# Summary

# Enrich Model Test on 2+3

Model: SVM

Feature: All features

## 3-way Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Train | Test | Precision | Recall | F-Measure |
| train2 | test1 | 0.833 | 0.842 | 0.835 |
|  | test2 | 0.835 | 0.806 | 0.817 |
|  | test3 | 0.716 | 0.71 | 0.713 |
|  | test4 | 0.841 | 0.82 | 0.828 |
|  |  |  |  |  |
| train3 | test1 | 0.789 | 0.776 | 0.782 |
|  | test2 | 0.797 | 0.744 | 0.764 |
|  | test3 | 0.87 | 0.873 | 0.868 |
|  | test4 | 0.868 | 0.751 | 0.782 |
|  |  |  |  |  |
| train23 | test1 | 0.837 | 0.847 | 0.838 |
|  | test2 | 0.832 | 0.813 | 0.821 |
|  | test3 | 0.865 | 0.869 | 0.864 |
|  | test4 | 0.88 | 0.808 | 0.828 |

In the summary10.24.2013, I have shown that enrich model is better than others (test1,2,4 with train2; test3 with train3). Now, train23 is even better.

## Combined Results (Highlight ones are better than “bestbyothers”)

“3way\_enrich\_train2\_allmetrics”, “3way\_enrich\_train3\_allmetrics”, “3way\_enrich\_train22\_allmetrics” are models trained 2, 3, 2+3 respectively (using distribution of labels).

“3way\_actngram\_allmetrics”, “3way\_enrich3\_allmetrics” used actngram and all features respectively (without distributional lables)



# New Models

## Voting Model

The idea is to use majority voting of different models. Here, I used Naïve Bayes, SVM, and Decision Tree.

## Self-Training Model

The idea of self-training is to train on the test set.

1. Train a model only on the train set
2. Use the model to get prediction on test set
3. Retrain the model on the train set and the test set with the predicted labels
4. Use the new model to predict the test set

### 3-wayResult

* Enrich3 is the best model we got
* Voting is even better than enrich3
* Self-training doesn’t help



## Combined results for voting

The new model is better than enrich3 without voting.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| method | schedule | test | accuracy | avgp | l2 | mrr |
| 3way\_enrich\_voting\_allmetrics | schedule1 | test1 | 0.928 | 0.9294 | 0.0998 | 0.941 |
| 3way\_enrich\_voting\_allmetrics | schedule1 | test2 | 0.9312 | 0.9324 | 0.0957 | 0.9465 |
| 3way\_enrich\_voting\_allmetrics | schedule1 | test3 | 0.9026 | 0.8925 | 0.1521 | 0.9132 |
| 3way\_enrich\_voting\_allmetrics | schedule1 | test4 | 0.8841 | 0.8836 | 0.1646 | 0.8853 |
| 3way\_enrich\_voting\_allmetrics | schedule2 | test1 | 0.8396 | 0.8435 | 0.2213 | 0.8692 |
| 3way\_enrich\_voting\_allmetrics | schedule2 | test2 | 0.7815 | 0.7852 | 0.3037 | 0.8284 |
| 3way\_enrich\_voting\_allmetrics | schedule2 | test3 | 0.6015 | 0.5675 | 0.6116 | 0.6503 |
| 3way\_enrich\_voting\_allmetrics | schedule2 | test4 | 0.7674 | 0.7663 | 0.3305 | 0.77 |
| 3way\_enrich\_voting\_allmetrics | schedule3 | test1 | 0.9087 | 0.9085 | 0.1293 | 0.9267 |
| 3way\_enrich\_voting\_allmetrics | schedule3 | test2 | 0.8862 | 0.8884 | 0.1578 | 0.9123 |
| 3way\_enrich\_voting\_allmetrics | schedule3 | test3 | 0.6599 | 0.6154 | 0.5438 | 0.7019 |
| 3way\_enrich\_voting\_allmetrics | schedule3 | test4 | 0.7325 | 0.7308 | 0.3807 | 0.7359 |

# New Data

## Differences between DSTC2 and DSTC1

|  |  |  |
| --- | --- | --- |
|  | DSTC2 | DSTC1 |
| domain | Bus | Restaurant |
| data sources | 1 group | 3 groups |
| slot values | Close (limited) | Open (Unlimited) |
| annotation | Only the Correct SLU | Whether each SLU is correct or not (Correct one might not be available) |
| evaluation metrics | Only 9 metrics | 1452 metrics by different schedule, slot, test set, metric |
| baseline | The best SLU among all SLU candidates | The best SLU among TOP 1 SLU candidate |
| Goal Changing | Allowed | Not allowed |

## Important Facts about DSTC2

### Combination of Dialog Manager and Speech Recognition

The 3 dialog managers (DM) are:

1. an MDP for tracking the dialog state, and a hand-crafted policy
2. a POMDP dialog state tracking, and a hand-crafted policy
3. a POMDP a policy learnt using reinforcement learning

The 2 speech recognizers (SR) are:

1. GMM-HMM model with artificially degraded acoustic models
2. full GMM-HMM model optimized for the domain

Combination

|  |  |  |
| --- | --- | --- |
| data set | size (calls) | combination |
| train | 1612 | (DM1+DM2)\*(SR1+SR2) |
| dev | 506 | (DM1+DM2)\*(SR1+SR2) |
| test | 1117 | DM3\*(SR1+SR2) |

So, to perform well on the test, we should know what the difference of between reinforcement learning and hand-crafted policy is.

### Number of Possible slot values

|  |  |  |  |
| --- | --- | --- | --- |
| slot | | number of values | value examples |
| Goals | area | 5 | centre,north,west,south,east |
| food | 91 | catalan,chinese,christmas,corsica,creative,crossover,cuban |
| name | 113 | ali baba,anatolia,ask,backstreet bistro,bangkok city,bedouin |
| pricerange | 3 | cheap,moderate,expensive |
| Method | | 4 | none,byconstraints,byname,finished,byalternatives |
| Requested slots | | 8 | addr,area,food,phone,pricerange,postcode,signature,name |

### Ontology is given

In DSTC2, the slot value cannot be directly derived from SLU, but needs some understanding.

Take the “method” slot for example. It has to be derived by the following rules:

1. The method becomes `by constraints' if the user gives a constraint specifying a goal for a particular slot. E.g. inform(food=chinese)
2. The method becomes `by alternatives' if the user issues a `reqalts' act.
3. The method becomes `by name' if the user either informs a value for the name slot, or requests information about an offered venue
4. The method changes to `Finished' if the user gives a `bye' act.

The ontology is given, which specified all possible values for each slot, which might be very useful for DSTC3 (in that case, no train data is given for a new data).

### 9 Interested Metrics

Only 9 metrics are interested in this challenge (3\*3).

3 slots:

1. “goals” (is the “joint” of all the four slots.)
2. “requested”,
3. “method”

3 metrics:

* 1. Accuracy
  2. L2
  3. Roc.v2\_ca05

## Baseline

3 baselines are given. 2 of them are from the organizer; the other is the adaptive version for DSTC2 from DSTC1 from one of the previous group [1].

## Results (only accuracy, l2 are considered now)



## MyModel

TODO

# Reference

[1] Z. Wang and O. Lemon. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In Proceedings of the 14th annual SIGdial Meeting on Dis-course and Dialogue, pages 423-432, Metz, France, 2013.