Identify Human Names in News Documents using Machine Learning Algorithms

Lan Bai, Chaoqun Mei, Yuzhe Ma

March 3, 2018

Abstract

We perform human name identification among a corpus of news articles using machine learning techniques. Based on the self-constructed features of the human names that appear in the BBC news corpus, we evaluate the performance of five learning algorithms with cross validation. The optimal algorithm achieves more than 90% on both the precision and the recall over the test articles, which shows that the machine learning is efficient in identifying human names in the text.

1 Data Source

The data comes from the BBC news corpus[1], which contains more than 2000 articles covering different aspects of human life, including business, entertainment, politics, technology and sport. To cut down the number of articles, we pick only the politics as our data source, which includes 320 articles.

2 Data Processing

The human names or positive examples in the 320 articles are marked up as ^Name^, where Name is the full name of some person, including his first name and the last name but not the prefix such as Mr. or Miss. We also include around 1000 non-person entity in the articles that are used as the negative examples. We let the first 220 articles be the training data **I**, the leftover being the test data **J**. In total we have 2039 mentions of Name. In the training data **I**, we have 1401 positive example and 618 negative examples while in the test data **J**, we have 638 positive examples and 293 negative examples. The constructed feature for each marked entity is a 7-dimensional vector, with each dimension being as follows:

(1). All the words in the entity is capitalized in the first letter, 1 being yes and 0 otherwise.

(2). The number of the words in the entity. For human names, this quantity is likely to be one or two.

(3). Whether there exists prefix in front of the entity, such as Mr. or Miss, 1 being yes and 0 otherwise. Note that this would be a strong indicator of whether the entity is a human name or not.

(4).Whether there exist some verbs that usually hand around human names, such as "said" and "feel". i being yes and 0 otherwise.

(5). Whether all letters are capital. 1 being yes and 0 otherwise. This feature could possibly distinguish human names from abbreviations of some other organizations or locations, etc. Usually, human names have value 0 on this feature while others would have 1, e.g. "BBC".

(6). Whether there exists "The" around the entity, since "The" rarely refers to a human.

(7). Whether there exist some word in the entity that appears also in a Black List, which is manually maintained by us.

[1] http://mlg.ucd.ie/datasets/bbc.html

3 Learning and Model Evaluation

We evaluate five models, including linear regression, logistic regression, support vector machine, decision tree and random forest. We perform 5-fold cross validation on all of them in order to pick the optimal algorithm. The precision, recall, and the F1 score of five models are shown in Table 1. Note that interestingly decision tree and random forest achieves almost the same result. Logistic regression and SVM also achieve the same value on the above three quantities.

Table 1: Measurement on training data

	Decision Tree	Random Forest	Linear Regression	Logistic Regression	SVM
Precision	0.910144	0.910144	0.968342	0.908938	0.908938
Recall	0.894617	0.893804	0.632078	0.892178	0.892178
F1 score	0.901481	0.901481	0.753617	0.899517	0.899517

The final classifier that we select is the decision tree. We then train on corpus **I** and test on corpus **J**, and compute the three quantities as above. On the test data, the precision, recall and F1 score are 0.905478, 0.942006, and 0.923381 respectively.