Name Entity Tagger using Maxent Classifier

Xihua Yang
Courant Institute of Mathematical Sciences
251 Mercer St, New York, NY 10012
xy644@nyu.edu

Abstract

Present a name entity tagger built with maximum entropy classifier. Different combinations of features are tested and compared by performance on CoNLL-2002[?] shared task corpus, which includes a Spanish dataset and a Dutch dataset. The classifier obtained 73.60% F_1-*measure on Spanish dataset and* 69.26% F_1-*measure on Dutch dataset.*

1. Introduction

1.1. Objective

Build a name entity tagger with maximum entropy classifier. The name tagger should be able to recognize the category of each word in CoNLL-2002 dataset.

1.2. External softwares and web resources

- NLTK 3.0.2[1]
- Megam[2]

1.3. System feature

- Language: Python 2.6
- Dataset: CoNLL-2002 shared task

2. Experiment

2.1. Dataset

The dataset comes from CoNLL-2002 shared task. It contains 4 kinds of name entities, which are Person, Location, Organization and Miscellany, they are labeled as "PER", "LOC", "ORG", and "MISC" respectively. The name entities are tagged using BIO tags in training data, "B" indicates the start of a new name entity, "I" indicates a word that is inside a name entity but not the begin word, and "O" means this word is outside any name entities. For example, the begining word of a Person type name entity will be tagged as "B-PER".

The dataset contains training and test data in Spanish and Dutch, Spanish training data has 264715 training samples, and Dutch training data has 202644 training samples.

During experiment, the performance of different feature sets are compared base on Spanish dataset. To avoid getting memory overflow and speed up training process, Megam package is used.

2.2. Previous Work

According to Tjong's paper[6], many techniques are adopted for CoNLL-2002 shared task, among which only Malouf[3] used maximum entropy classifier. In Malouf's paper, he reached F1-score of 73.66% on Spanish dataset and 68.08% on Dutch dataset.

2.3. Features

In previous homework assignment, I have found that using previous word, current word, next word, their POS tags and the combination of these features will result in a good performance, and previous BIO tag is also a good feature, therefore my initial feature set is:

```
previous word
current wor
next word
previous POS tag
current POS tag
next POS tag
previous word + current word
current word + next word
previous word + current word + next
    word
previous POS tag + current POS tag +
    next POS tag
previous BIO tag
```

F1-score is 59.93%

In Nadeau and Sekine's paper[5], they listed a bunch of possible features that could help improving performance of name entity taggers, including case, punctuation, digit, etc. Therefore in second iteration, I also detected whether the

word is a title word, whether the word contains a punctuation, and whether it contains a digit but is not a numerical word. Also, the length of word is added as a feature, now the feature set is:

```
previous word
current wor
next word
previous POS tag
current POS tag
next POS tag
previous word + current word
current word + next word
previous word + current word + next
    word
previous POS tag + current POS tag +
    next POS tag
previous BIO tag
istitle
punctuation
contain digit
word length
```

F1-score is 65.28%

Refering to Tkachenko and Simanovsky's paper[7], features like prefix and suffix are also useful. Since we don't know the common length of prefix and suffix in Spanish, I have to try different features. Length from 1 to 7 are tried for prefix, and 5 to 1 are tried for suffix. At last, it turned out that prefix with length 6, suffix with length 2 and 1 yielded best performance. The feature set is:

```
previous word
current wor
next word
previous POS tag
current POS tag
next POS tag
previous word + current word
current word + next word
previous word + current word + next
    word
previous POS tag + current POS tag +
    next POS tag
previous BIO tag
istitle
punctuation
contain digit
word length
prefix6
suffix2
suffix1
```

F1-score is 70.77%

Now we want to take more feature on the shape of word into consideration. In Nadeau and Sekine's paper, they mentioned that whether the word is upper case or mix case would also help. I did not take whether it is a lower case word as a feature, because that will mess up words like "Apple" and "apple". Now the feature set is:

```
previous word
current wor
next word
previous POS tag
current POS tag
next POS tag
previous word + current word
current word + next word
previous word + current word + next
    word
previous POS tag + current POS tag +
    next POS tag
previous BIO tag
istitle
punctuation
contain digit
word length
prefix6
suffix2
suffix1
is upper case
is mix case(not upper case and not
    lower case and not title)
```

F1-score is 72.61%

Also, we tried to stem the words and add them as features. Current word is more possible to be part of a name entity when previous word is name entity and current word is a noun, we also add a joint feature of current POS and previous BIO tag into feature set. For test set, previous prediction is taken as previous BIO tag. Besides, if adjacent words are title words, current word is likely to be a part of name entity if it is noun. Now the feature set is:

```
previous word
current wor
next word
previous POS tag
current POS tag
next POS tag
previous word + current word
current word + next word
previous word + current word + next
    word
previous POS tag + current POS tag +
    next POS tag
previous BIO tag
istitle
punctuation
contain digit
```

```
word length
prefix6
suffix2
suffix1
is upper case
is mix case(not upper case and not
    lower case and not title)
current POS + previous BIO
previous stem
current stem
next stem
previous istitle + next istitle +
    current pos
```

F1-score is 73.06%

If we also add whether previous word is title word and whether next word is title word, this could also help, because continuous title words could possibly form up a name entity. Now the feature set is:

```
previous word
current wor
next word
previous POS tag
current POS tag
next POS tag
previous word + current word
current word + next word
previous word + current word + next
    word
previous POS tag + current POS tag +
    next POS tag
previous BIO tag
istitle
punctuation
contain digit
word length
prefix6
suffix2
suffix1
is upper case
is mix case(not upper case and not
    lower case and not title)
current POS + previous BIO
previous stem
current stem
next stem
previous istitle + next istitle +
    current pos
```

F1-score is 73.60%

There are some other features tried but not resulting a good performance. I tried to add the upper case form of current word, lower case form of current word, the joint of previous POS tag and current POS tag, the joint of current POS tag and next POS tag. However, they yielded poorer performance. Their F1-scores are 71.12%, 71.49%, 68.74% and 70.58% respectively.

According to Mao's paper[4], we can try to adopt non-local features such as the position of word in sentence. I tried to add the position information into feature set, but this did not perform well on CoNLL-2002, the F1-score is 70.51%

References

[1] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[2] H. D. III. Notes on cg and lm-bfgs optimization of logistic regression. 2004. http://www.umiacs.umd.edu/ hal/docs/daume04cg-bfgs.pdf.

[3] R. Malouf. Markov models for language-independent named entity recognition. *Proceedings of CoNLL-2002*, pages 187–190, 2002.

[4] X. Mao, W. Xu, Y. Dong, S. He, and H. Wang. Using non-local features to improve named entity recognition recall. *The 21st Pacific Asia Conference on Language, Information and Computation : Proceedings*, 21, 2007.

[5] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[6] E. F. T. K. Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *Proceedings of CoNLL-2002*, pages 155–158, 2002.

[7] M. Tkachenko and A. Simanovsky. Named entity recognition: Exploring features. *Proceedings of KONVENS 2012*, 2012.