Benchmarking Natural Language Understanding Services for building Conversational Agents

Xingkun Liu, Arash Eshghi, Pawel Swietojanski and Verena Rieser

Abstract We have recently seen the emergence of several publicly available Natural Language Understanding (NLU) toolkits, which map user utterances to structured, but more abstract, Dialogue Act (DA) or Intent specifications, while making this process accessible to the lay developer. In this paper, we present the first wide coverage evaluation and comparison of some of the most popular NLU services, on a large, multi-domain (21 domains) dataset of 25K user utterances that we have collected and annotated with Intent and Entity Type specifications and which will be released as part of this submission.\(^1\) The results show that on Intent classification Watson significantly outperforms the other platforms, namely, Dialogflow, LUIS and Rasa; though these also perform well. Interestingly, on Entity Type recognition, Watson performs significantly worse due to its low Precision\(^2\). Again, Dialogflow, LUIS and Rasa perform well on this task.

1 Introduction

Spoken Dialogue Systems (SDS), or Conversational Agents are ever more common in home and work environments, and the market is only expected to grow. This has prompted industry and academia to create platforms for fast development of SDS, with interfaces that are designed to make this process easier and more accessible to those without expert knowledge of this multi-disciplinary research area.

One of the key SDS components for which there are now several such platforms available is the Natural Language Understanding (NLU) component, which maps individual utterances to structured, abstract representations, often called Dialogue Acts (DAs) or Intents together with their respective arguments that are usually Named Entities within the utterance. Together, the representation is taken to specify the semantic content of the utterance as a whole in a particular dialogue domain.

---

\(^1\) https://github.com/xliuhw/NLU-Evaluation-Data

\(^2\) At the time of producing the camera-ready version of this paper, we noticed the seemingly recent addition of a ‘Contextual Entity’ annotation tool to Watson, much like e.g. in Rasa. We’d therefore like to stress that this paper does not include an evaluation of this feature in Watson NLU.
In the absence of reliable, third-party – and thus unbiased – evaluations of NLU toolkits, it is difficult for users (which are often conversational AI companies) to choose between these platforms. In this paper, our goal is to provide just such an evaluation: we present the first systematic, wide-coverage evaluation of some of the most commonly used NLU services, namely: Rasa\(^4\), Watson\(^5\), LUIS\(^6\) and Dialogflow\(^7\). The evaluation uses a new dataset of 25k user utterances which we annotated with Intent and Named Entity specifications. The dataset, as well as our evaluation toolkit will be released for public use.

2 Related Work

To our knowledge, this is the first wide coverage comparative evaluation of NLU services - those that exist tend to lack breadth in Intent types, Entity types, and the domains studied. For example, recent blog posts \([3, 4]\), summarise benchmarking results for 4 domains, with only 4 to 7 intents for each. The closest published work to the results presented here is by \([1]\), who evaluate 6 NLU services in terms of their accuracy (as measured by precision, recall and F-score, as we do here) on 3 domains with 2, 4, and 7 intents and 5, 3, and 3 entities respectively. In contrast, we consider the 4 currently most commonly used NLU services on a large, new data set, which contains 21 domains of different complexities, covering 64 Intents and 54 Entity types in total. In addition, \([2]\) describe an analysis of NLU engines in terms of their usability, language coverage, price etc., which is complimentary to the work presented here.

3 Natural Language Understanding Services

There are several options for building the NLU component for conversational systems. NLU typically performs the following tasks: (1) Classifying the user Intent or Dialogue Act type; and (2) Recognition of Named Entities (henceforth NER) in an utterance\(^8\). There are currently a number of service platforms that perform (1) and (2): commercial ones, such as Google’s Dialogflow (formerly Api.ai), Microsoft’s LUIS, IBM’s Watson Assistant (henceforth Watson), Facebook’s Wit.ai, Amazon Lex, Recast.ai, Botfuel.io; and open source ones, such as Snips.ai\(^9\) and Rasa. As mentioned above, we focus on four of these: Rasa, IBM’s Watson, Microsoft’s LUIS and Google’s Dialogflow. In the following, we briefly summarise and discuss their various features. Table 1 provides a summary of the input/output formats for each of the platforms.

\(^3\) according to anecdotal evidence from academic and start-up communities  \(^4\) https://rasa.com/  
\(^5\) https://www.ibm.com/watson/ai-assistant/  \(^6\) https://www.luis.ai/home  
\(^7\) https://dialogflow.com/  
\(^8\) Note that, one could develop one’s own system using existing libraries, e.g. sklearn libraries http://scikit-learn.org/stable/, spaCy https://spacy.io/, but a quicker and more accessible way is to use an existing service platform.  
\(^9\) was not yet open source when we were doing the benchmarking, and was later on also introduced in https://arxiv.org/abs/1805.10190
Particular Entities and Entity types tend to be dependent on particular Intent types, e.g. with a ‘set alarm’ intent one would expect a time stamp as its argument. Therefore we think that joint models, or models that treat Intent & Entity classification together would perform better. We were unable to ascertain this for any of the commercial systems, but Rasa treats them independently (as of Dec 2018). (4) None of the platforms use dialogue context for Intent classification and NER - this is another significant limitation, e.g. in understanding elliptical or fragment utterances which depend on the context for their interpretation.

| Service | Input (Training)                                                                 | Output (Prediction)                                                                 |
|---------|---------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Rasa    | JSON or Markdown. Utterances with annotated intents and entities. Can provide synonym and regex features. | JSON. The intent and intent ranking with confidence. A list of entities without scores. |
| Dialogflow | JSON. List of all entity type names and values/synonyms. Utterance samples with annotated intents and entities. Need to specify the expected returning entities as parameters for each intent. | JSON. The intent and entities with values. Overall score returned, not specific to Intent or Entity. Other returned info related to dialogue app. |
| LUIS | JSON. Phrase list and regex patterns as model features, hierarchical and composites entities. List of all intents and entity type names. Utterance samples with annotated intents and entities. | JSON. The intent with confidence. A list of entities with scores |
| Watson | CSV. List of all utterances with Intent label. List of all Entities with values. No annotated entities in an utterance needed. | JSON. The intent with confidence. A list of entities and confidence for each. Other info related to dialogue app. |

Table 1: Input Requirements and Output of NLU Services

4 Data Collection and Annotation

The evaluation of NLU services was performed in the context of building a SDS, aka Conversational Interface, for a home assistant robot. The home robot is expected to perform a wide variety of tasks, ranging from setting alarms, playing music, search, to movie recommendation, much like existing commercial systems such as Microsoft’s Cortana, Apple’s Siri, Google Home or Amazon Alexa. Therefore the NLU component in a SDS for such a robot has to understand and be able to respond to a very wide range of user requests and questions, spanning multiple domains, unlike a single domain SDS which only understands and responds to the user in a specific domain.

4.1 Data Collection: Crowdsourcing setup

To build the NLU component we collected real user data via Amazon Mechanical Turk (AMT). We designed tasks where the Turker’s goal was to answer questions about how people would interact with the home robot, in a wide range of scenarios designed in advance, namely: alarm, audio, audiobook, calendar, cooking, datetime, email, game, general, IoT, lists, music, news, podcasts, general Q&A, radio, recommendations, social, food takeaway, transport, and weather.

The questions put to Turkers were designed to capture the different requests within each given scenario. In the ‘calendar’ scenario, for example, these pre-
designed intents were included: ‘set_event’, ‘delete_event’ and ‘query_event’. An example question for intent ‘set_event’ is: “How would you ask your PDA to schedule a meeting with someone?” for which a user’s answer example was “Schedule a chat with Adam on Thursday afternoon”. The Turkers would then type in their answers to these questions and select possible entities from the pre-designed suggested entities list for each of their answers. The Turkers didn’t always follow the instructions fully, e.g. for the specified ‘delete_event’ Intent, an answer was: “PDA what is my next event?”; which clearly belongs to ‘query_event’ Intent. We have manually corrected all such errors either during post-processing or the subsequent annotations.

The data is organized in CSV format which includes information like scenarios, intents, user answers, annotated user answers etc. (See Table 4 in Appendix). The split training set and test set were converted into different JSON formats for each platform according to the specific requirements of the each platform (see Table 1).

Our final annotated corpus contains 25716 utterances, annotated for 64 Intents and 54 Entity Types.

4.2 Annotation & Inter-annotator Agreement
Since there was a predetermined set of Intents for which we collected data, there was no need for separate Intent annotations (some Intent corrections were needed). We therefore only annotated the data for Entity Tokens & Entity Types. Three students were recruited to do the annotations. To calculate inter-annotator agreement, each student annotated the same set of 300 randomly selected utterances. Each student then annotated a third of the whole dataset, namely, about 8K utterances for annotation. We used Fleiss’s Kappa, suitable for multiple annotators. A match was defined as follows: if there was any overlap between the Entity Tokens (i.e. Partial Tokens Matching), and the annotated Entity Types matched exactly. We achieved moderate agreement ($\kappa = 0.69$) for this task.

5 Evaluation Experiments
In this section we describe our evaluation experiments, comparing the performance of the four systems outlined above.

5.1 Train & Test Sets
Since LUIS caps the size of the training set to 10K, we chose 190 instances of each of the 64 Intents at random. Some of the Intents had slightly fewer instances than 190. This resulted in a sub-corpus of 11036 utterances covering all the 64 Intents and 54 Entity Types. The Appendix provides more details: Table 5 shows the number of the sentences for each Intent. Table 6 lists the number of entity samples for each Entity Type. For the evaluation experiments we report below, we performed 10 fold cross-validation with 90% of the subcorpus for training and 10% for testing in each fold.10

10 We also note here that our dataset was inevitably unbalanced across the different Intents & Entities: e.g. some Intents had much fewer instances: iot wemo had only 77 instances. But this would affect the performance of the four platforms equally, and thus does not confound the results presented below.
5.2 System Versions & Configurations

Our latest evaluation runs were completed by the end of March 2018. The service API used was V1.0 for Dialogflow, V2.0 for LUIS. Watson API requests require data as a version parameter which is automatically matched to the closest internal version, where we specified 2017/04/21\(^\text{11}\). In our conversational system we run the open source Rasa as our main NLU component because it allows us to have more control over further developments and extensions. The evaluation done for Rasa was on Version 0.10.5, and we used its spacy\_sklearn pipeline which uses Conditional Random Fields for NER and sk-learn (scikit-learn) for Intent classifications. Rasa also provides other built-in components for the processing pipeline, e.g. MITIE, or latest tensorflow\_embedding pipeline.

6 Results & Discussion

We performed 10-fold cross validation for each of the platforms and pairwise t-tests to compare the mean F-scores of every pair of platforms. The results in Table 2 show the micro-average\(^\text{12}\) scores for Intent and Entity Type classification over 10-fold cross validation. Table 3 shows the micro-average F-scores of each platform after combining the results of Intents and Entity Types. Table 7 and Table 8 in the Appendix show the detailed confusion matrices used to calculate the scores of Precision, Recall and F1 for Intents and Entities.

| Intent | Prec | Rec | F1  |   | Entity | Prec | Rec | F1  |
|--------|------|-----|-----|---|--------|------|-----|-----|
| Rasa   | 0.863| 0.863| 0.863| 0.859| 0.694| 0.768|
| Dialogflow | 0.870| 0.859| 0.864| 0.782| 0.709| 0.743|
| LUIS   | 0.855| 0.855| 0.855| 0.837| 0.725| 0.777|
| Watson | 0.884| 0.881| 0.882| 0.835| 0.787| 0.488|

Table 2: Overall Scores for Intent and Entity

Table 3: Combined Overall Scores

Performing significance tests on separate Intent and Entity scores in Table 2 revealed: For Intent, there is no significant difference between Dialogflow, LUIS and Rasa. Watson F1 score (0.882) is significantly higher than other platforms ($p < 0.05$, with large or very large effects sizes - Cohen’s D). However, for Entities, Watson achieves significantly lower F1 scores ($p < 0.05$, with large or very large effects sizes - Cohen’s D) due to its very low Precision. One explanation for this is the high number of Entity candidates produced in its predictions, leading to a high number of False Positives\(^\text{13}\). It also shows that there are significant differences for Entity F1 score between Dialogflow, LUIS and Rasa. LUIS achieved the top F1 score (0.777) on Entities.

Table 3 shows that all NLU services have quite close F1 scores except for Watson which had significantly lower score ($p < 0.05$, with large or very large effects sizes

\(^{11}\) At the time of producing the camera-ready version of this paper, we noticed the seemingly recent addition of a ‘Contextual Entity’ annotation tool to Watson, much like e.g. in Rasa. We'd like to stress that this paper does not include an evaluation of this feature in Watson NLU

\(^{12}\) Micro-average sums up the individual TP, FP, and FN of all Intent/Entity classes to compute the average metric.

\(^{13}\) Interestingly, Watson only requires a list of possible entities rather than entity annotation in utterances as other platforms do (See Table 1)
- Cohen’s D) due to its lower entity score as discussed above. The significance test shows no significant differences between Dialogflow, LUIS and Rasa.

The detailed data analysis results in the Appendix (see Table 5 and Table 6) for fold-1 reveal that distributions of Intents and Entities are imbalanced in the datasets. Also, our data contains some noisy Entity annotations, often caused by ambiguities, which our simplified annotation scheme was not able to capture. For example, an utterance in the pattern “play xxx please” where xxx could be any entity from song name, audiobook name, radio name, podcasts name or game name, e.g., “play space invaders please” which could be annotated as [song name : space invaders] or [game name : space invaders]. This type of Intent ambiguity that can only be resolved by more sophisticated approaches that incorporate domain knowledge and the dialogue context. Nevertheless, despite the noisiness of the data, we believe that it represents a real-world use case for NLU engines.

7 Conclusion

The contributions of this paper are two-fold: First, we present and release a large NLU dataset in the context of a real-world use case of a home robot, covering 21 domains with 64 Intents and 54 Entity Types. Secondly, we perform a comparative evaluation on this data of some of the most popular NLU services – namely the commercial platforms Dialogflow, LUIS, Watson and the open source Rasa.

The results show they all have similar functions/features and achieve similar performance in terms of combined F-scores. However, when dividing out results for Intent and Entity Type recognition, we find that Watson has significant higher F-scores for Intent, but significantly lower scores for Entity Type. This was due to its high number of false positives produced in its Entity predictions. As noted earlier, we have not here evaluated Watson’s recent ‘Contextual Entity’ annotation tool.

In future work, we hope to continuously improve the data quality and observe its impact on NLU performance. However, we do believe that noisy data presents an interesting real-world use-case for testing current NLU services. We are also working on extending the data set with spoken user utterances, rather than typed input. This will allow us to investigate the impact of ASR errors on NLU performance.

References

1. Daniel Braun, Adrian Hernandez Mendez, Florian Matthes and Manfred Langen (2017) Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In: Proceedings of SIGDIAL 2017, 174–185.
2. Massimo Canonico and Luigi De Russis (2018) A Comparison and Critique of Natural Language Understanding Tools. In: Proceedings of CLOUD COMPUTING 2018.
3. Caroline Wisniewski, Clément Delpuech, David Leroy, Fran ois Pivan and Joseph Dureau (2017) Benchmarking Natural Language Understanding Systems. https://snips.ai/content/sdk-benchmark-visualisation/
4. Alice Coucke, Adrien Ball, Clément Delpuech, Clément Doumouro, Sylvain Raybaud, Thibault Gisselbrecht and Joseph Dureau (2017) Benchmarking Natural Language Understanding Systems: Google, Facebook, Microsoft, Amazon, and Snips. https://medium.com/snips-ai/benchmarking-natural-language-understanding-systems-google-facebook-microsoft-and-snips-2b88c0f9fbb1

14 Tables for other folds are omitted for space reason
5. Nguyen Trong Canh (2018) Benchmarking Intent Classification services – June 2018. https://medium.com/botfuel/benchmarking-intent-classification-services-june-2018-eb8684a1e55f
Appendix

We provide some examples of the data annotation and the training inputs to each of the 4 platforms in Table 4, Listing 1, 2, 3 and 4.

We also provide more details on the train and test data distribution, as well as the Confusion Matrix for the first fold (Fold_1) of the 10-Fold Cross Validation. Table 5 shows the number of the sentences for each Intent in each dataset. Table 6 lists the number of entity samples for each Entity Type in each dataset. Table 7 and Table 8 show the confusion matrices used to calculate the scores of Precision, Recall and F1 for Intents and Entities. The TP, FP, FN and TN in the tables are short for True Positive, False Positive, False Negative and True Negative respectively.

| userid | answerid | scenario | intent | answer | annotation |
|--------|----------|----------|--------|--------|------------|
| 1      | 2        | alarm    | set    | wake me up at [time: nine am] on [date: friday] |
| 2      | 558      | alarm    | remove | cancel my [time: seven am] alarm |
| 2      | 559      | alarm    | remove | remove the alarm set for [time: ten pm] |
| 2      | 561      | alarm    | query  | what alarms I have set |
| 502    | 12925    | calendar | query  | what is the time for [event_name: jimmy's party] |
| 633    | 17462    | calendar | query  | what is up in my schedule [date: today] |
| 2      | 564      | calendar | remove | please cancel all my events for [date: today] |
| 2      | 556      | play     | mute   | I'd like to hear [artist_name: queen's] [song_name: barcelona] |
| 63     | 2813     | play     | radio  | play a [radio_name: pop station] on the radio |
| 740    | 10087    | play     | podcasts | play my favorite podcast |
| 1      | 1064     | weather  | query  | tell me the weather in [place_name: barcelona] in [time: two days from now] |
| 92     | 1483     | weather  | query  | what is the current [weather_descriptor: temperature] outside |
| 284    | 10448    | email    | sendemail | send an email to [person: sarah] about [event_name: brunch] [date: today] |
| 649    | 264      | takeaway | order  | please order some [food_type: sushi] for [meal_type: dinner] |
| 38     | 2045     | takeaway | query  | search if the [business_type: restaurant] does [order_type: take out] |

Table 4: Data annotation example snippet

Listing 1: Rasa train data example snippet

```python
1
2    "rasa_nlu_data": {
3        "common_examples": [ {
4            "text": "lower the lights in the bedroom",
5            "intent": "iot_hue_lightdim",
6            "entities": [ {
7                  "start": 24,
8                  "end": 31,
9                  "value": "bedroom",
10                 "entity": "house_place"
11                  } ]
12        } ],
13    }
14    "text": "dim the lights in my bedroom",
15    "intent": "iot_hue_lightdim",
16    "entities": [ {
17                  "start": 21,
18                  "end": 28,
19                  "value": "bedroom",
20                 "entity": "house_place"
21                  } ]
```
Listing 2: LUIS train data example snippet

```json
{
  "intents": [
    {
      "name": "play_podcasts"
    },
    {
      "name": "music_query"
    },
    ....
  ],
  "entities": [
    {
      "name": "Hier2",
      "children": [
        "business_type", "event_name", "place_name",
        "time", "timeofday"
      ]
    },
    ....
  ],
  "utterances": [
    {
      "text": "call a taxi for me",
      "intent": "transport_taxi",
      "entities": [
        {
          "startPos": 7,
          "endPos": 10,
          "value": "taxi",
          "entity": "Hier9::transport_type"
        }]
    },
    ....
  ]
}
```

Listing 3: Watson train data example snippet

```plaintext
--- Watson Entity list ---
joke_type, nice joke_type, funny joke_type, sarcastic
relation, mum relation, dad person, ted
person, emma person, bina person, daniel bell
--- Watson utterance and Intent list ---
```
give me the weather for merced at three pm, weather_query
weather this week,weather_query
find weather report,weather_query
should i wear a hat today,weather_query
what should i wear is it cold outside,weather_query
is it going to snow tonight,weather_query

Listing 4: Dialogflow train data example snippet

```json
--- Dialogflow Entity list ---
{
  "id": "... ...",
  "name": "artist_name",
  "isOverridable": true,
  "entries": [ {
    "value": "aaron carter",
    "synonyms": [ "aaron carter" ],
    "value": "adele",
    "synonyms": [ "adele" ]
  } ],
  "isEnum": false,
  "automatedExpansion": true
}

--- Dialogflow "alarm_query" Intent annotation ---
{
  "userSays": [ {
    "id": "... ...",
    "data": [ {
      "text": "checkout ",
      "text": "today",
      "alias": "date",
      "meta": "@date",
      "userDefined": true
    }, {
      "text": " alarm of meeting"
    } ],
    "isTemplate": false,
    "count": 0
  } ]
}
```
| Intent                  | Total | Tran | Test | Intent                  | Total | Tran | Test | Intent                  | Total | Tran | Test | Intent                  | Total | Tran | Test |
|-------------------------|-------|------|------|-------------------------|-------|------|------|-------------------------|-------|------|------|-------------------------|-------|------|------|
| alarm_query             | 194   | 175  | 19   | general_negate          | 194   | 175  | 19   | play_music              | 194   | 175  | 19   |
| alarm_remove            | 117   | 106  | 11   | general_promise         | 194   | 175  | 19   | play_podcasts           | 194   | 175  | 19   |
| alarm_set               | 194   | 175  | 19   | general_gunky           | 194   | 175  | 19   | play_radio              | 194   | 175  | 19   |
| audio_volume_lown        | 80    | 72   | 8    | general_repeat          | 194   | 175  | 19   | play_currency           | 194   | 175  | 19   |
| audio_volume_medium      | 537   | 482  | 15   | general_tuning          | 1671  | 181  | 16   | qa_definition           | 194   | 175  | 19   |
| audio_volume_high        | 137   | 126  | 13   | general_finance         | 194   | 175  | 19   | qa_job                  | 194   | 175  | 19   |
| calendar_query           | 194   | 175  | 19   | iso_light_changes       | 194   | 175  | 19   | qa_math                | 148   | 134  | 14   |
| calendar_remove          | 194   | 175  | 19   | iso_light_dim           | 126   | 114  | 12   | qa_stock               | 194   | 175  | 19   |
| calendar_set             | 194   | 175  | 19   | iso_light_off           | 194   | 175  | 19   | qa_events              | 194   | 175  | 19   |
| cooking_recipe           | 194   | 175  | 19   | iso_light_ton           | 38    | 35   | 3    | re_locations           | 194   | 175  | 19   |
| date_time_convert        | 87    | 79   | 8    | iso_light_tonup         | 140   | 126  | 14   | req_movies             | 107   | 97   | 10   |
| date_time_query          | 194   | 175  | 19   | iso_wemo_on             | 98    | 89   | 9    | social_post             | 194   | 175  | 19   |
| email_add_contact        | 87    | 79   | 8    | iso_wemo_off            | 76    | 69   | 7    | social_query            | 183   | 165  | 18   |
| email_query              | 194   | 175  | 19   | less_recent_aad          | 194   | 175  | 19   | takeaway_order          | 194   | 175  | 19   |
| email_query_context      | 194   | 175  | 19   | less_query              | 194   | 175  | 19   | takeaway_order          | 194   | 175  | 19   |
| email_sendemail          | 194   | 175  | 19   | list_query              | 194   | 175  | 19   | transport_query         | 194   | 175  | 19   |
| general_affirm           | 194   | 175  | 19   | music_likeness          | 180   | 162  | 18   | transport_jaxi          | 181   | 163  | 18   |
| general_command_stop     | 194   | 175  | 19   | music_query             | 194   | 175  | 19   | transport_ticket        | 194   | 175  | 19   |
| general_contact          | 194   | 175  | 19   | music_settings          | 77    | 60   | 7    | transport_traffic       | 190   | 171  | 19   |
| general_healthcare       | 194   | 175  | 19   | new_query               | 194   | 175  | 19   | weather_query           | 194   | 175  | 19   |
| general_explain          | 194   | 175  | 19   | playaudiobook           | 194   | 175  | 19   |                       |       |      |     |
| general_lolke            | 122   | 110  | 12   | play_game               | 194   | 175  | 19   |

| Entity                  | Transet | Testset | Entity                  | Transet | Testset | Entity                  | Transet | Testset |
|-------------------------|---------|---------|-------------------------|---------|---------|-------------------------|---------|---------|
| alarm_type              | 14      | 0       | event_type              | 352     | 48      | person                  | 468     | 42      |
| app_name                | 32      | 5       | food_type               | 302     | 35      | personal_info           | 100     | 14      |
| artist_name             | 91      | 11      | game_name               | 133     | 17      | place_name              | 869     | 95      |
| audio_book_audio        | 40      | 1       | game_type               | 1       | 0       | play_setting            | 180     | 19      |
| audio_book_name         | 97      | 10      | general_frequency       | 27      | 5       | playlist_name           | 22       | 1       |
| business_game           | 199     | 19      | ingredient              | 17      | 4       | podcast_name            | 44       | 2       |
| business_type           | 199     | 19      | ingredient              | 17      | 4       | podcast_name            | 44       | 2       |
| change_amount           | 57      | 9       | joke_type               | 59      | 4       | radio_name              | 99       | 12      |
| clock                   | 31      | 3       | list_type               | 231     | 13      | relation                | 113      | 13      |
| color_type              | 135     | 11      | meal_type               | 37      | 0       | some_name               | 51       | 9       |
| cooking_type            | 10      | 0       | media_type              | 370     | 40      | time                    | 511      | 62      |
| currency_name           | 296     | 35      | move_game               | 18      | 0       | time_zone               | 59       | 7       |
| currency_name_data      | 905     | 85      | move_game               | 13      | 0       | time_zone               | 130      | 26      |
| definition_word         | 155     | 10      | music_album             | 5       | 10      | transport_agency        | 19       | 10      |
| device_type             | 33      | 41      | music_descriptor        | 17      | 2       | transport_descriptor    | 11       | 0       |
| drink_type              | 6       | 0       | music_genre             | 73      | 8       | transport_name          | 10       | 2       |
| email_address           | 38      | 5       | newscrop                | 75      | 9       | transport_type          | 363      | 35      |
| email_folder            | 17      | 1       | order_type              | 151     | 17      | weather_descriptor      | 95       | 14      |

Table 5: Data Distribution for Intents in Fold 1

Table 6: Data Distribution for Entities in Fold 1
| Intent                  | Rasa | Dialogflow | LUIS | Watson |
|------------------------|------|------------|------|--------|
| alarm                  | 11   | 11         | 16   | 15     |
| alarm_remove           | 11   | 11         | 16   | 15     |
| alarm_set              | 11   | 11         | 16   | 15     |
| audio_volume_down      | 11   | 11         | 16   | 15     |
| audio_volume_mute      | 11   | 11         | 16   | 15     |
| calendar_query         | 11   | 11         | 16   | 15     |
| calendar_remove        | 11   | 11         | 16   | 15     |
| calendar_set           | 11   | 11         | 16   | 15     |
| cooking_recipe         | 11   | 11         | 16   | 15     |
| date_time_offset       | 11   | 11         | 16   | 15     |
| email_old_contact      | 11   | 11         | 16   | 15     |
| email_query            | 11   | 11         | 16   | 15     |
| email_reply            | 11   | 11         | 16   | 15     |
| general_firm           | 11   | 11         | 16   | 15     |
| general_command_stop   | 11   | 11         | 16   | 15     |
| general_confirm        | 11   | 11         | 16   | 15     |
| general_donate         | 11   | 11         | 16   | 15     |
| general_explain        | 11   | 11         | 16   | 15     |
| general_joke           | 11   | 11         | 16   | 15     |
| general_pager          | 11   | 11         | 16   | 15     |
| general_pan          | 11   | 11         | 16   | 15     |
| general_repeat         | 11   | 11         | 16   | 15     |
| get_location           | 11   | 11         | 16   | 15     |
| get_ticket             | 11   | 11         | 16   | 15     |
| get_traffic            | 11   | 11         | 16   | 15     |
| get_news               | 11   | 11         | 16   | 15     |
| get_music              | 11   | 11         | 16   | 15     |
| music_query            | 11   | 11         | 16   | 15     |
| music_reply            | 11   | 11         | 16   | 15     |
| music_remind           | 11   | 11         | 16   | 15     |
| news_feed              | 11   | 11         | 16   | 15     |
| news_story             | 11   | 11         | 16   | 15     |
| play_and_speak         | 11   | 11         | 16   | 15     |
| play_game              | 11   | 11         | 16   | 15     |
| play_music             | 11   | 11         | 16   | 15     |
| play_podcasts          | 11   | 11         | 16   | 15     |
| play_radio             | 11   | 11         | 16   | 15     |
| play_video             | 11   | 11         | 16   | 15     |
| play_weather           | 11   | 11         | 16   | 15     |
| rec_and                  | 11   | 11         | 16   | 15     |
| recommendation_events   | 11   | 11         | 16   | 15     |
| recommendation_gateway  | 11   | 11         | 16   | 15     |
| recommendation_movies   | 11   | 11         | 16   | 15     |
| social_media           | 11   | 11         | 16   | 15     |
| social_story           | 11   | 11         | 16   | 15     |
| takeaway_order         | 11   | 11         | 16   | 15     |
| takeout_order          | 11   | 11         | 16   | 15     |
| takeout_quary          | 11   | 11         | 16   | 15     |
| transport_quary        | 11   | 11         | 16   | 15     |
| transport_taxi         | 11   | 11         | 16   | 15     |
| transport_ticket       | 11   | 11         | 16   | 15     |
| transport_weather      | 11   | 11         | 16   | 15     |
| weather_quary          | 11   | 11         | 16   | 15     |

Table 7: Confusion Matrix summary for Intents in Fold_1.
| Entity   | Data | Dialogflow | LUIS  | Watson |
|----------|------|------------|-------|--------|
| app_name | 2    | 1          | 1     | 1      |
| artist_name | 0    | 0          | 1     | 0      |
| audiobook_author | 0    | 0          | 1     | 0      |
| audiobook_name | 2    | 2          | 1     | 1      |
| business_name | 27   | 27         | 28    | 28     |
| business_type | 10   | 8          | 10    | 10     |
| change_amount | 0    | 0          | 1     | 0      |
| coffee_type | 1    | 1          | 1     | 1      |
| color_type | 8    | 3          | 1     | 1      |
| currency_name | 3    | 0          | 3     | 0      |
| date      | 9    | 8          | 9     | 9      |
| definition_word | 1    | 1          | 1     | 1      |
| device_type | 33   | 8          | 10    | 10     |
| email_address | 4    | 0          | 1     | 0      |
| email_folder | 1    | 0          | 1     | 0      |
| event_name | 22   | 23         | 23    | 23     |
| food_type | 11   | 12         | 13    | 13     |
| game_name | 7    | 8          | 7     | 7      |
| general_frequency | 1    | 1          | 1     | 1      |
| house_place | 24   | 23         | 24    | 24     |
| ingredient | 0    | 0          | 0     | 0      |
| joke_type | 3    | 3          | 3     | 3      |
| list_name | 9    | 4          | 10    | 10     |
| media_type | 27   | 24         | 27    | 27     |
| message_descriptor | 4    | 2          | 4     | 4      |
| music_genre | 6    | 2          | 6     | 6      |
| news_topic | 0    | 3          | 3     | 3      |
| order_type | 17   | 3          | 17    | 17     |
| person | 31   | 12         | 31    | 31     |
| personal_info | 2    | 2          | 2     | 2      |
| place | 5    | 5          | 5     | 5      |
| playing | 13   | 2          | 15    | 15     |
| playList_name | 0    | 0          | 0     | 0      |
| podcast_descriptor | 5    | 1          | 5     | 5      |
| poddata_name | 0    | 0          | 0     | 0      |
| radio_name | 4    | 2          | 4     | 4      |
| relation | 8    | 8          | 8     | 8      |
| song_name | 4    | 1          | 4     | 4      |
| time | 27   | 27         | 25    | 25     |
| time_zone | 23   | 23         | 24    | 24     |
| transport_agency | 10   | 0          | 10    | 10     |
| transport_name | 0    | 2          | 0     | 2      |
| transport_type | 3    | 4          | 3     | 4      |
| weather_descriptor | 5    | 5          | 5     | 5      |

Table 8: Confusion Matrix summary for Entities in Fold1