CS 839 Project Stage 1

Dongqiangzi Ye, Sek Cheong, Pan Wu

Entity type: Person Names, e.g. Mick Deats, Mick, Deats, Kerstin Dautenhahn (Our source database: BBC News article archive)

Our definition of name:

Full Name(contain both first name and last name, maybe contain middle name), only First Name, only Last Name

Total number of mentions marked up:	2799
The number of documents in set I:	247
The number of mentions in set I:	1747
The number of documents in set J:	123
The number of mentions in set J:	1052

Preprocessing:

1. We use 1-gram, 2-gram and 3-gram, our positive sample must start with Captal letter and only contain alphabets.
2. To balance negative and positive sample, we select only 30% negtive sample.
3. We extract words from all files based on stemming.

Features slection:

1. Use use the one-hot encoding for the word from a bag of words and we group the encode into bin of 5. The bag of words war created from the entire corpus of our documents. For example, the name "Peter Allen", the word "Peter" has index number of 6 in our bag of words and the word "Allen" has the index number of 11 in our bag of words. And let's say we have 1000 words in our bag of words. The vector for "Peter Allen" would be a 1000 dimensional

vector with the second and third element set to 1 and the rest of the elements set to 0, i.e. [0, 1, 1, ..., 0]

2. Is the word at the beginning of a sentence.
3. Is the word at the end of a sentence.
4. Is the word followed by a punctuation
5. Is the word followed by a pronoun such as 'who', 'whose', 'whom', etc
6. Is the word preceded by a title such as 'Mr', 'Mrs', 'Miss', 'Prof', 'Dr', etc
7. Is the word followed by a possessive form such as 's
8. Is the word preceded by a word that is likely followed by a name such as 'by', ‘include’, etc
9. Is the word followed by a past tense verb that is likely precede by a name such as 'said', 'told', etc

---

Cross validation:

Cross validation(num-bin=5)			
%	precision	recall	F1
Linear Regression	87.9	83	85.4
Logistic Regression	86.4	84.8	85.6
Decision Tree	88.2	81.9	85
Random Forest	88.7	80.7	84.5
SVM	0	0	0

Cross validation(num-bin=1)			
%	precision	recall	F1
Linear Regression	/	/	/
Logistic Regression	/	/	/
Decision Tree	/	/	/
Random Forest	90.6	78.7	84.2
SVM	/	/	/

Note: svm does not work with test data(bags of words)

classifier M(the type of the classifier that you selected after performing cross validation on set I):

Random forest

precision of M(on set I): 90.6%

recall of M(on set I): 78.7%

F1 of M(on set I): 84.2%

Test:

classifier X(the type of the classifier that you have finally settled on before the rule-based postprocessing step):

Random forest

precision of X(on set J): 93%

recall of X(on set J): 87%

F1 of X(on set J): 90%

---

rule-based post-processing:

We do not need rule-based post-processing, bacaue our precision exceeds 90% (we get 93%), and recall exceeds 60% (we get 87%).

what else can you possibly do to improve the accuracy:

1. add more training data
2. add more features
3. we can learn context