# Summery

# Why is bad on MRR?

Where, N is the total number of samples, is the rank of the correct answer. If the correct answer doesn’t appear,

In the 3way model, I have only one prediction, so, if the prediction is wrong, it gets 0. However, in other models, they can still get some points if the correct one is in other rank.

## Fix it

Instead of generating only one candidate, I generate three candidates, each one for -1, 0, 1, with the ranking by the probability given by SVM. The scores given by SVM are like “0.667 0.333 0.000” or “0.333 0.333 0.333”.

## New Results

By doing this, all the MRR scores increase without decreasing accuracy. Four of them are even better than “bestbyothers”.

However, the better the MRR scores, the worse the avgp and l2 scores. It makes since, because now, the new model split the confidences score to different SLU. Then, the l2 distance of the new score vector to the correct one will increase too.

I think it is pretty like gambling: you can win more if you want to take more risk by betting only a single choice, but at the same time, you will lose more if you are wrong.

The 3wat model is a gambling system, which has only one candidate. In that case, it can have a better l2 and avgp, but will have a low MRR.



# Trained 2+3

The results below show that training on 2+3 doesn’t give a better model.

Highlighted green like  means it is becoming better. If not it is worse.

Generally speaking, 23 get better and 25 get worse. The results are the same as the performance of the 3way classifier, which shows only the test4 is better when training on 2+3.





# Add new features

* ASR score
* Overlap of input and output (Cosine)
  + To predict whether the user is answering the correct question
  + To predict whether the user just repeats one of the examples in the output
* Overlap between current ASR and previous ASR
  + To predict repeat of the system (it might ask same question again and again)
* Team 6 features (Change raw scores to bins, and then count the number of scores that fall into the bins)
  + SLU score
  + Affirm Score
  + Negate Score
  + Max score
  + Acc Score (no bin)
    - slu + affirm – negate

The bin width is 0.1.

## 3way Result

Feature Set

* act (Binary features for each dialog acts)
* actngram (act + unigram asr)
* enrich1 (actngram + following)
  + Word Count
  + SLU Rank
  + Duration
  + speaking rate
  + slu score
  + slu rank
  + Confirmed?
  + Negated?
  + General Negated?
* enrich2 (enrich1 + following)
  + asr score
  + Overlap of input and output (Cosine)
  + Overlap between current ASR and previous ASR
* enrich3 (enrich2 + following)
  + Team 6 features (bins)
* bins (just the team 6 features)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Precision | Recall | F-Measure |
| act | Test1 | 0.804 | 0.794 | 0.798 |
|  | Test2 | 0.828 | 0.681 | 0.718 |
|  | Test3 | 0.792 | 0.78 | 0.783 |
|  | Test4 | 0.871 | 0.736 | 0.769 |
|  |  |  |  |  |
| actngram | Test1 | 0.832 | 0.841 | 0.835 |
|  | Test2 | 0.828 | 0.794 | 0.807 |
|  | Test3 | 0.851 | 0.856 | 0.852 |
|  | Test4 | 0.85 | 0.816 | 0.828 |
|  |  |  |  |  |
| enrich1 | Test1 | 0.829 | 0.84 | 0.832 |
| +speaking rate | Test2 | 0.829 | 0.797 | 0.809 |
| . | Test3 | 0.868 | 0.871 | 0.867 |
| . | Test4 | 0.834 | 0.804 | 0.816 |
|  |  |  |  |  |
| enrich2 | Test1 | 0.829 | 0.839 | 0.831 |
| +cos | Test2 | 0.829 | 0.798 | 0.81 |
| +asr score | Test3 | 0.869 | 0.872 | 0.867 |
| +cos2 | Test4 | 0.838 | 0.807 | 0.819 |
|  |  |  |  |  |
| enrich3 | Test1 | 0.833 | 0.842 | 0.835 |
| +bins | Test2 | 0.835 | 0.806 | 0.817 |
| +acc | Test3 | 0.87 | 0.873 | 0.868 |
|  | Test4 | 0.841 | 0.82 | 0.828 |
|  |  |  |  |  |
| bins | Test1 | 0.7 | 0.816 | 0.753 |
|  | Test2 | 0.725 | 0.826 | 0.772 |
|  | Test3 | 0.81 | 0.809 | 0.789 |
|  | Test4 | 0.723 | 0.839 | 0.776 |

The best results in F-measure are got by combing all features.

## Combined Result



## Observation

* The *enrich3* model is consistently better than *actngram* model except on test4.

The reason might be:

* + For 3way model, accuracy is more important, so F-measure cannot describe the overall performance. In addition, *actngram* has better accuracy on test4 compared to *enrich3*
* For test3, the *enrich3* model is much better but still cannot beat *bestbyother*

The reason might be:

* + In test3, only 4471 out of 13264 turns has correct label (It is very low)
  + 115 Errors
    - System: “I thought you said START OVER. Is that right?”
    - Input: yes
    - OutAct: expl-conf(act=restart)
    - My Prediction: 0, Actual: -1

I didn’t all slot-value pair in *expl-conf* as candidate, but I shouldn’t in this case, because act=restart doesn’t have required slot.

# New Data (a new domain: restaurant)

TODO