Evaluating N-best Calibration of Natural Language Understanding for Dialogue Systems

Ranim Khojah and **Alexander Berman**
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
guskhojra@student.gu.se,
alexander.berman@gu.se

Staffan Larsson
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
and Talkamatic AB
staffan@talkamatic.se

Abstract

A Natural Language Understanding (NLU) component can be used in a dialogue system to perform intent classification, returning an N-best list of hypotheses with corresponding confidence estimates. We perform an in-depth evaluation of 5 NLUs, focusing on confidence estimation. We measure and visualize calibration for the 10 best hypotheses on model level and rank level, and also measure classification performance. The results indicate a trade-off between calibration and performance. In particular, Rasa (with Sklearn classifier) had the best calibration but the lowest performance scores, while Watson Assistant had the best performance but a poor calibration.

1 Introduction

Natural Language Understanding (NLU) is an important component in dialogue systems. One of the typical tasks of NLU is intent classification: given a user utterance, the NLU returns a list of N hypotheses (an N-best list) ranked according to confidence estimates (a real number between 0 and 1). The highest ranking hypothesis is returned by the NLU as the predicted intent. Confidence estimates are also available for lower ranked hypotheses.

In this study, we evaluate confidence estimation in 5 NLU services, namely Watson Assistant, Language Understanding Intelligent Service (LUIS), Snips.ai and Rasa (with two pipelines Rasa-Sklearn and Rasa-DIET). We measure the calibration and the performance of NLUs on rank level (results for a specfic rank) and on model level (aggregated results of all ranks). *Calibration* here refers to the correlation between confidence estimates and accuracies, i.e. how useful the confidence estimate associated with a certain hypothesis is for predicting its accuracy.

To achieve our objectives, we conduct an exploratory case study on the 5 NLUs. We train the NLUs using a subset of a multi-domain dataset proposed by Liu et al. (2021). We measure the calibration of the NLUs on model and rank levels using reliability diagrams and correlation coefficient with respect to instance-level accuracy. We also measure the performance on a model level through accuracy and F1-score.

Our evaluation aims to facilitate NLU service selection and help dialogue system developers adapt their dialogue system to specific NLU services. For example, depending on the degree of calibration in an NLU, contextual or interactive disambiguation (clarification requests) can be an option. If confidence estimates reflect true accuracy, then if two (or more) hypotheses have similar confidence estimates, this may indicate the presence of an ambiguity in the user input (from the perspective of the NLU, i.e., disregarding dialogue context) that needs to be resolved. Conversely, if confidence estimates (especially those for non-top ranks) do not reflect accuracies, then even if the top two (or more) hypotheses have similar estimates, this may not be a reliable indication of ambiguity but rather be due to noise.

Our evaluation scripts are publicly available on GitHub[1] along with the dataset, enabling replication of the study and to ease building on it.

2 Related work

Current NLUs typically use machine-learning on natural-language data (i.e., the user utterances) to extract features (e.g., keywords, word counts and word embeddings) and predict the intent of the user accordingly (Jung, 2019; Shridhar et al., 2019).

NLU services are widely used by dialogue developers and allow them to create and train NLU models for dialogue systems. However, the task of

[1] https://github.com/ranimkhojah/confidence-estimation-benchmark

Proceedings of the SIGdial 2022 Conference, pages 582–594
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

choosing the best NLU service depends on the domain and context of the dialogue system. In prior work, benchmarks and evaluations have been performed to identify the best NLU service in different domains like software engineering (Abdellatif et al., 2021), meteorology (Canonico and De Russis, 2018), question answering (Braun et al., 2017) and others (McTear et al., 2016; Stoyanchev et al., 2016; Kar and Haldar, 2016; Koetter et al., 2019). Generally, these evaluation studies have been conducted to draw the trade-off line between different NLU services in terms of the usability of their user interfaces (Gregori, 2017), technical features (e.g., language and device support) (Koetter et al., 2019) and performance (Braun et al., 2017; Liu et al., 2021).

NLU performance is usually assessed via performance measures (e.g., accuracy, F1-score, etc.) which depend only on the top hypothesis returned by the NLU, and disregarding the associated confidence estimates. For example, an NLU that predicts 3 out of 10 intents incorrectly with high confidence estimates has the same performance as an NLU that predicts 3 out of 10 intents incorrectly with a low confidence estimation.

In earlier work, various methods for visualizing and measuring confidence calibration (the extent to which confidence estimates reflect true likelihoods) have been discussed. For example, Guo et al. (2017) and Vasudevan et al. (2019) visualize calibration of neural network models through reliability diagrams. As for quantitative metrics, one proposed measurement is statistical correlation between confidence estimate and some instance-level performance metric; Dong et al. (2018) use Spearman's correlation with respect to F1 score, while Vasudevan et al. (2019) use Pearson's correlation with respect to instance-level accuracy. A second option is to aggregate across instance-level calibration scores (so called proper scoring rules); examples include Brier score (Brier et al., 1950) and negative log-likelihood (Quinonero-Candela et al., 2005). A third approach involves partitioning confidence estimates into bins, assessing correlation for individual bins, and then aggregating across bin-level calibration results; one popular example of such an approach is Expected Calibration Error (ECE) (Naeini et al., 2015), which has been extended by Nixon et al. (2019) to assess calibration of all predictions rather than only the top one.

In this study, we apply some of the previously proposed calibration assessment methods – namely reliability diagrams and correlation with instance-level performance – to NLUs. In addition, we also measure calibration on rank level, enabling a more fine-grained analysis.

3 Background

When using an NLU, an utterance U is fed to the trained NLU, and the output normally includes the information in the following example:

```
{ 'utterance' : 'U' ,
  'top_intent': 'intent_1',
  'intent_ranking': {
    'intent_1': conf_1, # rank 1
    'intent_2': conf_2, # rank 2
    ... ,
    'intent_N': conf_N  # rank N
  }
}
```

The output of the NLU given an utterance U is a prediction consisting of the user utterance, the top intent and an intent ranking. The intent ranking consists of the N-best intent hypotheses along with their corresponding confidence estimates. The confidence estimates reflect how confident the NLU model is regarding each hypothesis.

Figure 1 illustrates how NLUs are used in dialogue systems, involving a scenario where a user asks a dialogue system a question within the home domain. The user utterance (which can be typed by the user in a chat or captured by a speech recognizer) is sent to an NLU service which performs intent classification on the user utterance and returns a prediction with the top intent and the intent ranking. The results are sent to a dialogue manager that decides how to steer the dialogue based on the output from the NLU and some dialogue policy. In case of a high estimated confidence for the most likely hypothesis, the dialogue manager integrates the user's intent, and information is sent to the natural-language generator that generates a response which is uttered back to the user.

A dialogue system can use confidence estimates as a basis for choosing a grounding strategy (e.g. asking a control question when confidence is low), ambiguity detection and handling (e.g. asking a clarification question if the top-ranked intents have similar confidence estimates) or re-scoring of hypotheses based on contextual information not available to the NLU but to the dialogue manager (such as dialogue state).

Different NLUs may have different ways of computing confidence estimates, possibly reflect-

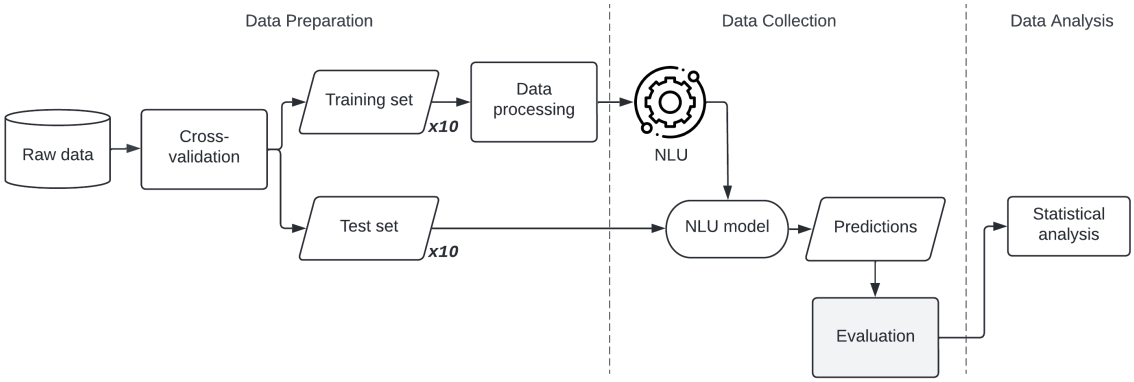


Figure 1: A dialogue system.

ing different notions of confidence. However, for the purpose of using the estimates in a dialogue system, we are interested in how well they reflect true probabilities. In section 4 we note variations in how confidence estimates are computed, but do not take these differences into account in our evaluation.

4 NLU services

NLU services can be used to construct the NLU component in a dialogue system. In this study, we chose NLU services (henceforth NLUs) based on the following criteria: i) can perform intent classification and ii) returns at least 10 top hypotheses in the output. We examine 5 NLUs: Watson Assistant (IBM, 2010), Language Understanding Intelligent Service (LUIS) (Microsoft, 2017), Snips (Snips, 2013), and Rasa (Rasa, 2016) (in two configurations).

Below, we briefly introduce the NLUs. Information about the NLUs, including the tested version, is summarized in table 1.

Watson Assistant Watson Assistant (henceforth Watson) is a cloud-based NLU developed by IBM. When parsing an utterance, Watson returns the top 10 hypotheses along with their confidence estimates. Confidence estimates are calculated independently for each intent that it has been trained on. In addition, Watson has an optional built-in "irrelevant" intent for out-of-scope (OOS) input.

LUIS LUIS (Language Understanding Intelligent Service) is provided by Microsoft and runs on the Azure cloud platform. LUIS trains an intent using provided positive examples and other intents as negative examples.

There is no limit in the number of hypotheses that LUIS returns; in other words, if the NLU is trained on N intents, then the intent ranking is of length N. A "None" intent for out-of-scope input is also supported, but requires the user to train it on example utterances.

Rasa Opensource Rasa is an open-source NLU provided by Rasa Technologies. It can run on different pipelines that are configurable which increases the flexibility of the NLU (Bocklisch et al., 2017). Rasa returns the top 10 hypotheses and their corresponding confidence estimates are normalized (they sum up to 1). Rasa does not offer a built-in out-of-scope intent.

In this study, we use with two different pipelines. The first pipeline uses the Sklearn intent classifier[2] while the second uses Dual Intent and Entity Transformer (DIET) (Bunk et al., 2020). We refer to the two pipelines above as Rasa-Sklearn and Rasa-DIET respectively.

Snips Snips is an AI voice platform for connected devices which provides an NLU for Python called Snips NLU (henceforth Snips). By default, Snips returns all hypotheses of all intents with confidence estimates, in addition to a "None" intent[3] for OOS input.

5 Dataset and data preparation

To conduct intent classification as a part of our evaluation, we build on the dataset proposed by Liu et al. (2021). The authors collect and annotate

[2] https://rasa.com/docs/rasa/components/#sklearnintentclassifier

[3] https://snips-nlu.readthedocs.io/en/latest/tutorial.html#the-none-intent

NLU	Packaging	Classifier Type	Version	OOS intent
Watson	Cloud-based service	Multiple-binary	Invoked in April 2022	Yes
LUIS	Cloud-based service	Multi-class	Invoked in April 2022	Yes
Snips	Open-source framework	Multi-class	v0.20.2	Yes
Rasa	Open-source framework	Multi-class	v2.4.3	No

Table 1: Summary of studied NLUs. (OOS = out of scope.)

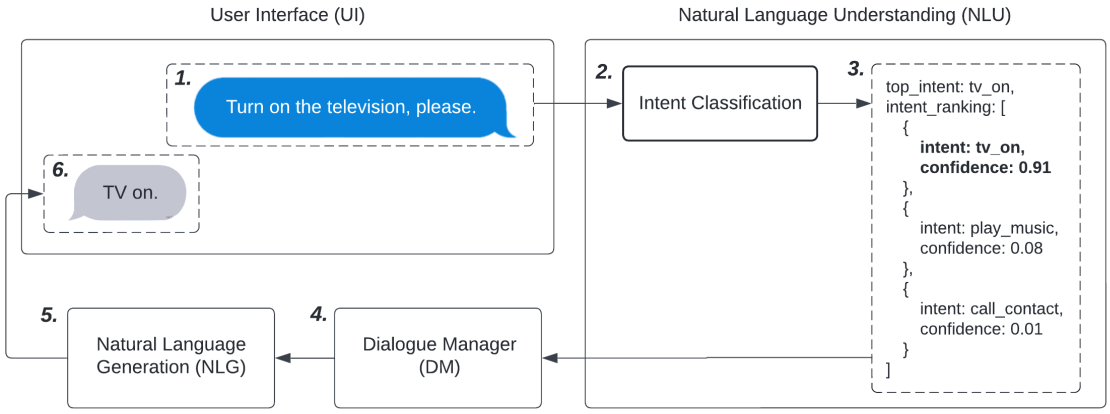


Figure 2: The evaluation process followed for each NLU to obtain the results; this process was repeated 5 times, 1 time per NLU model.

25716 user utterances for human-robot interaction and cover 64 intents, 18 scenarios and 21 domains. From this dataset, we select the 10 intents with the most examples (highest number of instances), yielding a total of 14962 utterances (see table 2).[4] We perform repeated random sub-sampling (Dubitzky et al., 2007) with 10 iterations to generate 10 random datasets; each dataset is divided with a 2:1 ratio into a training and testing sets respectively. (A breakdown by domain and/or scenario could also have been interesting, but was ruled out due to data sparsity.)

When analyzing the outputs from the NLUs, we exclude hypotheses with the OOS ("None"/"irrelevant") class in the intent ranking in order to ensure that all NLUs have the same intent ranking length and make their results comparable. (See section 8 for a discussion about OOS handling.)

6 Evaluation of confidence estimation

An overview of our study's execution is illustrated in figure 2. The evaluation is performed at two levels: rank and model. On rank level, the results are obtained for each rank across the NLUs, whereas on model level, the results of all ranks are aggregated.

The evaluation focuses on the calibration and performance of the NLUs. Calibration is measured using reliability diagrams and Spearman's correlation coefficient with respect to instance-level accuracy. The latter is measured through accuracy and F1-score. Evaluation is conducted for each split and results are averaged across splits.

6.1 Confidence calibration

Confidence calibration is the extent to which a model is able to produce confidence estimates that reflect the accuracy (true likelihood) of the respective intent hypotheses (Guo et al., 2017). For example, in a well-calibrated model, hypotheses with a confidence estimate of 0.7 are correct in 70% of the cases.

Reliability diagrams are visualizations of a model's calibration (Guo et al., 2017). They plot true likelihood (accuracy) of predictions as a function of confidence estimate. Hence, a perfectly-calibrated model is visualized as the identity func-

[4] Liu et al. (2021) provide user utterances in different forms: original (raw), with entity annotations, and normalized. In our study, we use the original user utterances.

Intent	Size	Example
query	5981	what's the time in australia
set	1748	wake me up at 9am on Friday
music	1205	start playing music from favourites
quirky	1088	I am not tired I am actually happy
factoid	1052	tell me comics of charlie chaplin
remove	986	cancel my 7am alarm
negate	939	you don't understand it right
sendemail	694	send a group mail to lookafter
explain	684	could you clarify me on it further
repeat	585	please let's start over
Total	14962 examples	

Table 2: Selected intents for the case study with their respective size (i.e. number of utterances) and one example utterance.

tion, and any deviation indicates miscalibration.

Reliability diagrams are plotted by partitioning predictions into bins, each of which represents a confidence range. In our study, we use 10 uniformly distributed bins, i.e. [0.0-0.1], [0.1-0.2], ...[0.9-1.0]. For each bin, mean confidence estimate and accuracy is calculated and plotted as a point.

Spearman's correlation coefficient In order to numerically measure the degree of calibration, we assess the correlation between confidence estimates (scores in the range 0-1) and instance-level accuracies (1 for correct classifications, 0 for incorrect classifications). More specifically, we measure the extent to which an increase in confidence estimate is associated with an increase in instance-level accuracy – in other words, the monotonicity of the relationship between confidence estimate and accuracy. The degree of monotonicity is measured using Spearman's correlation coefficient (Xiao et al., 2016).[5]

Given two variables (X and Y) of size N ($x_1, x_2, ..x_n$ and $y_1, y_2...y_n$ respectively), Spearman's correlation coefficient (ρ) is calculated through the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n is the number of samples, and d is the pairwise differences of the elements of the variables x_i and y_i.

[5]We choose Spearman's correlation rather than Pearson's correlation since our data is not normally distributed.

A perfectly-calibrated model has a Spearman's correlation coefficient of 1, while a correlation coefficient of 0 conveys a lack of correlation between confidence and accuracy.

Note that other approaches to numerically estimating calibration have been discussed in the literature, e.g. negative log-likelihood (Quinonero-Candela et al., 2005), Brier score (Brier et al., 1950) and expected calibration error (Nixon et al., 2019). Different measurement approaches have different advantages and weaknesses (Ashukha et al., 2020), and no gold standard seems to exist. In this study, we have opted for Spearman's correlation due to the fact that monotonicity in the relation between confidence estimate and accuracy is an important characteristic of good calibration. Spearman's correlation has been previously used to evaluate confidence scores for neural semantic parsers (Dong et al., 2018).

6.2 Performance

Since performance only considers the first rank, it can only be computed on a model level. To measure the performance, we use F1-score and accuracy. We use F1-score since it considers false positives and false negatives through precision and recall. Another reason is the unbalanced distribution of the example utterances across intents. We also include the accuracy since in this particular multi-domain dataset, false negatives have no major risks.

7 Results and analysis

In this section we present our results (averaged across the 10 splits). Our collected data are visual (reliability diagrams and calibration profiles) and numeric (Spearman's correlation, accuracy and F1-score). For our numeric results, we provide the average along with the standard deviation (SD), whereas for the visual results we provide the standard deviation in Appendix B to avoid cluttered diagrams.

7.1 Reliability diagrams

Calibration of the NLUs is visualized through reliability diagrams on model level (figure 3) and rank level (figures 4, 5, 6, 7). In the rank-level reliability diagrams, ranks 4-10 have been merged due to data sparsity.

Model-level results: On a model level (figure 3), all NLUs show a generally monotonic relationship between confidence and accuracy, except for Watson's lower ranges. In particular, Rasa-Sklearn is the closest to the gold standard, and is thus the best calibrated NLU according to this analysis. Moreover, Snips underestimates the true likelihood of predictions, while LUIS is over-confident. We observe a discrepancy in Watson's first 2 bins in the reliability diagram (figure 3) – a sudden underestimation followed by a drop that indicates an extreme overestimation.[6]

Rank-level results: On the first rank (figure 4), the NLUs are fairly well-calibrated in general. On ranks 2 (figure 5) and 3 (figure 6), the degree of calibration decreases (in comparison with the previous rank), for three of the NLUs (Watson, LUIS and Snips – all over-confident), while for the Rasa NLUs the trend seems inverted.

7.2 Calibration score and profile

The calculated Spearman's correlations between the confidence estimates and instance-level accuracy (table 3) show that Rasa-Sklearn has the highest Spearman's correlation with a score of ∼0.51, and is followed by LUIS, Rasa-DIET, Watson, and Snips with the lowest Spearman's correlation of ∼0.507. The difference between LUIS and Rasa-DIET is not significant, while the differences between each other pairs of NLUs are significantly different with a large effect size. (The entire list of t-test results is presented in table 6 in Appendix C.)

[6] As shown by figure 10 in Appendix A, Watson's first two bins are small in comparison with the other NLUs.

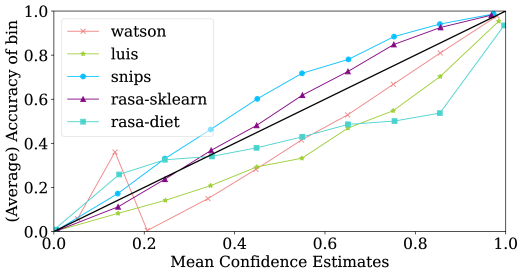


Figure 3: Model-level reliability diagram. The x-axis shows the mean confidence estimates in each bin, while the y-axis shows the mean accuracy of the confidence estimates in each bin (averaged across splits). The black diagonal line plots the identity function representing a gold standard of a perfectly-calibrated model.

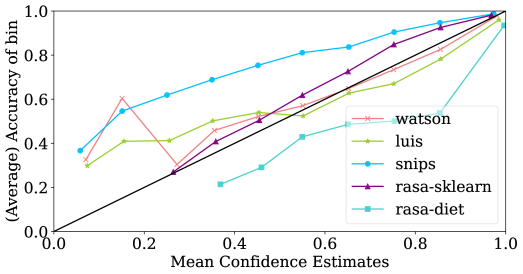


Figure 4: Rank-level reliability diagram on rank 1.

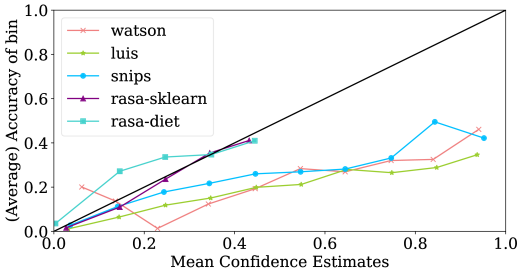


Figure 5: Rank-level reliability diagram on rank 2.

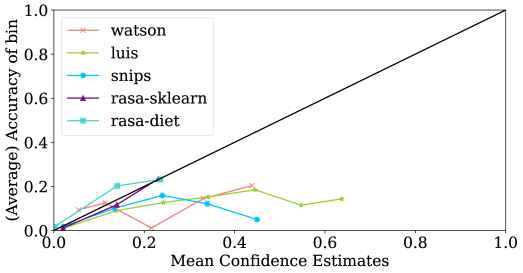


Figure 6: Rank-level reliability diagram on rank 3.

The model-level reliability diagram appears to resonate with the model-level calibration where

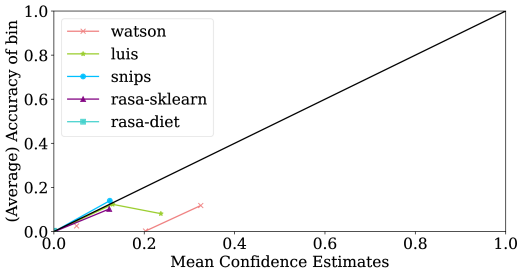


Figure 7: Rank-level reliability diagram on ranks 4-10.

Rasa-Sklearn shows the best calibration in the reliability diagram as well as the strongest monotonicity.

Figure 8 shows the *calibration profile* for each NLU – the Spearman's correlation coefficient as a function of rank. A perfect calibration profile (where calibration is perfect on each rank) would correspond to a straight line along the top of the diagram. In contrast, we can observe that all NLUs have noticeably lower Spearman's correlation for lower ranks. The decrease in Spearman's correlation for lower ranks may indicate that lower ranks are worse calibrated than higher ranks. However, there are reasons to treat these results with some caution.

We can note that the Spearman's correlation is generally lower on a rank level than on a model level. This can be explained by the fact that ranks extend across smaller ranges of confidence estimates (see model-level histogram in Appendix A), which increases variation in one of the correlated variables. Thus, it appears that a higher Spearman's correlation coefficient may be due to a larger variation in the confidence estimates. This may also explain that while figure 8 suggests a decrease in the calibration for lower ranks, the rank-level reliability diagrams show that Rasa-Sklearn and Rasa-DIET have better calibration in lower ranks. Still, on a model level, we take monotonicity to be a characteristic of well-calibrated NLUs. The stronger the monotonicity, the more one can trust an NLU's ranking of hypotheses in a prediction.

7.3 Performance

We measure the performance of the NLUs in intent classification by evaluating accuracy and F1-score. Performance is only evaluated on a model level since it considers the top hypothesis of the NLU's prediction. Our results of the accuracy and F1-scores are averaged across 10 splits for each NLU.

Accuracy: The results in table 4 show that Watson has the highest (∼0.92) and Rasa-Sklearn the lowest (∼0.87) accuracy. The accuracy scores of LUIS and Snips are not significantly different from each other, while all other differences between NLUs are statistically significant with a large effect size.

F1-score: The results in table 5 show that Watson has the highest (∼0.92) and Rasa-Sklearn the lowest (∼0.79) F1-score. All pairwise differences between the NLUs are significant with a large effect size. (The entire pairwise t-test results for accuracy and F1-score are included in table 7 in Appendix C.)

Our performance results are consistent with earlier work comparing Watson, LUIS and Rasa-Sklearn (Liu et al., 2021) that use the complete version of our dataset, and with Abdellatif et al. (2021) who use two datasets from the software engineering domain. However, the results differ from those in Braun et al. (2017) who use Telegram chatbot and StackExchange corpora in a question-answering domain and that has Watson as the worst performing NLU, and Rasa and LUIS on top.

A natural question at this point is whether calibration and performance are correlated. Figure 9 plots calibration (model-level Spearman's correlation) against model-level accuracy and F1 score. Judging from this, calibration and performance are not correlated, indicating a trade-off between calibration and performance (as previously reported for neural networks by Guo et al., 2017).

8 Discussion

In this study, we did not find support for any correlation between calibration and performance (judged by looking only at the top hypothesis). A consequence of this is that when it comes to choosing an NLU for a dialogue system, there is likely to be a trade-off between performance (good for getting the right interpretation) and calibration (good for detecting input that is ambiguous from the NLU perspective).

Differences in degree of calibration across ranks has been observed for all NLUs. Specifically, several of the NLUs are better calibrated for higher-ranking hypotheses than for lower-ranking ones.

NLU	Watson	LUIS	Snips	Rasa-Sklearn	Rasa-DIET
Mean	0.50838	0.50935	0.50669	**0.51024**	0.50906
Median	0.50851	0.50934	0.506491	**0.51026**	0.50888
***p*-value**	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
SD	0.00075	0.00055	0.00064	0.00046	0.00074

Table 3: Model-level calibration scores (Spearman's correlation coefficient ρ)

NLU	Watson	LUIS	Snips	Rasa-Sklearn	Rasa-DIET
Mean	**0.92287**	0.88726	0.88991	0.87263	0.90376
Median	**0.91997**	0.890405	0.89060	0.87866	0.89973
SD	0.00225	0.00417	0.00414	0.00386	0.003860

Table 4: (Averaged) accuracy scores of NLUs

NLU	Watson	LUIS	Snips	Rasa-Sklearn	Rasa-DIET
Mean	**0.92144**	0.88890	0.89029	0.79020	0.81890
Median	**0.91972**	0.89300	0.89166	0.79561	0.81716
SD	0.00234	0.00373	0.00407	0.00358	0.00331

Table 5: (Averaged) F1-scores of NLUs

For dialogue system developers, we may interpret this as indicating that it may be useful to look at the top two or three hypotheses when trying to detect ambiguity in input utterances. Looking at hypotheses ranked lower than 4 is likely to not be very informative. Fortunately, ambiguities are much more frequently 2-way (i.e. there are two possible interpretations of an input) or 3-way than 4-way or more.

It is worth stressing that one of the studied NLUs (Watson) is a multiple-binary classifier (it treats intents independently), while the others are multi-class (they treat intents as mutually exclusive). In this study, we do not investigate whether one type of classifier is more appropriate than another – presumably, both types have benefits and disadvantages. Nevertheless, since our dataset assumes a single correct class for a given utterance[7], our analysis may indirectly favour multi-class classifiers.

When interpreting our results, one should also consider that different NLUs handle out-of-scope (OOS) input differently. Specifically, among the studied NLUs only Rasa does not include an OOS intent. Our exclusion of out-of-scope intents from the intent rankings returned by the NLUs does not rule out the possibility that different OOS handling may have affected the result. A more level-playing field would have required all NLUs to either not consider OOS at all, or for all of them to be trained on the same OOS examples. Unfortunately, since Snips' OOS handling cannot be configured, neither of these options were available. (Larson et al. (2019) evaluated OOS detection for NLUs, but without considering confidence calibration.)

9 Conclusions and future work

We took established calibration measurement approaches and applied them to intent classification of publicly available NLUs. We also extended the chosen measurements with a rank-level analysis. Our findings show that the best calibrated NLU is Rasa-Sklearn and the least calibrated NLU is Snips, while Watson takes the lead as the best performing NLU and Rasa-Sklearn as the worst performing NLU. The results indicate a trade-off between confidence calibration and performance. We also showed differences in degree of calibration across ranks and discussed their implication

[7]Utterances in Liu et al.'s (2021) dataset, on which we build, are labelled with a single correct intent. There are cases of identical utterances for two different intents, but they are very rare (9 out of 25576 unique utterances).

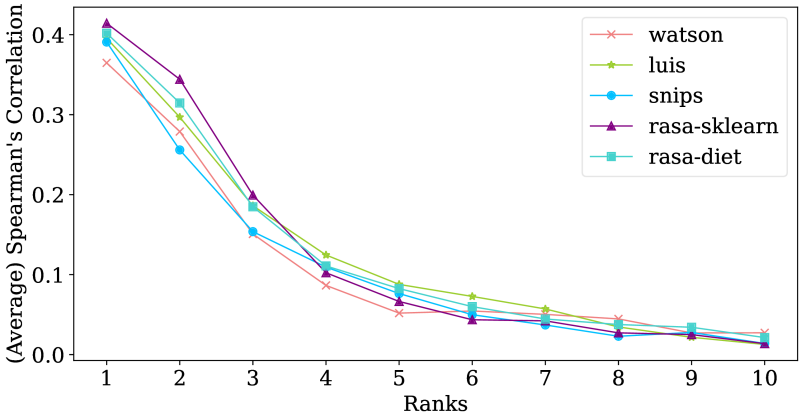
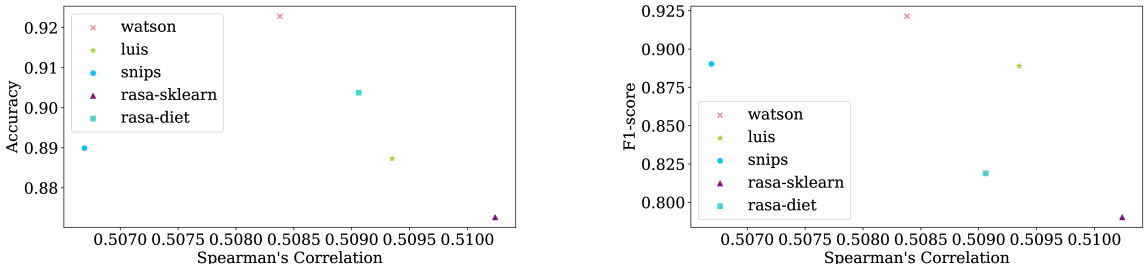


Figure 8: Calibration profiles for all NLUs (Spearman's correlation for ranks 1-10)



(a) Accuracy vs. Calibration

(b) F1-score vs. Calibration

Figure 9: (Model-level) accuracy (a) and F1-score (b) vs. calibration

for dialogue system development.

In future work, it would be interesting to extend the investigation with qualitative analyses of how differences in confidence estimation play out in concrete examples. It could also be valuable to find a better way of assessing how well the NLUs capture genuine ambiguity – something which is difficult with a dataset that assumes a single correct intent for a given utterance.

Acknowledgements

This work was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Ahmad Abdellatif, Khaled Badran, Diego Costa, and Emad Shihab. 2021. A comparison of natural language understanding platforms for chatbots in software engineering. *IEEE Transactions on Software Engineering*.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185.

Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. DIET: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*.

Massimo Canonico and Luigi De Russis. 2018. A comparison and critique of natural language understanding tools. *Cloud Computing*, 2018:120.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753.

Werner Dubitzky, Martin Granzow, and Daniel P Berrar. 2007. *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

Eric Gregori. 2017. Evaluation of modern tools for an OMSCS advisor chatbot. *SMARTech: smartech.gatech.edu*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

IBM. 2010. IBM Watson. Online available at: https://www.ibm.com/watson. Accessed on: 2022-04-14.

Sangkeun Jung. 2019. Semantic vector learning for natural language understanding. *Computer Speech & Language*, 56:130–145.

Rohan Kar and Rishin Haldar. 2016. Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799*.

Falko Koetter, Matthias Blohm, Monika Kochanowski, Joscha Goetzer, Daniel Graziotin, and Stefan Wagner. 2019. Motivations, classification and model trial of conversational agents for insurance companies. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,*, pages 19–30. INSTICC, SciTePress.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

M McTear, Z Callejas, and D Griol. 2016. The conversational interface: Talking to smart devices: Springer international publishing. *Doi: https://doi.org/10.1007/978-3-319-32967-3*.

Microsoft. 2017. LUIS (Language Understanding) - Cognitive Services. Online available at: https://www.luis.ai/home. Accessed on: 2022-04-14.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. *CVPR Workshops*, 2(7).

Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. 2005. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer.

Rasa. 2016. Rasa: Open source conversational AI. Online available at: https://rasa.com/. Accessed on: 2022-04-14.

Kumar Shridhar, Ayushman Dash, Amit Sahu, Gustav Grund Pihlgren, Pedro Alonso, Vinaychandran Pondenkandath, György Kovács, Foteini Simistira, and Marcus Liwicki. 2019. Subword semantic hashing for intent classification on small datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Snips. 2013. Snips.ai. Online available at: https://snips.ai/. Accessed on: 2022-04-14.

Svetlana Stoyanchev, Pierre Lison, and Srinivas Bangalore. 2016. Rapid prototyping of form-driven dialogue systems using an open-source framework. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–219.

Vishal Thanvantri Vasudevan, Abhinav Sethy, and Alireza Roshan Ghias. 2019. Towards better confidence estimation for neural models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7335–7339. IEEE.

Chengwei Xiao, Jiaqi Ye, Rui Máximo Esteves, and Chunming Rong. 2016. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14):3866–3878.

A Histograms of bin sizes

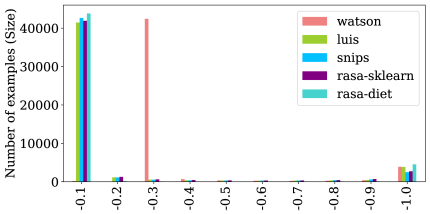


Figure 10: Model-level histogram

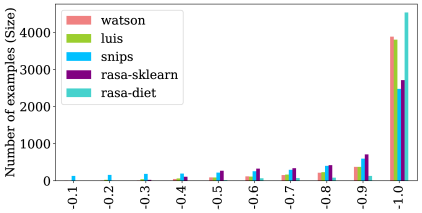


Figure 11: Rank-level (rank 1)

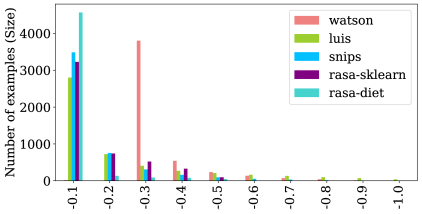


Figure 12: Rank-level (rank 2

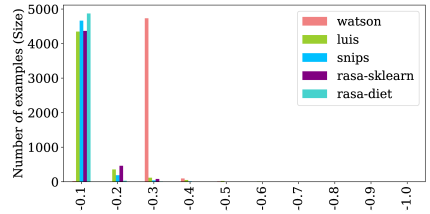


Figure 13: Rank-level (rank 3)

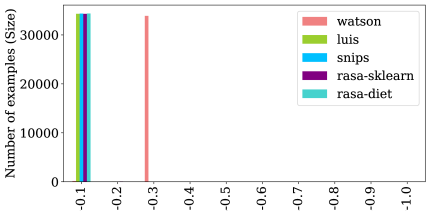


Figure 14: Rank-level (ranks 4-10)

B Reliability diagrams with standard deviation

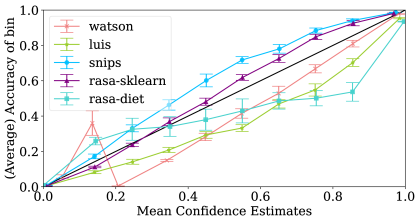


Figure 15: Model level

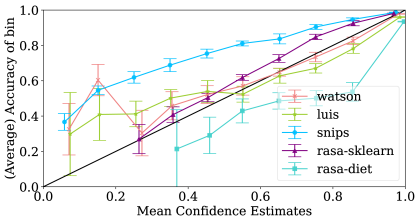


Figure 16: Rank level (rank 1)

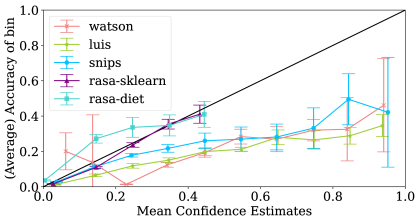


Figure 17: Rank level (rank 2)

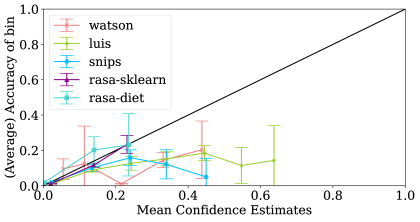


Figure 18: Rank level (rank 3)

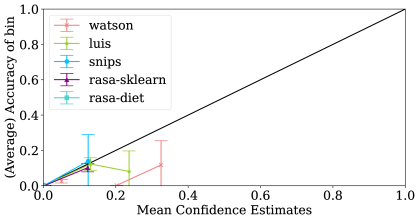


Figure 19: Rank level (ranks 4-10)

C T-test calculations

Pairwise Comp.	t-Statistic	*p*-value	df	Effect Size	SSD (*p*<.05)
(Watson, LUIS)	-3.1645	0.01147	9	L	Yes
(Watson, Snips)	4.9025	0.00084	9	L	Yes
(Watson, Rasa-Sklearn)	-5.4977	0.0003813	9	L	Yes
(Watson, Rasa-DIET)	-2.9555	0.01608	9	L	Yes
(LUIS, Snips)	25.569	<0.00001	9	L	Yes
(LUIS, Rasa-Sklearn)	-3.8306	0.00402	9	L	Yes
(LUIS, Rasa-DIET)	-78.645	0.2895	9	S	No
(Snips, Rasa-Sklearn)	-16.545	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	-7.8118	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-4.1319	0.002552	9	L	Yes

Table 6: T-test for pairwise NLUs' Spearman's correlation scores on a model level

Table 7: T-test for pairwise NLUs' performance

Pairwise Comp.	t Statistics	*p*-value	df	Effect Size	SSD (*p*<.05)
Accuracy					
(Watson, LUIS)	18.462	<0.00001	9	L	Yes
(Watson, Snips)	29.325	<0.00001	9	L	Yes
(Watson, Rasa-Sklearn)	25.059	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	12.82	<0.00001	9	L	Yes
(LUIS, Snips)	-0.62904	0.545	9	N	No
(LUIS, Rasa-Sklearn)	11.672	<0.00001	9	L	Yes
(LUIS, Rasa-DIET)	-7.2468	<0.00001	9	L	Yes
(Snips, Rasa-Sklearn)	13.889	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	-7.7684	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	18.968	<0.00001	9	L	Yes
F1-score					
(Watson, LUIS)	15.437	<0.00001	9	L	Yes
(Watson, Snips)	25.432	<0.00001	9	L	Yes
(Watson, Rasa-Sklearn)	79.213	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	73.47	<0.00001	9	L	Yes
(LUIS, Snips)	1.1095	0.296	9	S	No
(LUIS, Rasa-Sklearn)	95.383	<0.00001	9	L	Yes
(LUIS, Rasa-DIET)	49.549	<0.00001	9	L	Yes
(Snips, Rasa-Sklearn)	135.47	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	88.435	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	18.098	<0.00001	9	L	Yes

Pairwise Comp.	t Statistics	*p*-value	df	Effect Size	SSD (*p*<.05)
		Rank 1			
(Watson, LUIS)	-7.6715	<0.00001	9	L	Yes
(Watson, Snips)	-9.7613	<0.00001	9	L	Yes
(Watson, Rasa-Sklearn)	-11.441	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	-10.782	<0.00001	9	L	Yes
(LUIS, Snips)	1.2402	0.2463	9	S	No
(LUIS, Rasa-Sklearn)	-4.45	0.0016	9	L	Yes
(LUIS, Rasa-DIET)	-1.8668	0.09477	9	M	No
(Snips, Rasa-Sklearn)	-5.7598	0.0002729	9	L	Yes
(Snips, Rasa-DIET)	-3.0576	0.01362	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-6.754	<0.00001	9	L	Yes
		Rank 2			
(Watson, LUIS)	-3.2206	0.01048	9	L	Yes
(Watson, Snips)	6.4881	0.000113	9	L	Yes
(Watson, Rasa-Sklearn)	-17.398	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	-8.6273	<0.00001	9	L	Yes
(LUIS, Snips)	9.9936	<0.00001	9	L	Yes
(LUIS, Rasa-Sklearn)	-9.7455	<0.00001	9	L	Yes
(LUIS, Rasa-DIET)	-3.7508	<0.00001	9	L	Yes
(Snips, Rasa-Sklearn)	-17.882	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	-12.898	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-11.323	<0.00001	9	L	Yes
		Rank 3			
(Watson, LUIS)	-6.7607	<0.00001	9	L	Yes
(Watson, Snips)	-0.6851	0.5105	9	S	No
(Watson, Rasa-Sklearn)	-13.616	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	-6.2648	0.000147	9	L	Yes
(LUIS, Snips)	7.0407	<0.00001	9	L	Yes
(LUIS, Rasa-Sklearn)	-6.3356	0.0001352	9	L	Yes
(LUIS, Rasa-DIET)	0.46202	0.655	9	N	No
(Snips, Rasa-Sklearn)	-11.323	-7.0872	9	L	Yes
(Snips, Rasa-DIET)	-7.0872	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-4.6652	0.001177	9	L	Yes
		Rank 4-10			
(Watson, LUIS)	-5.9362	<0.00001	49	L	Yes
(Watson, Snips)	-0.72951	0.4692	49	N	No
(Watson, Rasa-Sklearn)	0.078179	0.938	49	N	No
(Watson, Rasa-DIET)	-3.3111	0.00175	49	S	Yes
(LUIS, Snips)	9.1052	<0.00001	49	L	Yes
(LUIS, Rasa-Sklearn)	8.087	<0.00001	49	L	Yes
(LUIS, Rasa-DIET)	3.9641	0.0002393	49	M	Yes
(Snips, Rasa-Sklearn)	1.2524	0.2164	49	N	No
(Snips, Rasa-DIET)	-4.1725	0.0001228	49	M	Yes
(Rasa-DIET, Rasa-Sklearn)	5.2551	<0.00001	49	M	Yes

Table 8: T-test for pairwise NLUs' Spearman's correlation scores on a rank level