Oreen Yousuf

Due: Fri, 12 Feb 2021 23:59

Attempt 1

Lab 1 – NLU for dialogue

Task 1: Evaluate NLU exploratively

1. My input was “turn off the light” and Rasa’s output, which I deem correct, was the following:

{

"text": "turn off the light",

"intent": {

"name": "turn\_off\_light",

"confidence": 0.532149265436195

},

"entities": [],

"intent\_ranking": [

{

"name": "turn\_off\_light",

"confidence": 0.532149265436195

},

{

"name": "turn\_on\_light",

"confidence": 0.31350596394002195

},

{

"name": "vacuum",

"confidence": 0.06391162700944744

},

{

"name": "move\_to\_trash",

"confidence": 0.05330748128987343

},

{

"name": "give",

"confidence": 0.037125662324462336

}

]

1. My new input was “turn the light switch off”, and Rasa’s output was the following:

{

"text": "turn the light switch off",

"intent": {

"name": "turn\_off\_light",

"confidence": 0.5465019814383855

},

"entities": [],

"intent\_ranking": [

{

"name": "turn\_off\_light",

"confidence": 0.5465019814383855

},

{

"name": "turn\_on\_light",

"confidence": 0.2812434248903737

},

{

"name": "vacuum",

"confidence": 0.0710866112896963

},

{

"name": "move\_to\_trash",

"confidence": 0.058760442037139975

},

{

"name": "give",

"confidence": 0.04240754034440447

}

]

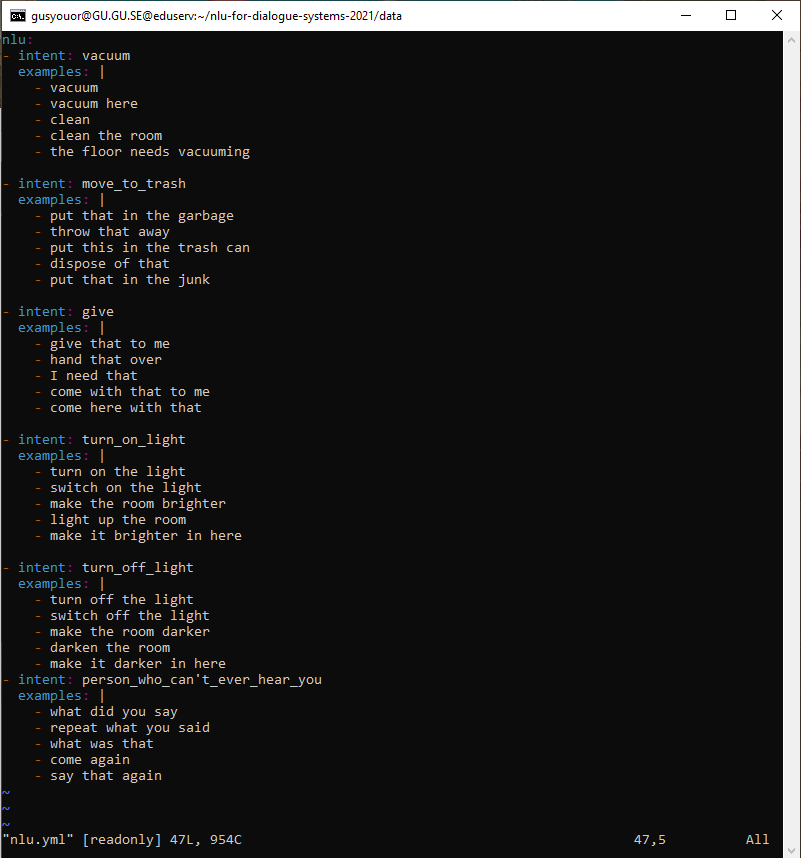
}

I judge the output as being correct, however the confidence of Rasa’s output changed for this new input. Surprisingly it went up. I think this is due to the inclusion of the word “switch” adding additional confidence when the phrase “turn off” was already included. Perhaps it gave more data pointing to the turn\_off\_light intent. I played around with another input, closer to my original utterance – “turn the light off” – and the confidence went down, but just by 0.00000001.

1. I tried this with three different utterances that were unrelated to the training data. My inputs were “buy a car”, “book the flight” and “look over there.” All three results from Rasa were much less confident in their interpretations than the actions took in part A and B, obviously. The outputs’ highest-ranking intents had corresponding confidences ranging from 30-40%. They’re seemingly random and nonsensical, but if I had to analyze their interpretations, I’d guess that they feed off of sentence/utterance patterns and/or Part of Speech (POS). For the second utterance I inputted, “book the flight”, the most confident intent in Rasa’s output was for the “intent: give”. This is perhaps that the pattern of my utterance has a similar POS order (verb, article, noun) or maybe it’s just because “flight” is similar to “light.”

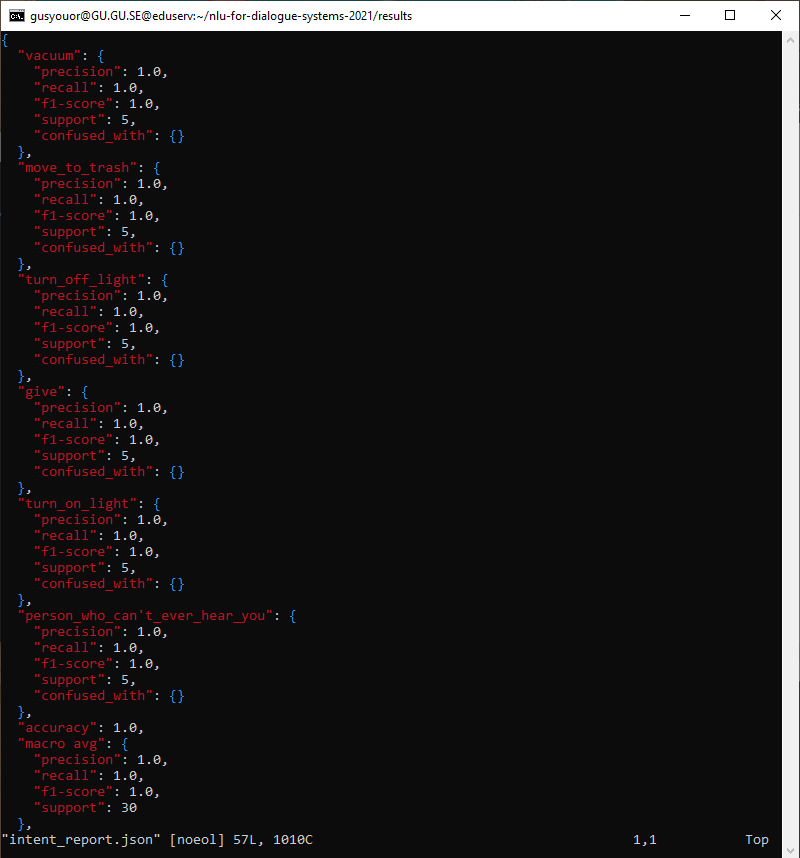
Task 2: Add an intent

The updated nlu.yml file, with my new intent, person\_who\_can’t\_ever\_hear\_you, is at the bottom of the below picture:

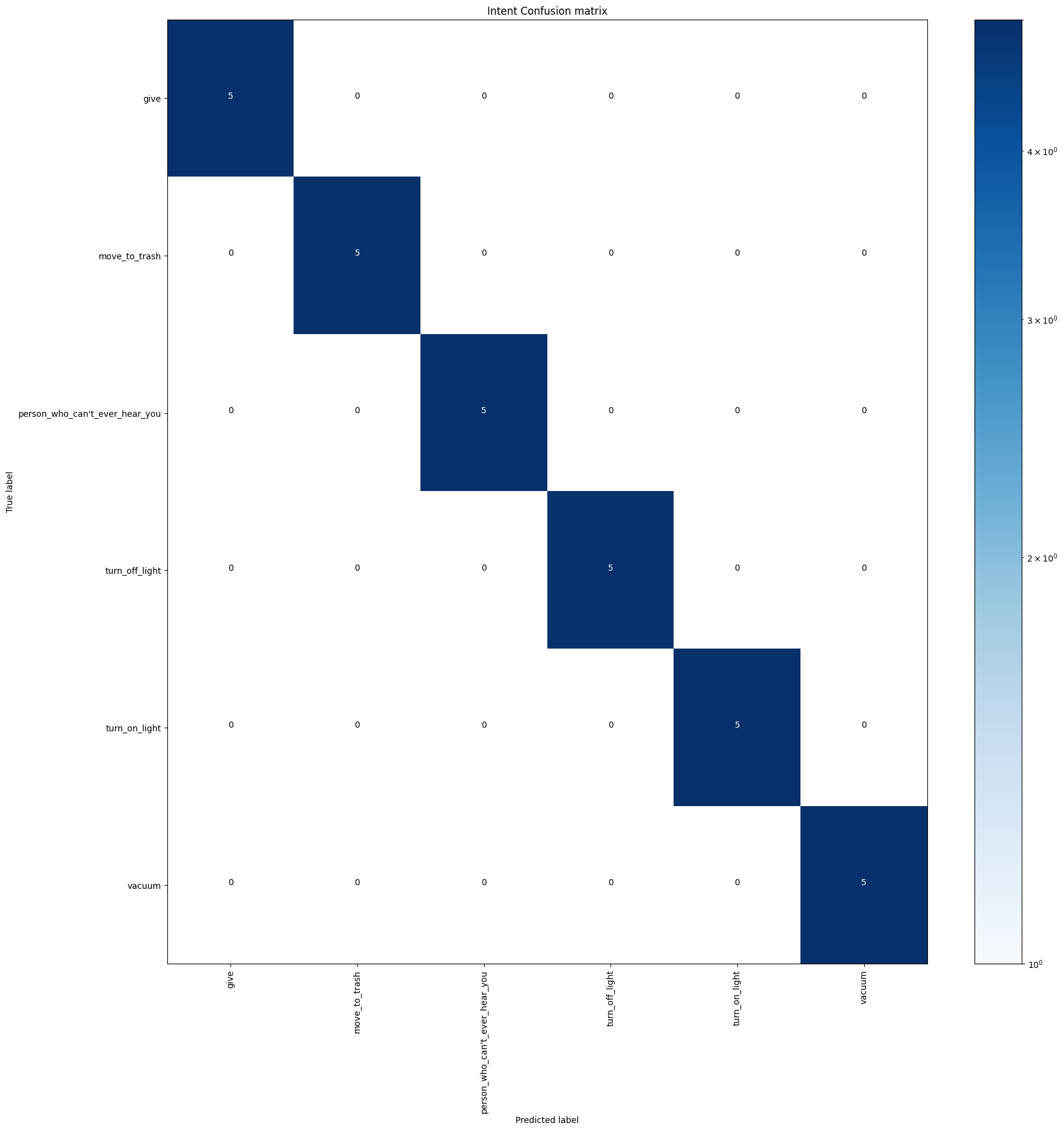


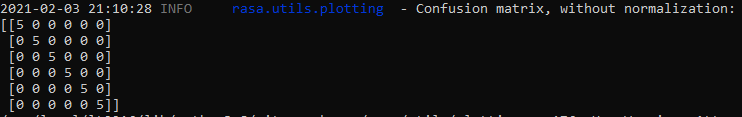
Task 3: Evaluate NLU systematically

1. Precision and recall for macro average (i.e., average precision and recall across all intents):



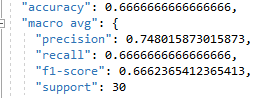
Confusion matrix for intents:





The confusion matrix looks like the two pictures directly above. They have “perfect” scores because the model was tested on the same dataset as its training data, hence the “5” in the middle. Because the confusion matrix evaluates the performance of a classification mode, comparing actual target values with those predicted by the machine learning model, the “perfect” 5 in a diagonal trajectory shows that the model was 100% (1.0) accurate in precision and recall.

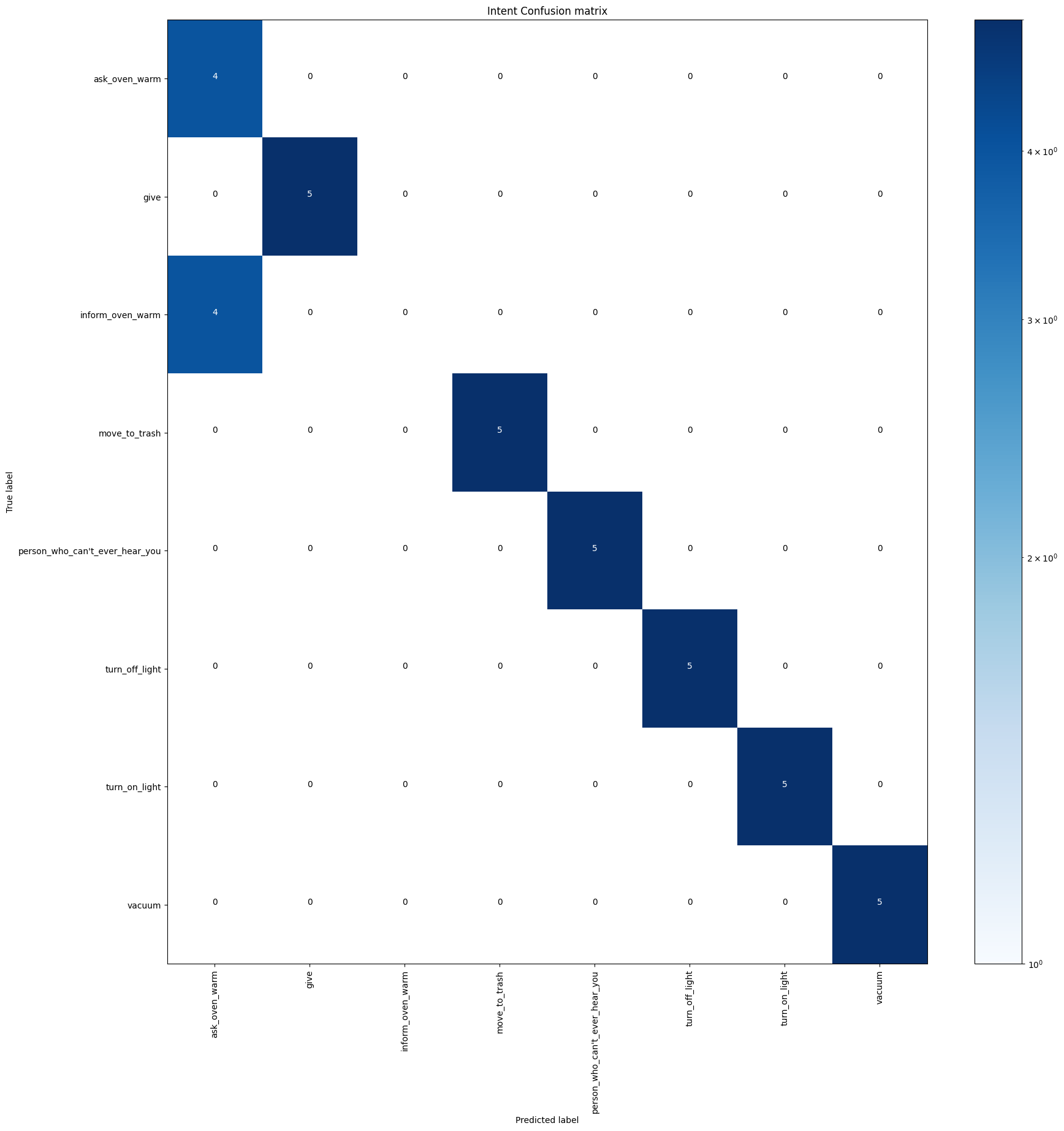
1. Step A yielded us 1.0 (100%) precision, recall and, as a result of the these, f-1 score. The only useful thing about having the testing data for a model be the same as said model’s training data is to ensure that the model is accurate and didn’t suffer from intermediate alterations between training and testing. Other than that, it isn’t very useful; continuous use of this can lead to overfitting, which is hard to realize after several successive tests as your model’s performance on the test set is good. Testing different data than your training data on the model is healthy for your model as it’ll allow you to accurately evaluate whether or not overfitting is occurring.
2. Below is the precision and recall for the macro average:



1. Step C attempted to divide the entire data set into two subsets of training and test data sets. Compared with step A, the cross-validation creates a model that generalizes better to new data. This doesn’t give us completely accurate and precise predictions, as some predictions are incorrect. Pleasantly however, the model performs roughly equally well on the test data as it does on the training data and the model does not overfit the training data.

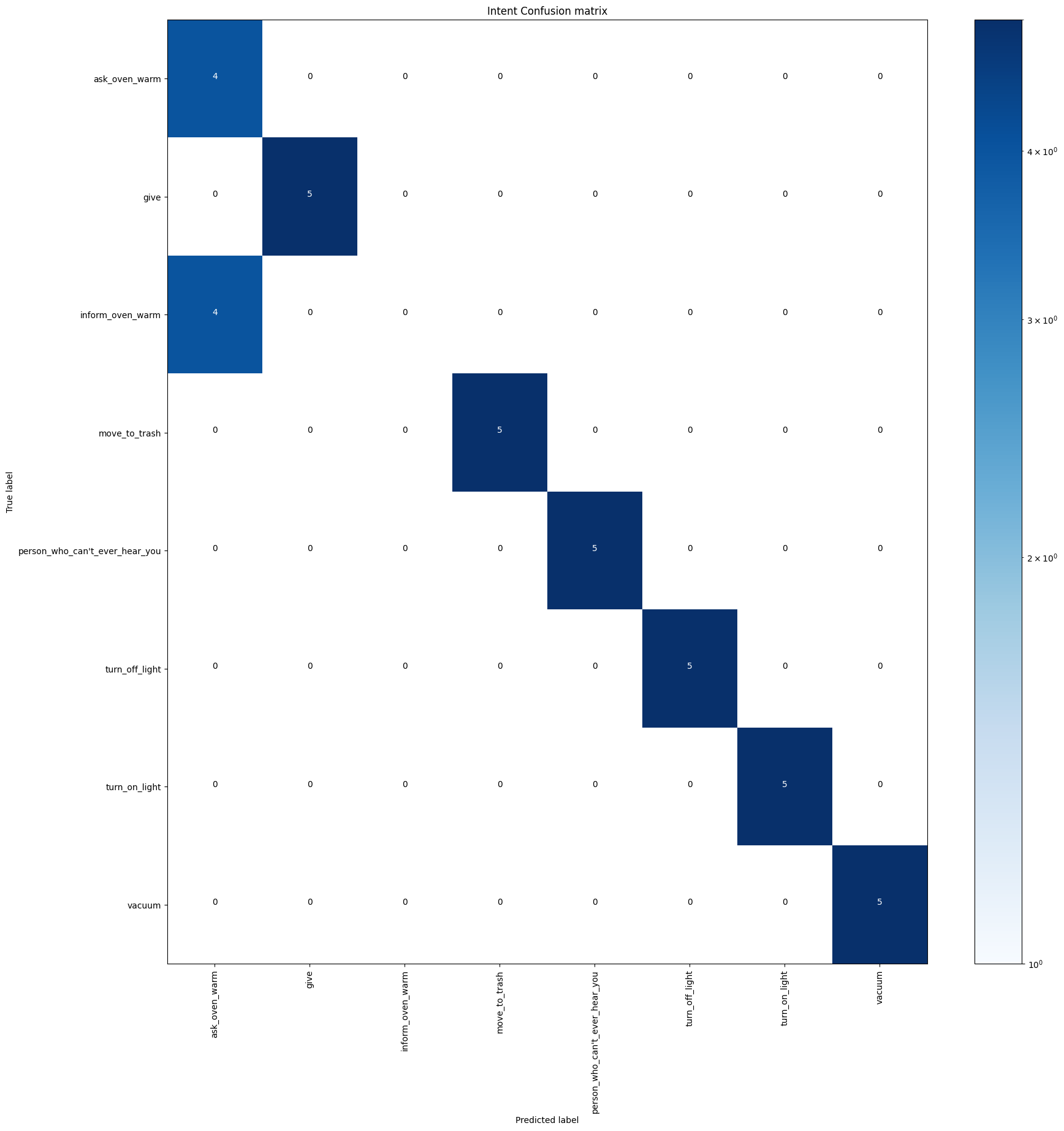
Task 4: Overcome a limitation

1. Confusion matrix for intents:



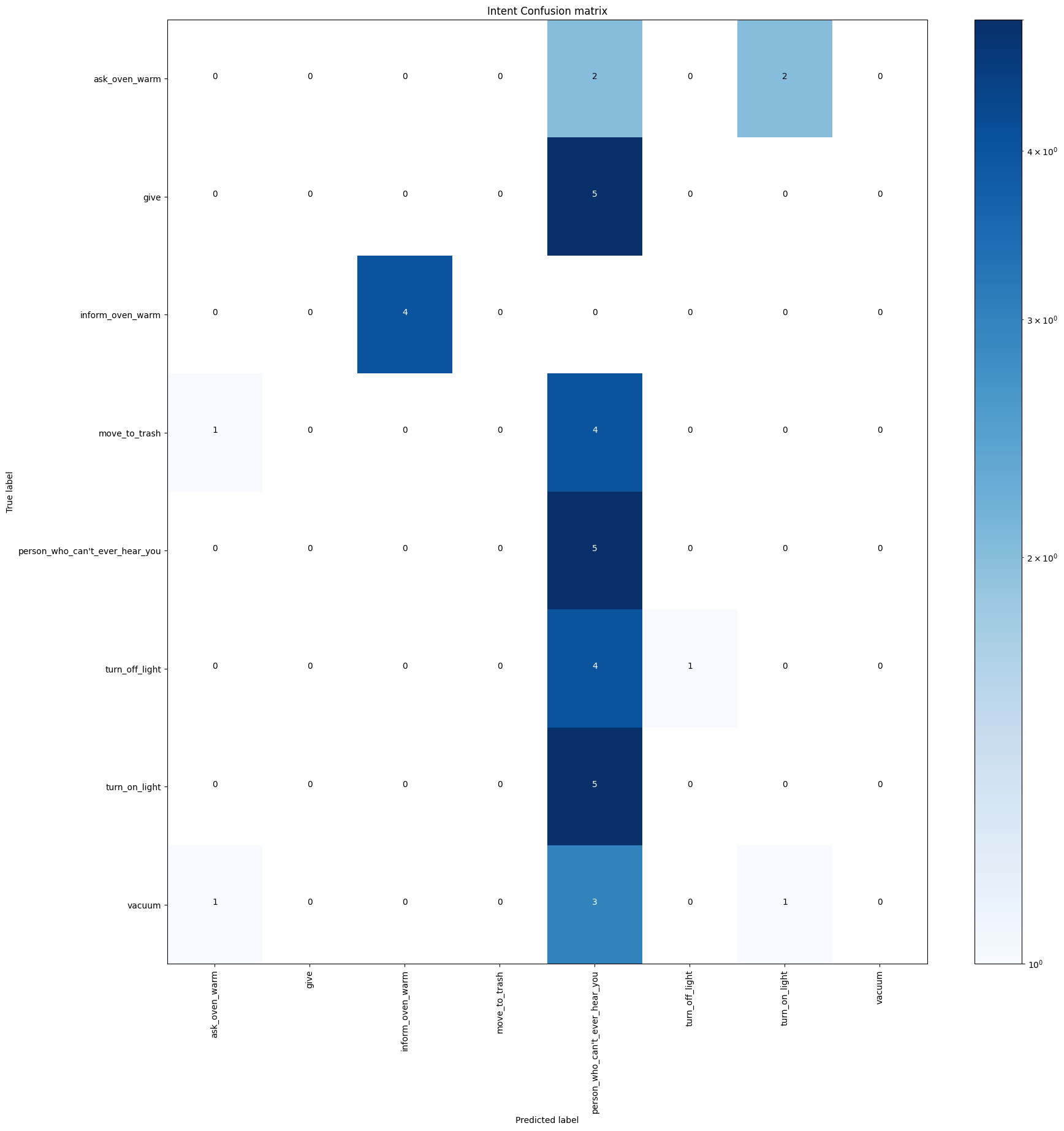
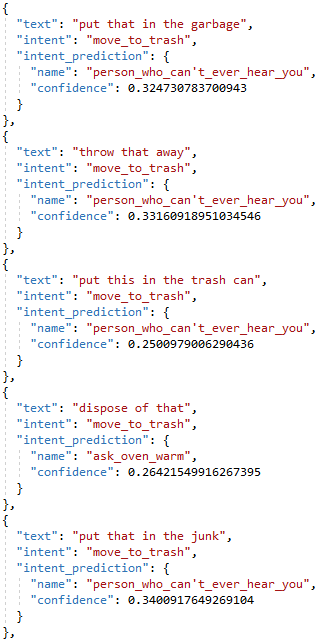
There are intent classification errors for the stove-related intents, namely that the “ask” and “inform” intents are switched; predicting each other incorrectly. This can reasonably be explained by the almost identical words in each intent’s list of utterances.

1. Changing the way sentence vectors are calculated from mean pooling to max pooling didn’t have an affect on the outcome. Below are the confusion matrix and intent errors after training the model after changing to max pooling:



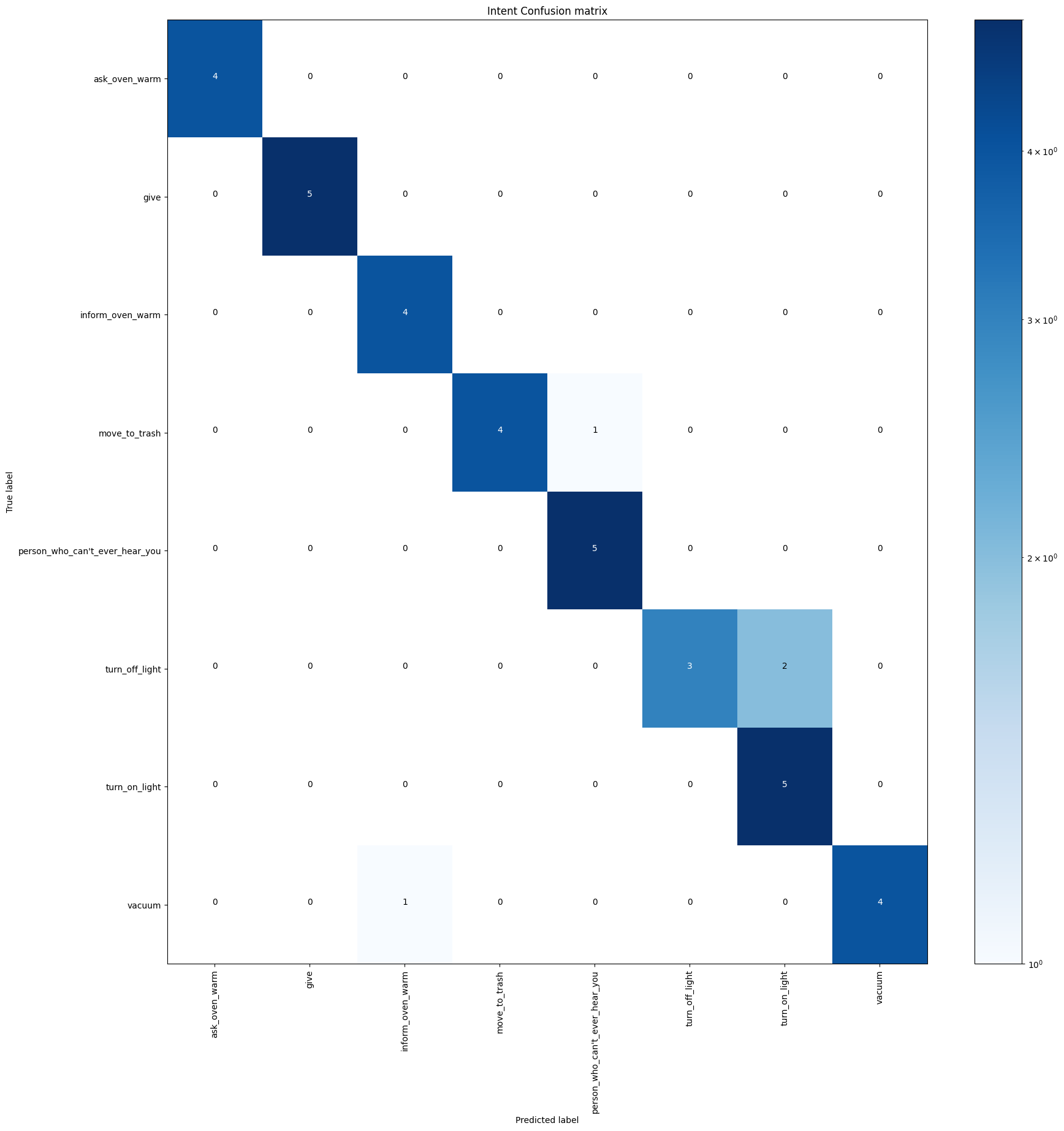


1. I believe the misclassifications of intents occur in steps A and B simply due to how similar the vocabulary is between the intents “ask\_oven\_warm” and “inform\_oven\_warm”. Quite literally every single word is the same between each corresponding utterance; word order is the only slight difference made to meet the grammar needed for a question (for ask\_oven\_warm) and for a statement (for inform\_oven\_warm). More importantly, the model that was used to train this model used word vectors but doesn’t preserve sentence order, so this explains why the stove intents were confused for each other - they could potentially look the same to the model, as the model only sees the same word vectors from the two intents.
2. Below are the confusion matrix and a snippet of the full intent errors after testing the model with the training data after training it on the DIET attention-based model:

There are several intent classification errors. Something of note is that the model is predicting the new intent I inserted at a high rate for the majority of intents tested on.

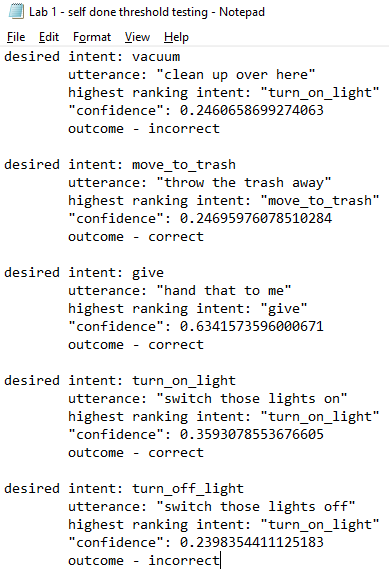
1. I increased the epochs to 50 and it yielded no incorrect predictions of the stove intents pairing to each other. Confusion matrix when epochs changed to 50:



1. The DIET model includes the pre-trained model “distilbert” and utilizes ngrams for its tokenization, which is better than the previous model which simply made use of word vectors and lacked memory preservation of sentence order. More importantly, I believe, the use of epochs, a hyperparameter that allows the DIET model to traverse through the entire training dataset, aids immensely. As the number of epochs increased, the learning algorithm gained the ability to work through the dataset and update the internal model parameters, which I believe allows it to make a more accurate prediction than the previous one.

Task 5: Pinpoint confidence threshold (\*)

My suggested threshold for my system’s dialogue manager is 0.25. My decision stems from a log of experiments I conducted on the model where I inputted utterances related to pre-determined intents that weren’t included in the utterance dataset. Below is a snippet of the log:



There is a noticeable downward tick in correct outcomes when the model’s highest-ranking intent has a confidence level around the early-to-mid-20 percentile (0.2-0.24). This snippet is reflective of the remaining portion of the log. A confidence threshold would allow the model to mitigate uncertainty in proclaiming an incorrect intent by asking the user for aid in whether or not it should progress through with the impending prediction is has on hold. To set my confidence threshold as X = 0.25 would ensure the avoidance of a great deal of incorrect outcomes based on my experiments.