CS 638

Stage IV

Shixuan Fan: sfan33@wisc.edu
Zhenyu Zhang: zzhang546@wisc.edu
Yuting Liu: liu487@wisc.edu

Part A. Refinement of the blocked set

In last stage, we have a blocked table C of size 100k which is too big to sample enough positive data. Current method we used to block table is
(1) Edit distance of Phone Number less than 1;
(2) Edit distance of restaurant name less than 1;
(3) Two Address has the same value.
Take the union of (1)(2)(3), the size of the new blocked table C reduced to 908 tuples. We also filled missing value of Phone with 000-000-0000, filled the missing value of Price with 0.

Part B. Analysis

1. Obtain the Precision, recall, and F-1 of five learning methods after performing cross validation for the first time for these methods on the development set I

Learning Methods	Precision	Recall	F-1
Decision Tree	0.982658959538	0.960451977401	0.971428571429
Random Forests	0.988439306358	0.966101694915	0.977142857143
SVM	0.987804878049	0.915254237288	0.950146627566
Naïve Bayes	0.909090909091	0.960451977401	0.934065934066
Logistic Regression	0.988235294118	0.949152542373	0.968299711816

2. After cross validation, chose the learning-based matcher

As the table shown above, the method of random forests provides the highest precision, recall, and F-1 among all methods.

3. Report all debugging iterations and cross validation iterations.

For each debugging iteration

(a) what is the matcher that you are trying to debug, and its precision/recall/F-1
(b) what kind of problems you found, and what you did to fix them
(c) the final precision/recall/F-1 reached

For each cross-validation iteration

(a) matchers tried to evaluate using the cross validation
(b) precision/recall/F-1 of those

We didn't provide any debugging iterations and cross validation iterations because our performances of the result with different metrics are quite good enough.

4. The final best learning-based matcher selected, and its precision/recall/F-1.

The final best learning-based matcher is Random forest.

Precision	Recall	F1
0.988439306358	0.966101694915	0.977142857143

5. For each of five learning methods, train it on the evaluation set I, then report its precision/recall/F-1 on the development set J.

Learning Methods	Precision	Recall	F-1
Decision Tree	0.934065934066	1.0	0.965909090909
Random Forests	0.965909090909	1.0	0.982658959538
SVM	0.976470588235	0.976470588235	0.976470588235
Naïve Bayes	0.976744186047	0.988235294118	0.982456140351
Logistic Regression	0.976470588235	0.976470588235	0.976470588235

6. Use final best matcher Y* and train it on I. Report its precision/recall/F-1 on J.

We chose the method of random forests and it's result is shown as following:

Precision	Recall	F1
0.965909090909	1.0	0.982658959538

7. List the final set of features that you are using in your feature vectors.

Features including:

- Rate: Absolute difference of two rates as feature
- Price: Edit distance of two prices.
- Zip-code: Firstly calculate absolute difference value of Zip-code in table A and B. If zero, we use 0 as the feature value. If the value is less than 20, we use 1 as the feature value. For other cases, we use 5 as the feature value of zip-code.
- State: 0 if they are equal, else let the feature value be 1.
- City: 0 if they are equal, else let the feature value be 1.

- phone: Edit distance of two strings.
- Delivery: 0 if they are equal, else let the feature value be 1.
- Take-out: 0 if they are equal, else let the feature value be 1.
- Name: TF-IDF value.
- Address: TF-IDF value.

8. Discuss why you can't reach higher precision, recall, F-1.

To be honest, it took a long time for us to label the data. It turned out to have a very good result at the very first iteration of the cross validation, which saves us quite a long time for debugging and repeating processes of building the classifier.
Since our precision, recall, and F1 are closed to 100%. We all know that there are several reasons that limit the possibility of the perfect matcher. There may be some system errors. To be specific, there will be some restaurants that moved to some other places, or maybe they just closed due to the unsuccessful running. There may be some realistic problems leading to the result of the incorrect information, and it's hard for the websites to follow up with the real-time information of all restaurants. Different restaurants will have different speed to check the incorrect information and update the information, sometimes there may be some incorrect information existing on the website. We have to tolerate the system errors then.