# Summary

# Feature Vector

Models are trained on Train2.

Feature Set: Acts + Unigram

## Decision Tree

ASR\_yes <= 0

| ASR\_correct <= 0

| | ASR\_54c <= 0

| | | request = 0

| | | | ASR\_61c <= 0

| | | | | ASR\_28x <= 0

| | | | | | ASR\_61b <= 0

| | | | | | | ASR\_64 <= 0

| | | | | | | | ASR\_61a <= 0: -1 (4870.0/397.0)

| | | | | | | | ASR\_61a > 0

| | | | | | | | | ASR\_hays <= 0: 1 (121.0/50.0)

| | | | | | | | | ASR\_hays > 0: -1 (24.0/3.0)

| | | | | | | ASR\_64 > 0: 1 (79.0/35.0)

| | | | | | ASR\_61b > 0: 1 (68.0/28.0)

| | | | | ASR\_28x > 0: 1 (38.0/10.0)

| | | | ASR\_61c > 0

| | | | | ASR\_leaving <= 0: 1 (273.0/86.0)

| | | | | ASR\_leaving > 0: -1 (21.0/1.0)

| | | request = 1

| | | | ASR\_from <= 0

| | | | | ASR\_in <= 0

| | | | | | ASR\_to <= 0

| | | | | | | ASR\_thirty <= 0

| | | | | | | | ASR\_at <= 0

| | | | | | | | | ASR\_downtown <= 0

| | | | | | | | | | ASR\_now <= 0

| | | | | | | | | | | ASR\_duquesne <= 0

| | | | | | | | | | | | ASR\_homestead <= 0

| | | | | | | | | | | | | ASR\_pittsburgh <= 0

| | | | | | | | | | | | | | ASR\_mckeesport <= 0

| | | | | | | | | | | | | | | ASR\_oakland <= 0

| | | | | | | | | | | | | | | | ASR\_hill <= 0

| | | | | | | | | | | | | | | | | ASR\_m <= 0

| | | | | | | | | | | | | | | | | | ASR\_braddock <= 0

| | | | | | | | | | | | | | | | | | | ASR\_clock <= 0

| | | | | | | | | | | | | | | | | | | | ASR\_side <= 0

| | | | | | | | | | | | | | | | | | | | | ASR\_swissvale <= 0

| | | | | | | | | | | | | | | | | | | | | | ASR\_airport <= 0

| | | | | | | | | | | | | | | | | | | | | | | ASR\_murray <= 0

| | | | | | | | | | | | | | | | | | | | | | | | ASR\_fifth <= 0: -1 (948.0/239.0)

| | | | | | | | | | | | | | | | | | | | | | | | ASR\_fifth > 0: 1 (22.0/7.0)

| | | | | | | | | | | | | | | | | | | | | | | ASR\_murray > 0: 1 (25.0/10.0)

| | | | | | | | | | | | | | | | | | | | | | ASR\_airport > 0: 1 (27.0/10.0)

| | | | | | | | | | | | | | | | | | | | | ASR\_swissvale > 0: 1 (25.0/6.0)

| | | | | | | | | | | | | | | | | | | | ASR\_side > 0: 1 (29.0/10.0)

| | | | | | | | | | | | | | | | | | | ASR\_clock > 0: 1 (46.0/16.0)

| | | | | | | | | | | | | | | | | | ASR\_braddock > 0: 1 (39.0/11.0)

| | | | | | | | | | | | | | | | | ASR\_m > 0: 1 (72.0/23.0)

| | | | | | | | | | | | | | | | ASR\_hill > 0: 1 (45.0/13.0)

| | | | | | | | | | | | | | | ASR\_oakland > 0: 1 (74.0/18.0)

| | | | | | | | | | | | | | ASR\_mckeesport > 0: 1 (79.0/20.0)

| | | | | | | | | | | | | ASR\_pittsburgh > 0: 1 (39.0/8.0)

| | | | | | | | | | | | ASR\_homestead > 0: 1 (43.0/9.0)

| | | | | | | | | | | ASR\_duquesne > 0: 1 (42.0/5.0)

| | | | | | | | | | ASR\_now > 0: 1 (188.0/24.0)

| | | | | | | | | ASR\_downtown > 0: 1 (148.0/17.0)

| | | | | | | | ASR\_at > 0: -1 (43.0/7.0)

| | | | | | | ASR\_thirty > 0: -1 (34.0/2.0)

| | | | | | ASR\_to > 0: -1 (32.0/1.0)

| | | | | ASR\_in > 0: -1 (32.0/1.0)

| | | | ASR\_from > 0: -1 (41.0/1.0)

| | ASR\_54c > 0: 1 (138.0/28.0)

| ASR\_correct > 0: 0 (48.0/14.0)

ASR\_yes > 0

| expl-conf = 0: -1 (106.0)

| expl-conf = 1: 0 (1486.0/209.0)

## Feature Selection

CfsSubsetEval + GreedyStepwise

28 most significant features

|  |
| --- |
| inform |
| expl-conf |
| affirm |
| ASR\_54c |
| ASR\_60e |
| ASR\_61f |
| ASR\_6d |
| ASR\_aiken |
| ASR\_bellevue |
| ASR\_down |
| ASR\_ebo |
| ASR\_forty |
| ASR\_harwick |
| ASR\_hobart |
| ASR\_mont |
| ASR\_moroeville |
| ASR\_now |
| ASR\_opposite |
| ASR\_out |
| ASR\_s |
| ASR\_township |
| ASR\_wal-mart |
| ASR\_yes |
| ASR\_chesterfield |
| ASR\_found |
| ASR\_information |
| ASR\_today |
| ASR\_university |

# Comparison with Training Set: train2, train3 and train2+3

Model: SVM

Feature: Unigram + Act

Trained on 2, 3, 2+3 respectively

Train1, Test1 and Test2 are from Group A; Test3 and Train3 are both from Group B; Test4 is from Group C. Thus, it might be better to predict Test3 based on Train3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Precision | Recall | F-Measure |
| Train3 | Test1 | 0.824 | 0.796 | 0.805 |
|  | Test2 | 0.823 | 0.754 | 0.776 |
|  | Test3 | 0.851 | 0.856 | 0.852 |
|  | Test4 | 0.848 | 0.778 | 0.8 |
| Train2 | Test1 | 0.832 | 0.841 | 0.835 |
|  | Test2 | 0.828 | 0.794 | 0.807 |
|  | Test3 | 0.664 | 0.634 | 0.645 |
|  | Test4 | 0.85 | 0.816 | 0.828 |
| Train2+3 | Test1 | 0.828 | 0.836 | 0.829 |
|  | Test2 | 0.821 | 0.789 | 0.801 |
|  | Test3 | 0.829 | 0.827 | 0.828 |
|  | Test4 | 0.858 | 0.816 | 0.83 |

Train on 3 on largely improves performance on Test3.

# Other Metrics

I chose four metrics:

* Accuracy: Percent of turns in which the tracker’s 1-best hypothesis is correct. This measures raw 1-best accuracy [The bigger the better]
* Mean reciprocal rank (mrr): Average of 1/R, where R is the rank of the first correct hypothesis. This measures the quality of the ranking [The bigger the better]
* Brier score (l2): The L2 norm between the vector of scores output by dialog state tracker and a vector with 1 in the position of the correct item, and 0 elsewhere. If there are multiple correct items, the correct item which has been assigned the highest score is chosen. This measures the calibration of the scores – ie the extent to which they are good probabilities. [The smaller the better]
* Average score (avgp): This measures the average score assigned to the correct item. [The bigger the better]

## Results:

The 3way\_actnagram model is better on 27 metrics compared to 21 for “best models by others”



Highlighted cell mean it is better.

# CRF model (Conditional Random Field)

The assumption is that the ngam + svm is modeling the performance of SLU and ASR; The CRF is modeling the performance of dialog management. It is because CRF considers the history of dialog.

Data Preparing:

* Split the data into individual dialogs. If the “turn\_index” is 0, then it is the beginning of a new dialog
* Each dialog is a sample is CRF.
* The task of CRF is to get the sequence of labels in the dialog which has the max probability.

## Results:

Act (30)

Unigram (26) by Feature Selection

Metric: Accuracy

Slots: “all” and “joint”

Methods:

* Baseline: top SLU
* Bestbyother: best results got by all entries submitted (not necessary by the same team)
* 3way: Unigram + SVM
* 3way\_actngram: Act + Unigram + SVM (Test3 is trained on Train3; others are trained on train2)
* 3way\_crf\_act: Act + CRF
* 3way\_crf\_actngram: Act + Selected Unigram + CRF

See the table below:

For each row, if the cell is highlighted , it means that it is better than “bestbyother”.



# Error Analysis for 3-way model

Model: SVM

Feature Set: Unigram + Act

## Confusion Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| train | test | a | b | c | <-classified as |
| train2 | train2 | 5635 | 227 | 270 | a=-1 |
|  |  | 151 | 1338 | 18 | b=0 |
|  |  | 325 | 0 | 1381 | c=1 |
|  |  |  |  |  |  |
|  |  | a | b | c | <-classified as |
| train2 | test1 | 5996 | 203 | 423 | a=-1 |
|  |  | 197 | 1816 | 1 | b=0 |
|  |  | 781 | 0 | 668 | c=1 |
|  |  |  |  |  |  |
|  |  | a | b | c | <-classified as |
| train2 | test2 | 5982 | 197 | 1261 | a=-1 |
|  |  | 228 | 1839 | 18 | b=0 |
|  |  | 529 | 7 | 796 | c=1 |
|  |  |  |  |  |  |
|  |  | a | b | c | <-classified as |
| train3 | test3 | 5453 | 225 | 1762 | a=-1 |
|  |  | 203 | 1851 | 31 | b=0 |
|  |  | 447 | 6 | 879 | c=1 |
|  |  |  |  |  |  |
|  |  | a | b | c | <-classified as |
| train2 | test4 | 6897 | 1591 | 964 | a=-1 |
|  |  | 205 | 671 | 505 | b=0 |
|  |  | 1582 | 7 | 842 | c=1 |

For test1, test2, test3, the most significant errors are confusing “-1” and “1”.

For test4, the most significant errors are “-1 -> 0” and “1->-1”.

## Error Examples:

For Test1, Checked 587 errors

Only considered the samples which,

* The correct SLU doesn’t appear in the any of SLUs or System output

Error Types:

* H: Annotation Error
* A: ASR Error
* D: Dialog Management
* S: SLU Error

### Distribution of errors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | H | A | D | S |
| count | 100 | 467 | 8 | 12 |

Examples are in the next page.

Examples: