

HINT3: Raising the bar for Intent Detection in the Wild

Gaurav Arora

Jio Haptik
gaurav@haptik.ai

Chirag Jain

Jio Haptik
chirag.jain@haptik.ai

Manas Chaturvedi

Jio Haptik
manas.chaturvedi@haptik.ai

Krupal Modi

Jio Haptik
krupal@haptik.ai

Abstract

Intent Detection systems in the real world are exposed to complexities of imbalanced datasets containing varying perception of intent, unintended correlations and domain-specific aberrations. To facilitate benchmarking which can reflect near real-world scenarios, we introduce 3 new datasets created from live chatbots in diverse domains. Unlike most existing datasets that are crowdsourced, our datasets contain real user queries received by the chatbots and facilitates penalising unwanted correlations grasped during the training process. We evaluate 4 NLU platforms and a BERT based classifier and find that performance saturates at inadequate levels on test sets because all systems latch on to unintended patterns in training data.

1 Introduction

Over the last few years, task-oriented dialogue systems have gained increasing traction for applications like personal assistants, automated customer support agents, etc. This has led to the availability of several commercialised and/or open conversational bot building platforms. Most popular systems today involve intent detection as a vital part of their Natural Language Understanding (NLU) pipeline. Recent advances in transfer learning (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019) has enabled systems that perform quite well on existing benchmarking datasets (Larson et al., 2019; Casanueva et al., 2020).

Definitions of intent often vary across users, tasks and domains. Perception of intent could range from a generic abstraction such as “Ordering a product” to extreme granularity such as “Enquiring for a discount on a specific product if ordered using a specific card”. Additionally, factors such as imbalanced data distribution in the training set, assumptions during training data generation, diverse

background of domain experts involved in defining the classes make this task more challenging. During inference, these systems may be deployed to users with diverse cultural backgrounds who might frame their queries differently even when communicating in the same language. Furthermore, during inference, apart from correctly identifying in-scope queries, the system is expected to accurately reject out-of-scope (Larson et al., 2019) queries, adding on to the challenge.

Most existing datasets for intent detection are generated using crowdsourcing services. To accurately benchmark in real-world settings, we release 3 new single-domain datasets, each spanning multiple coarse and fine grain intents, with the test sets being drawn entirely from actual user queries on the live systems at scale instead of being crowdsourced. On these datasets, we find that the performance of existing systems saturates at unsatisfactory levels because they end up learning spurious patterns from the training dataset instead of generalising to the perceived meanings of intents.

We evaluate 4 NLU platforms - Dialogflow¹, LUIS², Rasa NLU³, Haptik⁴ and a BERT (Devlin et al., 2019) based classifier on all 3 datasets and highlight gaps in language understanding. We further probe into queries where all the current systems fail and question the efficacy of the current approach of learning. Additionally, we repeat all our experiments on the subset of training data and show a performance drop in all the systems despite retaining relevant and sufficient utterances in the training subset. We’ve made our datasets and code freely accessible on GitHub to promote

¹<https://cloud.google.com/dialogflow>

²<https://www.luis.ai/>

³<https://github.com/RasaHQ/rasa/>

⁴<https://haptik.ai>

⁵Access requests for signup on Haptik are processed via contact form at <https://haptik.ai/contact-us/>

Dataset	Example Intents		Example Queries		
	Type	Label	Train	Test	
				In-scope	Out-of-Scope
SOF Mattress	Generic	OFFERS	What are the available offers Give me some discount	Any other offers	If I order now do i get 20% discount in lockdown period
	Specific	100_NIGHT_TRIAL_OFFER	What is the 100-night offer Trial offer on customisation	Free 100 days trial	I need try
Curekart	Generic	RECOMMEND_PRODUCT	Sir I want to fast gain weight. i'm beginner in gym	I need Beginners hair multivitamin Which is the best whey protein	Role of electrolytes powder Is it help for sperm count
			For biceps and tricep muscle growth supplements	Can i get rivamal 120 ml at my home	Can diabetic patient have it
			Connect with agent My transaction is incorrect My sign up bonus is incorrect Did not receive my amount	chat with customer service agent PowerPlay11 rummy issue Suddenly balance gone Why it is showing wrong balance	Application is not responding during team joining Why my current basketball teams being shown??

Table 1: Few examples of Intents and Queries in Train and Test set in HINT3 dataset

transparency and reproducibility⁶.

2 Prior Work

Despite intent detection being an important component of most dialogue systems, very few datasets have been collected from real users. Web Apps, Ask Ubuntu and Chatbot datasets from (Braun et al., 2017) contain a limited number of intents (<10), oversimplifying the task. More recent datasets like HWU64 from (Liu et al., 2019) and CLINC150 from (Larson et al., 2019) span a large number of intents in multiple domains but are generated using crowd sourcing services hence are limited in diversity in user expressions which arise from but not limited to domain specific presumptions, context from how and where the bot is made available, paraphrases emerging from cultural and ethnic diversity of user base, conversational slang, etc. Our work has some similarity with CLINC150, in that they also highlight the problem of out-of-scope intent detection and with BANKING77 from (Casanueva et al., 2020) that focuses on a single domain. However, all three - HWU64, CLINC150,

⁶<https://github.com/hellohaptik/HINT3>

Dataset	#Intent	#Queries			
		Train		Test	
		Full	Subset	in-scope	oos
SOFMattress	21	328	180	231	166
Curekart	28	600	413	452	539
Powerplay11	59	471	261	275	708

Table 2: Statistics of the 3 datasets in HINT3

BANKING77 offer relatively large and well balanced training set which might not be always feasible to collect for every new domain. For all datasets mentioned so far, recent works have reported a reasonably high performance (>90% average) for in-scope queries. Despite this, gaps in language understanding become apparent when such systems are deployed. Datasets introduced in this paper and further analysis of results attempts to recognise critical gaps in language understanding and calls for further research into more robust methods.

3 Datasets

We introduce HINT3, a collection of datasets shown in Table 2 - **SOFMattress**, **Curekart** and **Powerplay11** each containing diverse set of intents in a single domain - mattress products retail, fitness

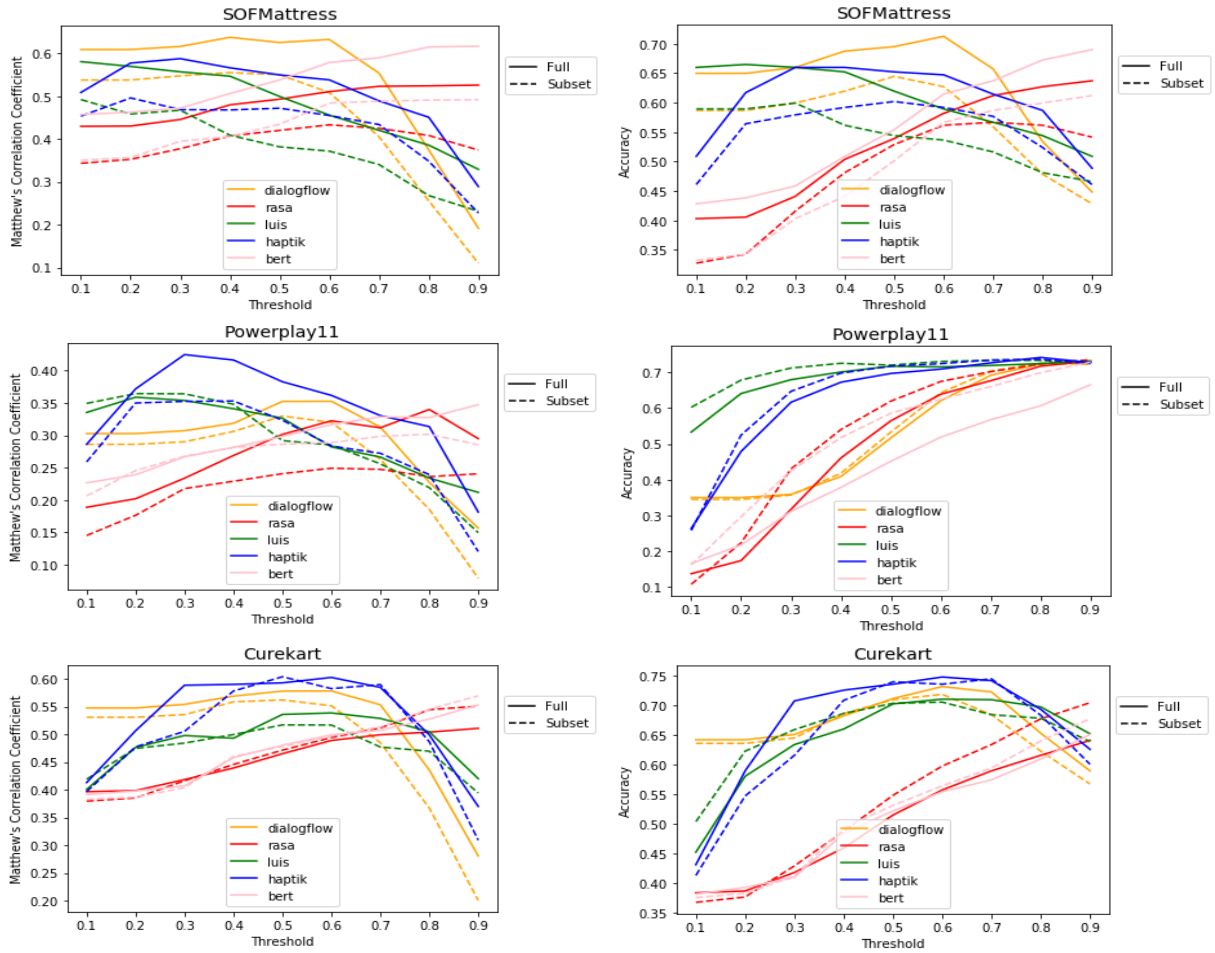


Figure 1: Matthew’s Correlation Coefficient and Accuracy across all datasets and platforms

supplements retail and online gaming respectively. Table 1 shows few example intents of varying granularity in HINT3 dataset, along with examples of training queries created by domain experts and in-scope, out-of-scope queries received from real users.

3.1 Training Data Collection

Training data is prepared by a team of domain experts trying to emulate real users after in-depth research of historical user queries. The experts do not create an explicit set of out of scope queries primarily because the universe of such queries is infinitely big. Training datasets show class imbalance, occurrence of domain specific words, acronyms⁷. All training data queries are in English.

Dataset Variants

In addition to **Full** training sets, we create **Subset** versions for each training set. For each class, after retaining the first query we iterate over the

rest, discarding a query if it has an entailment score (Bowman et al., 2015) greater than 0.6 in both directions with any of the queries retained so far i.e. the subset version has the following property

$$E(x_a, x_b) \leq 0.6 \wedge E(x_b, x_a) \leq 0.6;$$

$$a \neq b, a \in [1, |\hat{X}_i|], b \in [1, |\hat{X}_i|] \forall I$$

where I is the set of all intents, \hat{X}_i is the set of queries retained for class i , $E(h, p)$ is the entailment scoring function with h as hypothesis and p as premise. We use ELMo model trained on SNLI (Peters et al., 2018; Parikh et al., 2016)⁸ for $E(h, p)$. These are intended to evaluate performance with only semantically different sentences in the training set as ideally systems should already understand semantically similar queries to the ones present in the training set.

⁷github.com/hellohaptik/HINT3/tree/master/data_exploration

⁸<https://demo.allennlp.org/textual-entailment>

	SOFMattress		Curekart		Powerplay11	
	Full	Subset	Full	Subset	Full	Subset
Dialogflow	73.1	65.3	75.0	71.2	59.6	55.6
RASA	69.2	56.2	84.0	80.5	49.0	38.5
LUIS	59.3	49.3	72.5	71.6	48.0	44.0
Haptik	72.2	64.0	80.3	79.8	66.5	59.2
BERT	73.5	57.1	83.6	82.3	58.5	53.0

Table 3: Inscope Accuracy at low threshold=0.1 for Full and Subset data variants

3.2 Test Data Collection and Annotation

Our test sets contain the first message received by live systems from real users over a period of 15 days. Inter-annotator agreement was 75.8%, 80.0% and 73.4% for SOFMattress, Curekart and Powerplay11 respectively and conflicts were resolved by domain experts. One major reason for low inter-annotator agreement was unclear criteria for defining an intent which sometimes lead to overlapping intents of different levels of granularity, even after we had made sure to manually merge any conflicting or highly similar intents in the training data.

Directly coming from real users our test set queries also contain messaging slangs, acronyms, spelling mistakes, grammatical mistakes and usage of code-mixed languages⁷. Queries in non-Latin script or code-mixed languages were marked as out of scope (labelled as NO_NODES_DETECTED). Since live chat systems don't cater all the queries related to a brand, our test set contains relevant out-of-scope queries received from users about that domain. Any identifiable information of users, brands was replaced with made-up values in both train and test sets.

4 Benchmark Evaluation

We evaluated the performance of our datasets on platforms like Dialogflow, LUIS, RASA and Haptik in addition to evaluating performance on BERT. All layers of BERT were fine-tuned with a learning rate of 4e-5 for up to 50 epochs with a warmup period of 0.1 and early stopping.

4.1 Out-Of-Scope (OOS) prediction

We use thresholds on the model's probability estimate for the task of predicting whether a query is OOS. We show performance on thresholds ranging from 0.1 to 0.9 at an interval of 0.1 to show the maximum performance a model can achieve irrespective of how we choose the threshold.

4.2 Metrics

We consider Accuracy and Matthew's Correlation Coefficient⁹ as overall performance metrics for the systems. We use OOS recall (Larson et al., 2019) to evaluate performance on OOS queries and accuracy of in-scope queries to evaluate performance on in-scope queries.

5 Results

Figure 1 presents results for all systems, for both Full and Subset variations of the dataset. Best Accuracy on all the datasets is in the early 70s. Best MCC for the datasets varies from 0.4 to 0.6, suggesting the systems are far from perfectly understanding natural language.

In Table 3, we consider in-scope accuracy at a very low threshold of 0.1, to see if false positives on OOS queries would not have mattered, what's the maximum in-scope accuracy that current systems are able to achieve. Our results show that even with such a low threshold, the maximum in-scope accuracy which systems are able to achieve on Full Training set is pretty low, unlike the 90+ in-scope accuracies of these systems which have been reported on other public datasets like CLINC150 in (Larson et al., 2019). And, the in-scope accuracy is even worse for the Subset of the training data.

Table 5 shows percentage drop in in-scope accuracy on subset data across all systems as compared to in-scope accuracy on full data. The drop varies from 0.6% to 22.3% across datasets and platforms. In an ideal world, this drop should be close to 0 across all datasets, as if the system understands the meaning of queries in training data, its performance should not get affected at all by removing queries in training data which are semantically similar to the ones already present.

Analyzing few example queries which failed on all platforms in Table 4 suggests that these models

⁹https://scikit-learn.org/stable/modules/model_evaluation

Test query	True label	Top predicted label	Sample training queries for True label	Sample training queries for predicted label
Ergo 7272 inches price?	MATTRESS_COST	L,H,D,R: ERGO_FEATURES	<ul style="list-style-type: none"> • Price of mattress • Custom size cost 	<ul style="list-style-type: none"> • Features of Ergo mattress • Tell me about SOF Ergo mattress
Trail option are there	100_NIGHT_TRIAL_OFFER	L,H,D: COD R: EMI	<ul style="list-style-type: none"> • Trial details • How to enroll for trial 	<ul style="list-style-type: none"> • Can I get COD option? • Can it deliver by COD
I require 75 inch 57 inch. Is it available?	SIZE_CUSTOMIZATION	L: DISTRIBUTORS H,D,R: WHAT_SIZE_TO_ORDER	<ul style="list-style-type: none"> • Will I get an option to Customise the size • How can I order a custom sized mattress 	<ul style="list-style-type: none"> • Want to know the custom size chart • Show me all available sizes
20 % discount available on emi	OFFERS	L,H,D,R: EMI	<ul style="list-style-type: none"> • Want to know the discount • Tell me about the latest offers 	<ul style="list-style-type: none"> • You guys provide EMI option? • No cost EMI is available?
How will u deliver with this LockDown in place ?	NO_NODES_DETECTED	L,H,D,R: CHECK_PINCODE		<ul style="list-style-type: none"> • Do you deliver to my pincode • Will you be able to deliver here
Covid19 how can you deliver	NO_NODES_DETECTED	L,H,D,R: CHECK_PINCODE		

Table 4: Few examples of test queries in SOFMattress which failed on all platforms, **L:** LUIS, **H:** Haptik, **D:** Dialogflow, **R:** Rasa. NO_NODES_DETECTED is the out-of-scope label.

	SOF Mattress	Curekart	Power play11
Dialogflow	10.6	5.0	6.7
RASA	18.7	4.1	10.5
LUIS	16.8	1.2	8.3
Haptik	11.3	0.6	10.9
BERT	22.3	1.5	9.4

Table 5: Percentage **drop** in Inscope Accuracy at low threshold=0.1 in Subset data as compared to Full

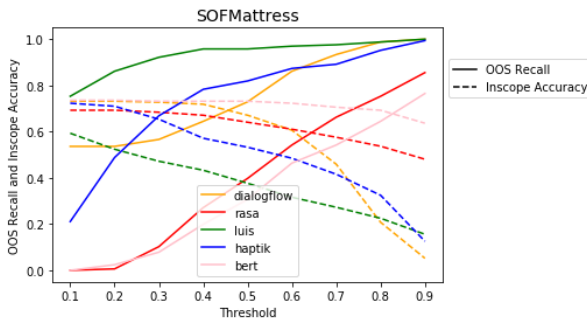


Figure 2: Out-of-Scope (OOS) Recall at the cost of Inscope Accuracy for SOFMattress Full dataset

aren’t actually “understanding” language or capturing “meaning”, instead capturing spurious patterns in training data, as was also pointed in (Bender and Koller, 2020). Predicting based on these spurious patterns, which models latch on to during training, leads to models having high confidence even on OOS queries. Figure 2 shows this behaviour on SOFMattress Full dataset, as significant percentage of OOS queries have high confidence scores on all systems, except LUIS, for which it is at the cost of in-scope accuracy.

6 Conclusion

This paper analyzed intent detection on 3 new datasets consisting of both in-scope and out-of-scope queries received on 3 live chat bots over a period of 15 days. Our findings indicate that

there’s a significant gap in performance on crowd-sourced datasets vs in a real world setup. NLU systems don’t seem to be actually “understanding” language or capturing “meaning”. We believe our analysis and dataset will lead to developing better, more robust dialogue systems.

Acknowledgments

We are grateful to Bot Analysts at Haptik, especially Aaron Dsouza¹⁰, who helped us open-source HINT3 datasets. We also want to thank clients of Haptik who allowed us to share queries received on their bots with the research community.

References

- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- 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, Saarbrücken, Germany. Association for Computational Linguistics.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. [Efficient intent detection with dual sentence encoders](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

¹⁰Reachable at aaron.dsouza@haptik.ai

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 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, Hong Kong, China. Association for Computational Linguistics.
- Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. [Benchmarking natural language understanding services for building conversational agents](#). In *Proceedings of the Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages xxx–xxx, Ortigia, Siracusa (SR), Italy. Springer.
- Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. [A decomposable attention model for natural language inference](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.