: 62.87
LOC: precision: 54.69%; recall: 73.61%; FB1: 62.75 1183
MISC: precision: 45.74%; recall: 26.84%; FB1: 33.83 223
ORG: precision: 70.64%; recall: 61.23%; FB1: 65.60 1308
PER: precision: 76.88%; recall: 60.87%; FB1: 67.94 878

Use full sized training data (esp.train):

esp.testa

processed 46167 tokens with 3076 phrases; found: 2943 phrases; correct: 2292.
accuracy: 97.10%; precision: 77.88%; recall: 74.51%; FB1: 76.16
LOC: precision: 80.61%; recall: 74.48%; FB1: 77.42 887
MISC: precision: 68.52%; recall: 42.53%; FB1: 52.48 162
ORG: precision: 76.61%; recall: 79.69%; FB1: 78.12 1270
PER: precision: 79.01%; recall: 77.76%; FB1: 78.38 624

esp.testb

processed 47696 tokens with 3877 phrases; found: 3641 phrases; correct: 2650.
accuracy: 95.47%; precision: 72.78%; recall: 68.35%; FB1: 70.50
LOC: precision: 64.75%; recall: 76.91%; FB1: 70.31 1044
MISC: precision: 55.33%; recall: 35.53%; FB1: 43.27 244
ORG: precision: 74.91%; recall: 71.44%; FB1: 73.13 1439
PER: precision: 83.26%; recall: 68.62%; FB1: 75.23 914

(b) Conclusion:

As the size of training data increases, the general accuracy, precision, recall and FB1 score all increases. The improvement of accuracy is less obvious, and the improvement of recall is more obvious.
But generally speaking, because the original size of the training data is relatively large, the improvement of FB1 score from ¼ size training data and full size training data is not very large. (67.7 -> 77.4 for testa, 57.7->70.5 for testb). Which means if the training data is not very closed correlated, a relatively small size of training data can result in a reasonably good performance.

Q3

(a) Baseline 1: Simple Heuristics

Use simple features to recognize named entities:

PERSON
- Special title: Mr., Mrs., Miss., … ect.
- Initial capital: Jack, Mary, Green, … ect.

ORGANIZATION
- All capital

LOCATION
- Special preposition: in, at, around, ... ect.
- Initial capital: Paris, Beijing,... ect

pseudo-code:

```
Heuristic_NER(current_token):
    if All_Capital(current_token):
        return ORGANIZATION
    if (current_token.previous() in {Mr., Mrs., Miss, ... }) and
    Initial_Capital(current_token):
        return PERSON
    if (current_token.previous() in {in, at, around, ...}) and
    Initial_Capital(current_token):
        return LOCATION
```

(a) Baseline 2: Naive Bayes

Tag lexical category of the token, and apply Naive Bayes on the following features[1]:

A. Tokens that are turned into all upper-case, in a window of ± 2

B. Tokens themselves, in a window of ± 2

C. The previous two predicated tags, and the conjunction of the previous tag and the current token

D. Initial capitalization of tokens in a window of ± 2

E. More elaborated word type information: initial capitalization, all capitalization, all digitals, or digital containing punctuations

pseudo-code:

```
Naive_Bayes_training(train_set):
        for every class C in {PERSON, ORGANIZATION, LOCATION}:
                for every feature F in {A,B,C,D,E}:
                        p(F|C) = (number of training data in class C and contains F) /
                (number of all training data in C)
        for every class C in {PERSON, ORGANIZATION, LOCATION}:
                p(C) = (number of training data in class C) / (number of all training data)
        return all p(F|C), p(C)

Naive_Bayes_predicting(token):
        for every class c in {PERSON, ORGANIZATION, LOCATION}:
                unnormalized p'(c) = p(A|c)*p(B|c)*p(C|c)*p(D|c)*p(E|c)*p(c)
        return c with the largest p'(c)
```

Q3

Hidden States: the field we want to extract: (i) title, (ii) authors, (iii) publication date, (iv) others
Observation: different features of the text in scientific papers. For example: (i) all capital, (ii) initial capital, (iii) numbers, dashes,... (iv) place in the text. ect.

References

[1] Tong Zhang, David Johnson. A Robust Risk Minimization based Named Entity Recognition System