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 25 K user utterances that we have collected and annotated with Intent and Entity Type specifications and which will be released as part of this submission (https://github.com/xliuhw/NLU-Evaluation-Data). 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 (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 threfore like to stress that this paper does *not* include an evaluation of this feature in Watson NLU.). Again, Dialogflow, LUIS and Rasa perform well on this task.

(Work done when Pawel was with Emotech North LTD).

X. Liu (✉) · A. Eshghi · V. Rieser
Heriot-Watt University, Edinburgh EH14 4AS, UK
e-mail: x.liu@hw.ac.uk

A. Eshghi
e-mail: a.eshghi@hw.ac.uk

V. Rieser
e-mail: v.t.rieser@hw.ac.uk

P. Swietojanski
The University of New South Wales, Sydney, Australia
e-mail: p.swietojanski@unsw.edu.au

E. Marchi et al. (eds.), *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, Lecture Notes in Electrical Engineering 714,
https://doi.org/10.1007/978-981-15-9323-9_15

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.

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[1] NLU services, namely: Rasa,[2] Watson,[3] LUIS[4] and Dialogflow.[5] The evaluation uses a new dataset of 25 k 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, 5], 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.

[1] According to anecdotal evidence from academic and start-up communities.

[2] https://rasa.com/.

[3] https://www.ibm.com/watson/ai-assistant/.

[4] https://www.luis.ai/home.

[5] https://dialogflow.com/.

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.[6] 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[7] 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.

Table 1 Input requirements and output of NLU services

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

[6]Note that, one could develop one's own system using existing libraries, e.g. sk_learn libraries http://scikit-learn.org/stable/, spaCy https://spacy.io/, but a quicker and more accessible way is to use an existing service platform.

[7]Was not yet open source when we were doing the benchmarking, and was later on also introduced in https://arxiv.org/abs/1805.10190.

(1) All four platforms support Intent classification and NER; (2) None of them support Multiple Intents where a single utterance might express more than one Intent, i.e. is performing more than one action. This is potentially a significant limitation because such utterances are generally very common in spoken dialogue; (3) 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.

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 and 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 8 K 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 and 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.[8]

5.2 System Versions and 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`.[9] 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 and 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[10] 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. Tables 7 and 8 in the Appendix show the detailed confusion matrices used to calculate the scores of Precision, Recall and F1 for Intents and Entities.

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

[8] 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.

[9] 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. Wed like to stress that this paper does *not* include an evaluation of this feature in Watson NLU.

[10] Micro-average sums up the individual TP, FP, and FN of all Intent/Entity classes to compute the average metric.

Table 2 Overall scores for intent and entity

	Intent			Entity		
	Prec	Rec	F1	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**	0.354	0.787	**0.488**

Table 3 Combined overall scores

	Prec	Rec	F1
Rasa	0.862	0.787	0.822
Dialogflow	0.832	0.791	0.811
LUIS	0.848	0.796	0.821
Watson	0.540	0.838	**0.657**

of False Positives.[11] 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—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 Tables 5 and 6) for fold-1[12] 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, posdcasts_name or game_name, e.g. "play space invaders please" which could be annotated the entity 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.

[11] Interestingly, Watson only requires a list of possible entities rather than entity annotation in utterances as other platforms do (See Table 1).

[12] Tables for other folds are omitted for space reason.

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.

Appendix

We provide some examples of the data annotation and the training inputs to each of the 4 platforms in Table 4, Listings 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. Tables 7 and 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.

Listing 1 Rasa train data example snippet

```
{
  "rasa_nlu_data": {
    "common_examples": [   {
        "text": "lower the lights in the
           bedroom",
        "intent": "iot_hue_lightdim",
        "entities": [    {
            "start": 24,
            "end": 31,
            "value": "bedroom",
            "entity": "house_place"
```

```
            } ]   },
        {
          "text": "dim the lights in my bedroom"
             ,
          "intent": "iot_hue_lightdim",
          "entities": [   {
              "start": 21,
              "end": 28,
              "value": "bedroom",
              "entity": "house_place"
            } ]    },
         ...  ...
        ]
}
```

Table 4 Data annotation example snippet

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]
653	17462	Calendar	Query	What is up in my schedule [date: today]
2	564	Calendar	Remove	Please cancel all my events for [date: today]
2	586	Play	Music	I'd like to hear [artist_name: queen's] [song_name: barcelona]
65	2813	Play	Radio	Play a [radio_name: pop station] on the radio
740	19087	Play	Podcasts	Play my favorite podcast
1	1964	Weather	Query	Tell me the weather in [place_name: barcelona] in [time: two days from now]
92	3483	Weather	Query	What is the current [weather_descriptor: temperature] outside
394	10448	Email	Sendemail	Send an email to [person: sarah] about [event_name: brunch] [date: today]
4	649	Email	Query	Has the [business_name: university of greenwich] emailed me
2	624	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 5 Data distribution for intents in Fold_1

Intent	Total	Train	Test	Intent	Total	Train	Test	Intent	Total	Train	Test
Alarm_query	194	175	19	General_negate	194	175	19	Play_music	194	175	1
Alarm_remove	117	106	11	General_praise	194	175	19	Play_podcasts	194	175	1
Alarm_set	194	175	19	General_quirky	194	175	19	Play_radio	194	175	19
Audio_volume_down	80	72	8	General_repeat	194	175	19	qa_currency	194	175	19
Audio_volume_mute	157	142	15	iot_cleaning	167	151	16	qa_definition	194	175	19
Audio_volume_up	139	126	13	iot_coffee	194	175	19	qa_factoid	194	175	19
Calendar_query	194	175	19	iot_hue_lightchange	194	175	19	qa_maths	148	134	14
Calendar_remove	194	175	19	iot_hue_lightdim	126	114	12	qa_stock	194	175	19
Calendar_set	194	175	19	iot_hue_lightoff	194	175	19	rec_events	194	175	19
Cooking_recipe	194	175	19	iot_hue_lighton	38	35	3	rec_locations	194	175	19
Datetime_convert	87	79	8	iot_hue_lightup	140	126	14	rec_movies	107	97	10
Datetime_query	194	175	19	iot_wemo_off	98	89	9	Social_post	194	175	19
Email_addcontact	87	79	8	iot_wemo_on	76	69	7	Social_query	183	165	18
Email_query	194	175	19	Lists_createoradd	194	175	19	Takeaway_order	194	175	19
Email_querycontact	194	175	19	Lists_query	194	175	19	Takeaway_query	194	175	19
Email_sendemail	194	175	19	Lists_remove	194	175	19	Transport_query	194	175	19
General_affirm	194	175	19	Music_likeness	180	162	18	Transport_taxi	181	163	18
General_commandstop	194	175	19	Music_query	194	175	19	Transport_ticket	194	175	19
General_confirm	194	175	19	Music_settings	77	70	7	Transport_traffic	190	171	19
General_dontcare	194	175	19	News_query	194	175	19	Weather_query	194	175	19
General_explain	194	175	19	Play_audiobook	194	175	19				
General_joke	122	110	12	Play_game	194	175	19				

Table 6 Data distribution for entities in Fold_1

Entity	Trainset	Testset	Entity	Trainset	Testset	Entity	Trainset	Testset
Alarm_type	14	0	Event_name	352	48	Person	468	42
App_name	32	5	Food_type	302	25	Personal_info	100	14
Artist_name	91	11	Game_name	133	17	Place_name	869	95
Audiobook_author	10	1	Game_type	1	0	Player_setting	190	19
Audiobook_name	97	10	General_frequency	27	5	Playlist_name	22	1
Business_name	394	41	House_place	259	25	Podcast_descriptor	67	6
Business_type	199	19	Ingredient	17	4	Podcast_name	44	2
Change_amount	57	9	Joke_type	59	4	Radio_name	99	12
Coffee_type	31	4	List_name	211	13	Relation	127	13
Color_type	135	11	Meal_type	37	0	Song_name	51	9
Cooking_type	10	0	Media_type	370	40	Time	511	62
Currency_name	296	35	Movie_name	18	0	Time_zone	59	7
Date	905	85	Movie_type	13	0	Timeofday	150	26
Definition_word	158	16	Music_album	1	0	Transport_agency	59	10
Device_type	353	41	Music_descriptor	17	2	Transport_descriptor	11	0
Drink_type	6	0	Music_genre	72	8	Transport_name	10	2
Email_address	38	5	News_topic	75	9	Transport_type	363	35
Email_folder	17	1	Order_type	151	17	Weather_descriptor	95	14

Table 7 Confusion matrix summary for intents in Fold_1

	Rasa				Dialogflow				LUIS				Watson			
Intent	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN
Alarm_query	17	1	2	1056	19	0	0	1057	18	2	1	1055	19	0	0	1057
Alarm_remove	11	0	0	1065	10	2	1	1063	9	0	2	1065	11	0	0	1065
Alarm_set	18	3	1	1054	17	4	2	1053	17	3	2	1054	17	3	2	1054
Audio_volume_down	7	1	1	1067	8	0	0	1068	7	0	1	1068	8	0	0	1068
Audio_volume_mute	13	1	2	1060	14	0	1	1061	12	1	3	1060	14	1	1	1060
Audio_volume_up	12	3	1	1060	13	0	0	1063	12	3	1	1060	12	3	1	1060
Calendar_query	11	10	8	1047	13	18	6	1039	11	6	8	1051	10	8	9	1049
Calendar_remove	17	0	2	1057	18	1	1	1056	18	2	1	1055	19	1	0	1056
Calendar_set	16	2	3	1055	14	2	5	1055	14	4	5	1053	16	3	3	1054
Cooking_recipe	15	1	4	1056	11	2	8	1055	13	4	6	1053	15	1	4	1056
Datetime_convert	5	2	3	1066	7	4	1	1064	7	2	1	1066	8	2	0	1066
Datetime_query	17	4	2	1053	18	9	1	1048	17	4	2	1053	18	4	1	1053
Email_addcontact	8	3	0	1065	8	0	0	1068	8	0	0	1068	8	2	0	1066
Email_query	17	1	2	1056	18	1	1	1056	15	3	4	1054	17	2	2	1055
Email_querycontact	11	4	8	1053	13	3	6	1054	14	4	5	1053	14	3	5	1054
Email_sendemail	17	1	2	1056	16	1	3	1056	16	4	3	1053	17	2	2	1055

(continued)

Table 7 (continued)

	Rasa				Dialogflow				LUIS				Watson			
General_affirm	19	1	0	1056	19	0	0	1057	19	0	0	1057	19	1	0	1056
General_commandstop	19	0	0	1057	18	1	1	1056	19	0	0	1057	19	1	0	1056
General_confirm	19	1	0	1056	19	0	0	1057	19	0	0	1057	19	0	0	1057
General_dontcare	19	0	0	1057	19	1	0	1056	18	1	1	1056	19	2	0	1055
General_explain	19	1	0	1056	19	0	0	1057	18	0	1	1057	19	2	0	1055
General_joke	11	0	1	1064	12	0	0	1064	12	0	0	1064	12	0	0	1064
General_negate	18	0	1	1057	19	0	0	1057	19	1	0	1056	19	0	0	1057
General_praise	18	1	1	1056	19	0	0	1057	19	1	0	1056	18	1	1	1056
General_quirky	11	22	8	1035	4	2	15	1055	8	16	11	1041	7	9	12	1048
General_repeat	19	0	0	1057	19	1	0	1056	19	0	0	1057	19	0	0	1057
iot_cleaning	14	1	2	1059	13	6	3	1054	16	1	0	1059	16	1	0	1059
iot_coffee	18	3	1	1054	18	1	1	1056	18	0	1	1057	19	1	0	1056
iot_hue_lightchange	15	1	4	1056	14	3	5	1054	15	4	4	1053	13	3	6	1054
iot_hue_lightdim	12	0	0	1064	11	0	1	1064	10	1	2	1063	11	1	1	1063
iot_hue_lightoff	17	2	2	1055	15	0	4	1057	17	1	2	1056	17	2	2	1055
iot_hue_lighton	3	3	0	1070	3	3	0	1070	2	3	1	1070	3	3	0	1070
iot_hue_lightup	9	1	5	1061	11	1	3	1061	11	0	3	1062	11	2	3	1060
iot_wemo_off	9	2	0	1065	8	4	1	1063	9	4	0	1063	9	2	0	1065
iot_wemo_on	5	2	2	1067	5	1	2	1068	4	3	3	1066	6	1	1	1068
Lists_createoradd	16	2	3	1055	16	6	3	1051	16	5	3	1052	18	3	1	1054
Lists_query	16	3	3	1054	16	5	3	1052	16	3	3	1054	14	2	5	1055
Lists_remove	17	1	2	1056	18	3	1	1054	18	2	1	1055	18	0	1	1057

(continued)

Table 7 (continued)

	Rasa				Dialogflow				LUIS				Watson			
Music_likeness	12	4	6	1054	13	5	5	1053	13	3	5	1055	14	1	4	1057
Music_query	13	0	6	1057	11	3	8	1054	10	4	9	1053	11	2	8	1055
Music_settings	6	2	1	1067	4	2	3	1067	7	0	0	1069	7	2	0	1067
News_query	13	9	6	1048	10	4	9	1053	13	3	6	1054	14	1	5	1056
Play_audiobook	16	3	3	1054	13	8	6	1049	17	1	2	1056	16	2	3	1055
Play_game	15	5	4	1052	13	2	6	1055	13	2	6	1055	13	2	6	1055
Play_music	13	4	6	1053	16	5	3	1052	12	11	7	1046	12	14	7	1043
Play_podcasts	17	0	2	1057	14	1	5	1056	16	0	3	1057	17	1	2	1056
Play_radio	15	1	4	1056	15	2	4	1055	17	1	2	1056	15	2	4	1055
qa_currency	17	1	2	1056	16	0	3	1057	18	0	1	1057	18	0	1	1057
qa_definition	19	0	0	1057	13	2	6	1055	18	0	1	1057	18	1	1	1056
qa_factoid	10	13	9	1044	7	9	12	1048	15	15	4	1042	14	8	5	1049
qa_maths	14	2	0	1060	12	2	2	1060	13	4	1	1058	14	1	0	1061
qa_stock	19	2	0	1055	19	1	0	1056	19	0	0	1057	19	1	0	1056
Recommendation_events	13	2	6	1055	14	6	5	1051	16	3	3	1054	15	2	4	1055
Recommendation_locations	16	1	3	1056	15	1	4	1056	17	2	2	1055	16	1	3	1056
Recommendation_movies	8	2	2	1064	8	2	2	1064	9	1	1	1065	10	2	0	1064
Social_post	18	3	1	1054	17	4	2	1053	18	1	1	1056	19	1	0	1056
Social_query	16	5	2	1053	14	8	4	1050	17	3	1	1055	17	3	1	1055
Takeaway_order	12	0	7	1057	16	2	3	1055	16	4	3	1053	16	1	3	1056
Takeaway_query	18	6	1	1051	19	3	0	1054	16	2	3	1055	18	3	1	1054
Transport_query	16	3	3	1054	17	3	2	1054	13	3	6	1054	14	5	5	1052
Transport_taxi	17	2	1	1056	17	1	1	1057	18	0	0	1058	18	1	0	1057
Transport_ticket	16	1	3	1056	17	0	2	1057	16	1	3	1056	16	2	3	1055
Transport_traffic	18	1	1	1056	18	1	1	1056	18	1	1	1056	19	2	0	1055
Weather_query	16	2	3	1055	12	2	7	1055	13	5	6	1052	13	2	6	1055

Table 8 Confusion matrix summary for entities in Fold_1

	Rasa				Dialogflow				LUIS				Watson			
Entity	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN
App_name	3	0	2	1071	2	1	3	1070	3	0	2	1071	4	10	1	1061
Artist_name	3	0	8	1065	5	1	6	1064	4	2	7	1063	3	1	8	1064
Audiobook_author	0	0	1	1075	0	0	1	1075	0	0	1	1075	0	0	1	1075
Audiobook_name	2	3	8	1063	6	2	4	1064	5	1	5	1065	6	3	4	1063
Business_name	25	12	16	1027	32	8	9	1029	32	5	9	1031	29	30	12	1008
Business_type	15	2	4	1055	13	1	6	1056	14	5	5	1054	16	45	3	1014
Change_amount	7	0	2	1067	6	2	3	1065	8	2	1	1065	6	12	3	1056
Coffee_type	1	0	3	1072	2	1	2	1071	2	0	2	1072	2	4	2	1068
Color_type	8	2	3	1063	8	1	3	1064	8	1	3	1064	9	26	2	1042
Currency_name	25	0	10	1058	14	0	21	1058	28	4	7	1056	31	12	4	1049
Date	77	8	8	983	74	25	11	969	78	9	7	984	80	30	5	971
Definition_word	7	2	9	1058	10	3	6	1057	11	4	5	1056	6	104	10	961
Device_type	33	0	8	1035	24	10	17	1027	33	6	8	1029	38	76	3	963
Email_address	4	0	1	1071	4	1	1	1070	3	2	2	1071	1	0	4	1071
Email_folder	1	0	0	1075	1	0	0	1075	1	0	0	1075	1	0	0	1075
Event_name	27	4	21	1024	25	25	23	1005	24	6	24	1023	30	56	18	973
Food_type	13	3	12	1048	16	5	9	1046	16	4	9	1047	17	16	8	1040
Game_name	7	2	10	1057	11	2	6	1057	12	0	5	1059	9	2	8	1057
General_frequency	1	1	4	1070	0	0	5	1071	2	0	3	1071	3	3	2	1069
House_place	22	1	3	1050	22	10	3	1042	24	1	1	1050	25	18	0	1033
Ingredient	0	0	4	1072	1	0	3	1072	0	1	4	1072	1	3	3	1069

(continued)

Table 8 (continued)

	Rasa				Dialogflow				LUIS				Watson			
Joke_type	3	1	1	1071	3	0	1	1072	3	2	1	1070	2	53	2	1019
List_name	9	7	4	1056	6	2	7	1061	10	5	3	1058	7	56	6	1010
Media_type	29	4	11	1033	26	24	14	1013	31	11	9	1026	34	81	6	961
Music_descriptor	0	0	2	1074	0	0	2	1074	0	0	2	1074	0	4	2	1070
Music_genre	6	1	2	1067	7	2	1	1066	6	1	2	1067	7	8	1	1060
News_topic	0	2	9	1065	3	3	6	1064	2	4	7	1063	3	18	6	1049
Order_type	14	3	3	1056	12	3	5	1056	13	2	4	1057	17	8	0	1051
Person	31	14	11	1021	31	12	11	1023	30	7	12	1028	27	36	15	999
Personal_info	5	0	9	1063	5	1	9	1062	7	4	7	1059	12	58	2	1011
Place_name	65	22	30	971	66	17	29	976	71	5	24	986	76	39	19	961
Player_setting	13	2	6	1056	9	3	10	1055	16	7	3	1052	18	71	1	988
Playlist_name	0	0	1	1075	0	0	1	1075	0	0	1	1075	0	0	1	1075
Podcast_descriptor	5	1	1	1069	4	1	2	1069	5	2	1	1068	5	9	1	1061
Podcast_name	0	0	2	1074	0	0	2	1074	1	2	1	1072	0	111	2	968
Radio_name	4	2	8	1063	6	2	6	1062	7	5	5	1060	2	17	10	1048
Relation	8	0	5	1063	6	4	7	1059	7	1	6	1063	10	4	3	1059
Song_name	4	1	5	1066	5	2	4	1065	3	1	6	1066	3	13	6	1055
Time	53	3	9	1013	45	18	17	1002	49	12	13	1010	55	119	7	928
Time_zone	2	0	5	1071	3	1	4	1070	2	1	5	1070	6	63	1	1019
Timeofday	23	3	3	1047	13	3	13	1047	22	4	4	1047	26	4	0	1046
Transport_agency	10	0	0	1066	10	0	0	1066	10	0	0	1066	10	0	0	1066
Transport_name	0	0	2	1074	0	0	2	1074	0	0	2	1074	0	0	2	1074
Transport_type	35	1	0	1040	14	1	21	1041	34	4	1	1039	35	7	0	1035
Weather_descriptor	5	1	9	1063	7	3	7	1061	7	2	7	1062	8	12	6	1053

Listing 2 LUIS train data example snippet

```
{
   "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

```
---- 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

```
---- 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
    },
   ... ...
] }
```

References

1. Braun D, Mendez AH, Matthes F, Langen M (2017) Evaluating natural language understanding services for conversational question answering systems. In: Proceedings of SIGDIAL 2017, pp 174–185
2. Canonico M, Russis LD (2018) A comparison and critique of natural language understanding tools. In: Proceedings of CLOUD COMPUTING 2018
3. Coucke A, Ball A, Delpuech C, Doumouro C, Raybaud S, Gisselbrecht T, Dureau J (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-2b8ddcf9fb19
4. Canh NT (2018) Benchmarking intent classification services. June 2018. https://medium.com/botfuel/benchmarking-intent-classification-services-june-2018-eb8684a1e55f
5. Wisniewski C, Delpuech C, Leroy D, Pivan F, Dureau J (2017) Benchmarking natural language understanding systems. https://snips.ai/content/sdk-benchmark-visualisation/