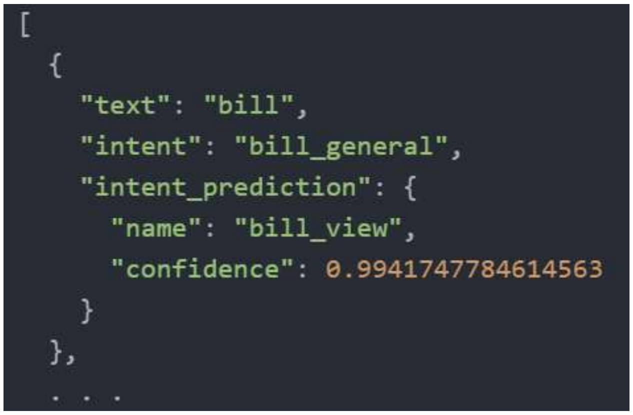
Cross-Validation Report Builder

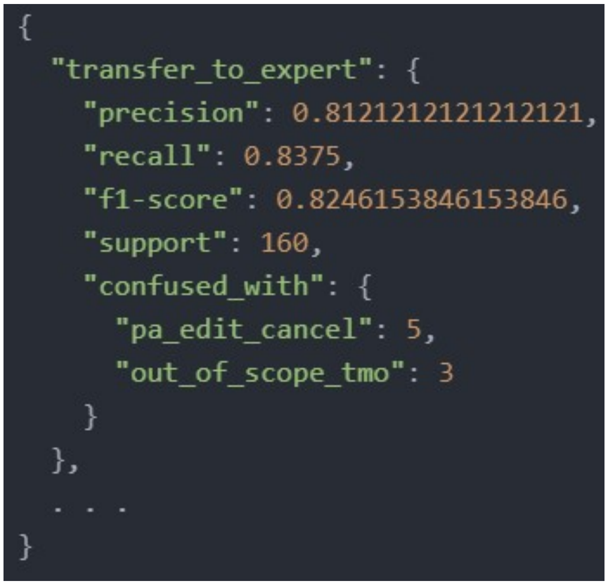
Pre-Use Setup:

This script provides a series of more human-readable output from two Rasa-generated reports, `intent_errors.json` and `intent_report.json`. The files are expected to be in their respective formats in *Figs 1 and 2*.



```
[
  {
    "text": "bill",
    "intent": "bill_general",
    "intent_prediction": {
      "name": "bill_view",
      "confidence": 0.9941747784614563
    }
  },
  . . .
```

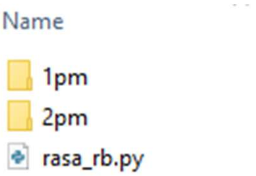
Fig 1: intent_errors.json format



```
{
  "transfer_to_expert": {
    "precision": 0.8121212121212121,
    "recall": 0.8375,
    "f1-score": 0.8246153846153846,
    "support": 160,
    "confused_with": {
      "pa_edit_cancel": 5,
      "out_of_scope_tmo": 3
    }
  },
  . . .
}
```

Fig 2: intent_report.json format

These files should be found in folders generated by Rasa during its self-evaluation. In order to use this script, place it within the directory with the folders generated by Rasa, as seen in *Fig 3*. The script currently only work when 2 reports are present.



Name

1pm

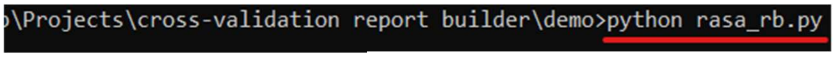
2pm

rasa_rb.py

Fig 3: directory setup

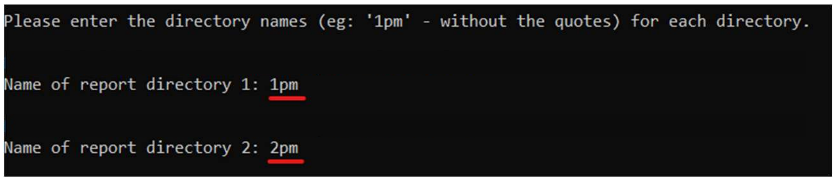
Running the file:

From the CLI, while within the directory hosting the rasa_rg.py file and report directories, run the script (*fig 4*). After starting, follow the instructions to enter the existing directory names and desired output labels for each report (*figs 5 and 6*, respectively). You will be asked to confirm these entries before continuing. The script will update you as it is building the report before creating a new directory, report_summaries, where the built reports will be placed.



```
\Projects\cross-validation report builder\demo>python rasa_rb.py
```

Fig 4: run command

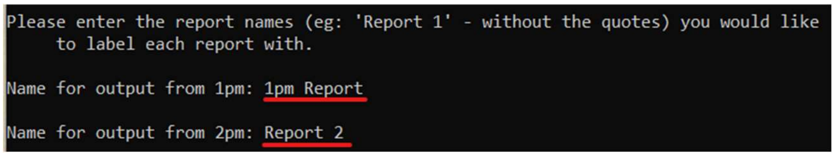


```
Please enter the directory names (eg: '1pm' - without the quotes) for each directory.

Name of report directory 1: 1pm

Name of report directory 2: 2pm
```

Fig 5: filling directory names



```
Please enter the report names (eg: 'Report 1' - without the quotes) you would like
    to label each report with.

Name for output from 1pm: 1pm Report

Name for output from 2pm: Report 2
```

Fig 6: filling output labels

Script Output:

Within the new report_summaries directory, you will find folders with the labels for each report you entered in addition to a pdf comparing the two reports (*fig 7*). Within the folders there will be a "*label* errors" pdf and "*label* figs" pdf (*fig 8*).

Label errors.pdf is simply a series of tables for each instance of an incorrectly predicted intent by the models. The tables contain the text, correct intent, incorrectly labeled intent, and confidence of that incorrect prediction (*fig 9*). The tables attempt to present the errors in a more human-readable format by dynamically wrapping text and changing font sizes, and are limited to 27 table rows per page to prevent overlapping in the case of many long text entries.

Within *label* figs.pdf, the first figures are summary tables of intent accuracy for the given model report. The "confused with" column displays only the two top intents the model confused the current-row intent with. The "samples" column is how many NLU samples the model had to train that particular intent. The tables are sorted by f1-score of each intent, and in the last row is a highlighted weighted average recall for all intents which Rasa uses as the overall "accuracy" of model predictions.

The following figure, Intent f1 Ranking, is a visualization of the previous tables for a less precise but easier digestable sumary of intent performance from the model (*fig 12*). This is followed by un-labeled and labeled scattergrams of the same data, but in another format that also helps visualize the relationship between number of samples for an intent and its precision, recall, and f1 performance (*fig 11)*.

The next two tables display confidence averages for errors with each intent, one being for the confidence when an intent was missed (predicted something other than the correct intent) and the model confidence when an intent was incorrectly predicted (predicted instead of the correct intent), as well as how many occurences the average is representing (*fig 13*). The follow and final table, Intent Pairings, summarizes the intents that were most often confused together and the average confidence of all occurances for a particular pairing (*fig 14*).

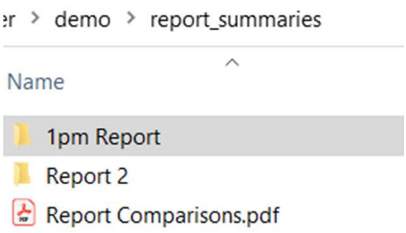


Fig 7: report_summaries folder

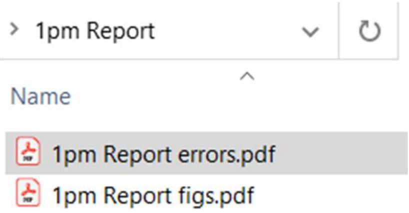
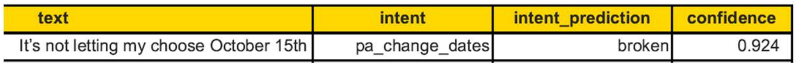
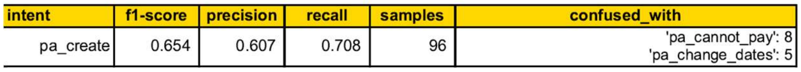


Fig 8: within a report folder



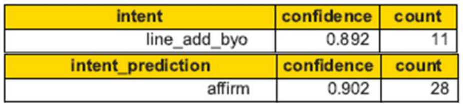
text	intent	intent_prediction	confidence
It's not letting my choose October 15th	pa_change_dates	broken	0.924

Fig 9: Report Intent Errors table from "label errors.pdf"



intent	f1-score	precision	recall	samples	confused_with
pa_create	0.654	0.607	0.708	96	'pa_cannot_pay': 8 'pa_change_dates': 5

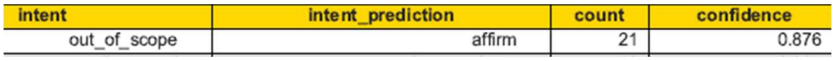
Fig 10: Report Intent Accuracy summary table from "label figs.pdf"



intent	confidence	count
line_add_byo	0.892	11

intent_prediction	confidence	count
affirm	0.902	28

Fig 13: tables for intent error confidence averages



intent	intent_prediction	count	confidence
out_of_scope	affirm	21	0.876

Fig 14: table for mistaken intent pairings

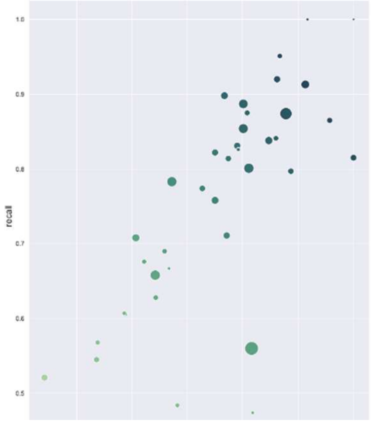


Fig 11: scattergram from "label figs.pdf"

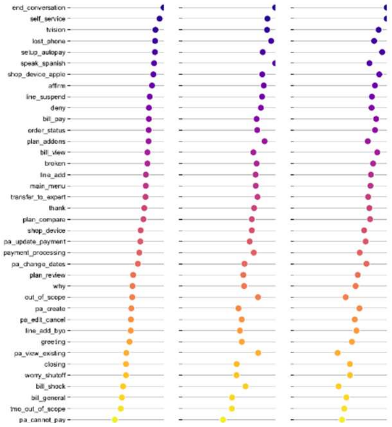


Fig 12: intent ranking from "label figs.pdf"

Back in the report_summaries directory, Report Comparison.pdf has several figures which summarize and compare the performance of each model report analyzed by rasa_rb. The first two are their respective scattergrams overlaid, again unlabeled and labeled, with lines added to indicate the performance shift of each intent tracked (*fig 16*). Green lines indicate an improvement in both recall and precision, brown lines are an improvement in at least one direction, while red lines are a decrease in the performance of both metrics. The next figure, Intent Accuracy Mean Comparison, summarizes the overall performance of each model with the mean of each field marked for each report (*fig 15*). The final figure is a confusion matrix which summarizes the change in intent confusions between each report. Red cells indicate two intents that were confused more often while green cells represent fewer confusions, with the integers for each being the difference between the two reports (*fig 17*).

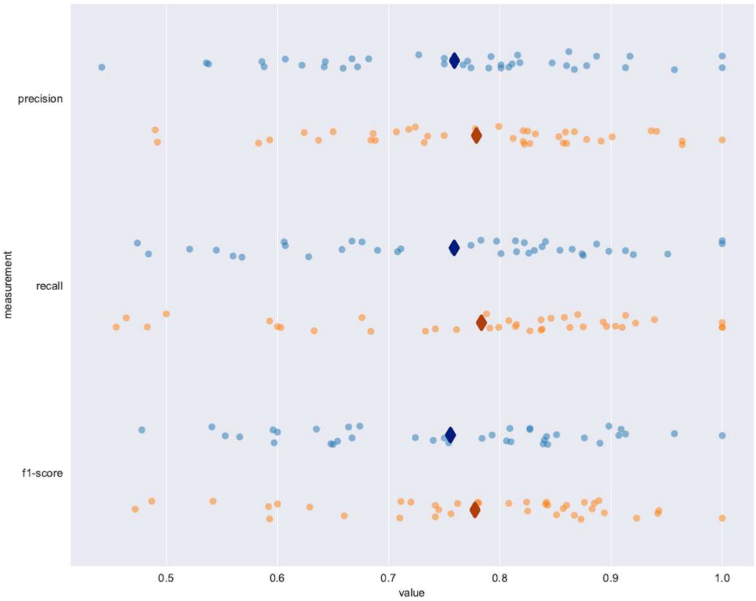


Fig 15: intent accuracy mean comparisons

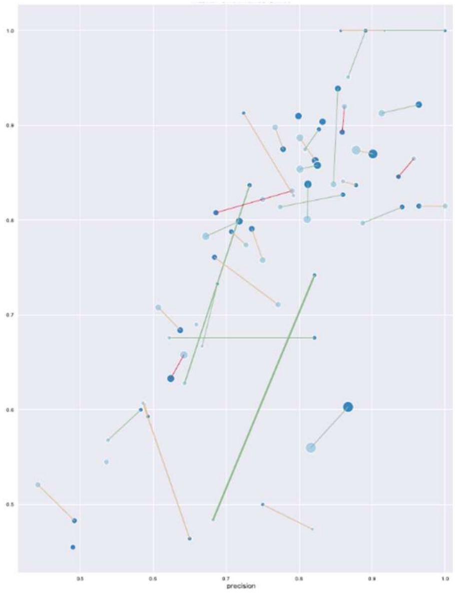


Fig 16: report scattergram shifts

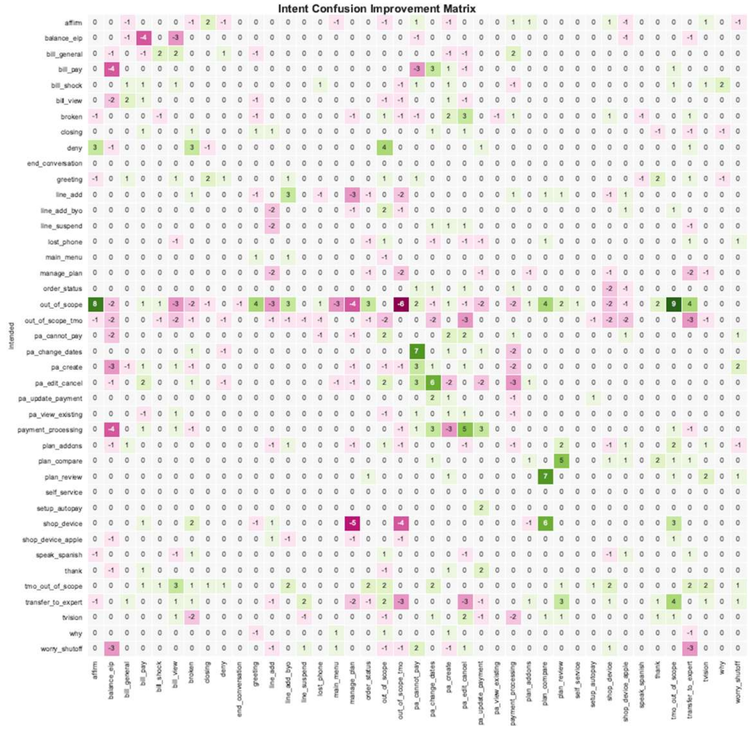


Fig 17: Confusion improvement matrix