Iterative Feature Mining for Constraint-Based Data Collection to Increase Data Diversity and Model Robustness

Stefan Larson[*♣†] **Anthony Zheng**[♠†] **Anish Mahendran**[♠†] **Rishi Tekriwal**[♠†]
Adrian Cheung[♦] **Eric Guldan**[♠†] **Kevin Leach**[♠†] **Jonathan K. Kummerfeld**[♠†]

♣ Rosegold AI
Ann Arbor, MI, USA

♠ University of Michigan
Ann Arbor, MI, USA

♦ Clinc, Inc.
Ann Arbor, MI, USA

Abstract

Diverse data is crucial for training robust models, but crowdsourced text often lacks diversity as workers tend to write simple variations from prompts. We propose a general approach for guiding workers to write more diverse text by iteratively constraining their writing. We show how prior workflows are special cases of our approach, and present a way to apply the approach to dialog tasks such as intent classification and slot-filling. Using our method, we create more challenging versions of test sets from prior dialog datasets and find dramatic performance drops for standard models. Finally, we show that our approach is complementary to recent work on improving data diversity and that training on data collected with our approach leads to more robust models.

1 Introduction

Crowdsourcing is widely used to collect data, including cases where workers are writing new text, such as questions (Rajpurkar et al., 2016), dialog (Budzianowski et al., 2018), and captions (Russakovsky et al., 2015). To avoid repetition of short labels for images, von Ahn and Dabbish (2004) proposed using a taboo list, preventing workers from writing labels that previous workers had written. This idea has since been applied to emotion annotation (Pearl and Steyvers, 2010) and word association (Vickrey et al., 2008; Lafourcade and Joubert, 2012). However, in all of these cases the constraint is that there cannot be an exact match with another label. This limits the approach to tasks where workers write a single word or a short phrase. Meanwhile, recent work on dialog has found that crowdsourced data can have limited diversity (Jiang et al., 2017; Kang et al., 2018; Larson et al., 2019a). This limited diversity has dramatic consequences, as models trained on such data may not generalize well to unseen or uncommon inputs.

We present a generalization of the taboo list idea that can be applied to longer text like sentences. First, rather than features in the taboo list being complete labels, we allow them to be anything, e.g., for intent classification, each feature in the list is a single word that the worker can't use in their new utterance. To create the taboo list, we propose using a simple model to find over-represented features in the data collected so far. Second, rather than having a 1-1 mapping of taboo lists to examples we allow any mapping, e.g., for intent classification we have a taboo list for each intent. To show how this idea improves diversity for longer text, we apply it to crowdsourcing paraphrases for two standard dialog tasks: intent classification and slot-filling.

We evaluate our approach in two ways. First, we generate new test sets for several standard intent classification and slot-filling dialog datasets. We find that results on our new test sets are dramatically lower than on the standard test sets, indicating these standard datasets do not provide data of sufficient diversity to train robust models. Second, we compare our approach to another recent effort to improve diversity in dialog data (Larson et al., 2019a). We collect data with both approaches, a baseline, and a mixture of all three, then evaluate models on all combinations of training and test sets. The mixed approach performs best, indicating that the two approaches complement each other by encouraging different types of diversity.

Simply collecting enormous datasets may be a way to develop robust models, but it is certainly not sample efficient. Without any guidance, workers will mainly write examples that are in the head of the distribution of expressions, only slowly filling in the long tail (if at all). This work provides a method to encourage crowd workers to cover the long tail by using constraints to promote diversity. Our results show that by collecting more diverse data, we can produce more robust and therefore useful models.

[*]Corresponding email: stefan.dataset@gmail.com.
[†]Work performed while author was employed by Clinc.

2 Related Work

Crowdsourcing Dialog Data: Data for most recent task-oriented dialog datasets (Coucke et al., 2018; Gupta et al., 2018; Liu et al., 2019; Larson et al., 2019b), and custom dialog agents (Han et al., 2013; Iyer et al., 2017; Campagna et al., 2017; Ravichander et al., 2017; Shah et al., 2018) has been written by crowd workers via paraphrasing. Recent work has shown that diverse training data is important for robust dialog systems (Kang et al., 2018) and that a range of factors impact the diversity of utterances (Wang et al., 2012; Jiang et al., 2017). There has been some work on improving diversity using outlier detection (Larson et al., 2019a), and our idea is orthogonal to this approach.

Taboo Lists: von Ahn and Dabbish (2004)'s ESP game introduced the taboo list idea that we extend. In their game, a pair of players label an image with a single word up to 13 characters long. If they write the same label, it becomes a label for the image and is added to a taboo list for future players looking at that image. Of the papers in the ACL Anthology that cite their work, 38 cite the general idea of a game-with-a-purpose, but do not use the taboo idea; 25 cite the dataset released with the paper; two have the paper in the references but not in the main text; three use the taboo idea in new games. Two of the new games use static taboo lists defined by the researchers (Pearl and Steyvers, 2010; Vickrey et al., 2008), while the third uses the ESP game approach, but applies it to a new task (Lafourcade and Joubert, 2012). Being based on exact matching limits the range of tasks the taboo idea can apply to. Our work overcomes this limitation. Concurrent work by Yaghoub-Zadeh-Fard et al. (2020) also uses taboo lists to encourage diversity in paraphrases, but they use simple frequency-based taboo word selection, and do not apply their approach to intent classification and slot-filling data.

Adversarial Methods: Our work is related to generation of adversarial examples. Recent work has shown that inserting text can confuse question answering models (Jia and Liang, 2017; Wallace et al., 2019), as can one-word changes to sentences that require world knowledge (Glockner et al., 2018), and changing syntax can confuse pretrained models (Iyyer et al., 2018). The methodology of our first experiment is similar to this work, as we show that models trained on existing crowdsourced datasets perform poorly on the more diverse test sets that we collect.

3 Taboo Data Collection

We propose a general iterative algorithm for data collection that encourages diversity. By introducing constraints, we can force writers to go beyond the most obvious response to a prompt. This increases the diversity of data, which is crucial for the creation of robust models. The general idea works as follows:

- Start with a set of prompts and an empty list of taboo features for each prompt.
- Collect new crowdsourced responses for each prompt while telling workers not to use features from the taboo list for that prompt.
- Identify new taboo features for each prompt.
- Stop or return to the second step above.

This algorithm can be varied in four key ways:

1. The type of prompt.
2. The type of features we make taboo.
3. The method of mining taboo features.
4. The mapping from taboo features to prompts.

Within this framework, the ESP game involves (1) prompts that are images, (2) taboo features that are complete labels assigned to images, (3) making all labels assigned to a prompt taboo features, and (4) having a separate taboo list for each prompt. However, our algorithm is more general than this, enabling use across a range of other tasks with suitable choices of these four properties. For example: the prompts could be text, tables, or audio; the features could be words, longer n-grams, parse structures, or named entity types; the mining method could be a statistical model, rules, or done by other workers; and the mapping of features to prompts could be many-to-one, one-to-many, or many-to-many.

3.1 Application to Dialog Tasks

We consider two dialog tasks: intent classification and slot-filling. In both cases, we have (1) either example utterances or scenarios as prompts, (2) words as taboo features, (3) taboo words identified using a model, and (4) a set of taboo words for each dialog intent or slot type. For intent classification, (3) is achieved by training a linear SVM with a bag-of-words representation on all of the data. For each intent label we take the highest weighted words over a certain frequency (5 in our experiments) in the SVM model and make them taboo words. The intuition for this approach is that the SVM identifies tokens that are over-represented within a label set and so may lead models to learn

Taboo	Paraphrases of *What is the capital of Florida?*
-	*what city is the state capital of florida* *what is florida's capital*
florida	*what is the capital of FL* *what is the capital of the sunshine state*
capital	*what is florida's statehouse city* *where is the seat of government in florida*
what	*i would like to know the capital of florida* *tell me the name of florida's capital*

Table 1: Crowdsourced paraphrases with variation in the taboo features workers could not use.

only surface cues. Similarly, for slot-filling, (3) is achieved by training a CRF with token features on all of the data. For each slot we use tokens with high weights that are from the context (not the slot itself) as taboo words. In the slot-filling case, we restrict to context words because slot diversity can be introduced by substituting values from a list.

As a motivating example of how this can encourage diversity, consider the crowdsourced paraphrases in Table 1. In all cases, workers received the same prompt, but in the first section they had no constraints and in the other three sections they were not permitted to use a particular taboo word. All of the paraphrases are accurate, but the type of changes depends heavily on the taboo word. Without a taboo word, paraphrases are very similar to the prompt. For "florida", crowd workers used real-world knowledge of nicknames and acronyms to refer to the state, but kept the rest of the sentence the same. For "capital", they again used world knowledge, but also started modifying the rest of the sentence. For "what", they were forced to make substantial changes to the sentence. More examples can be found in Appendix A.

4 Experiments

To demonstrate our approach we consider two experiments. First, we show that existing published datasets are brittle, with training sets that are not sufficiently diverse to train robust models. Second, we show how our approach can be used to collect more robust training data from scratch.

4.1 Challenge Test Versions of Current Datasets

As discussed in Section 2, most existing datasets were crowdsourced with a fixed set of prompts and no taboo constraints, which leads to limited diversity in the data. As a result, models trained on the data may be brittle, failing when tested on new data in the same domain. To test this, we use our taboo approach to create new test sets. If the original training set is diverse then models will achieve high performance on the new test set. We also measure the vocabulary size of each new test set, hypothesizing that as the number of taboo words increases, so does the vocabulary size.

It would be very expensive to collect new versions of every intent and slot type in every dataset, so we randomly sample a subset for our experiments. We crowdsourced the paraphrases using Amazon Mechanical Turk. Paraphrases were checked by hand to ensure they were semantically valid. We collected 3 paraphrases per prompt. We consider NewTable (Jaech et al., 2016), Facebook (Gupta et al., 2018), Snips (Coucke et al., 2018), ATIS (Hemphill et al., 1990), Liu et al. (2019), and Larson et al. (2019b). These cover restaurant and flight booking, home media control, and general knowledge queries. More details can be found in Appendices B and C.

4.2 Robust Data Collection for New Datasets

Our second experiment involves bootstrapping datasets from scratch. We compare four data collection approaches:
1) **`same`**: static prompts, the standard approach.
2) **`unique`**: Larson et al. (2019a)'s approach. They collect data in several rounds, with new prompts chosen using outlier detection to get samples from underrepresented regions in the space of utterances.
3) **`taboo`**: our proposed approach from Section 3.
4) **`mixed`**: a random sample from each approach, with the same amount of total data.

For intent classification, we use the data from Larson et al. (2019a) for `same` and `unique`. For slot-filling we considered three domains, flight booking, money transfer, and restaurant booking, but display results for the first two in Appendix F due to lack of space (the trends were very similar).

We conducted three rounds of data collection using each method on each dataset. The first round was shared across all three methods. The second and third rounds were collected using either the same prompt (`same`) or new prompts (`unique` and `taboo`). Crowd workers were asked to write five paraphrases for each prompt in intent classification and three for each prompt in slot-filling.

Following Larson et al. (2019a) and advice in Gorman and Bedrick (2019), we average results across 10 runs with different random train/test

Dataset	Original Test Example	Taboo Words	Paraphrase Written with Taboo Constraints
Facebook	*where is the closest back road exit*	station, where	*find me the closest back road exit*
	how long will i'll be in traffic	long, time	*what is the period i'll be in traffic for*
	did black bear wash flood last night?	icy, flooding	*did black bear wash overflow with water last night?*
Larson	*how much in taxes will i owe*	taxes, tax	*how much do i owe uncle sam?*
	this charge is bs	fraud, fraudulent	*this charge is a mistake.*
	when should my tires be changed	tires, change	*when do i need new shoes on my car?*
Snips	*play a sixties soundtrack*	hear, play	*put on a sixties soundtrack*
	weather in kaneville maryland	forecast, weather	*atmospheric conditions showing for kaneville maryland*
	play all things must pass.	hear, play	*i want to listen to all things must pass*

Table 2: Examples where BERT gets the original utterance right, but our paraphrase wrong. The paraphrases were crowdsourced using our *taboo* method, which requires crowd workers to avoid using certain words in their paraphrases. Note that taboo words are defined for each intent and so do not always occur in the prompt sentence.

# Taboo	ATIS	Snips	Larson	FB	Liu
0	88.2	96.9	93.7	77.4	93.4
2	55.0	94.9	80.4	72.6	84.5
4	56.7	93.3	73.8	65.6	85.7
6	51.0	93.2	74.5	64.9	76.5

(a) Intent Classification Accuracy.

# Taboo	ATIS	Snips	FB	Newtable
0	92.6	68.3	93.5	92.6
2	83.6	73.6	76.0	87.8
4	81.4	64.2	47.3	80.6
6	78.0	67.8	42.9	78.1

(b) Slot-filling F_1.

Table 3: Results of testing on paraphrased test sets using taboo paraphrases. There is a substantial drop in performance observable across almost all datasets as the number of restricted words increases.

# Taboo	ATIS	Snips	Larson	FB	Liu
0	559	1341	946	808	409
2	575	1394	1098	922	423
4	598	1514	1225	894	484
6	668	1495	1345	977	432

(a) Intent Classification Dataset Vocabulary Size

# Taboo	ATIS	Snips	FB	Newtable
0	249	308	264	226
2	276	302	313	210
4	293	305	368	251
6	283	299	352	281

(b) Slot-filling Dataset Vocabulary Size

Table 4: Vocabulary sizes (number of token types) in each paraphrased dataset for a given number of taboo words. In almost all cases, the vocabulary size increases with the number of taboo words.

splits. In each case, the test data is drawn only from the second and third rounds of data collection to ensure there is no train-test data overlap across methods (since the first round data is shared).

4.3 Models

In both experiments, we use standard models: BERT (Devlin et al., 2019) for intent classification, and a Bi-LSTM for slot-filling. The Appendices show results using an SVM and FastText for intent classification, which showed the same trends as BERT, though more severe. More model details can be found in Appendix D.

5 Results

5.1 Challenge Versions of Current Test Sets

Tables 3a and 3b show the impact of collecting more diverse test cases using our approach. Performance consistently decreases as the number of taboo words increases from 0 to 4. Even with just two taboo words, the median performance drop for BERT is 9 points. In the worst case, ATIS, it drops 33.2 points. The shift is even more severe for FastText (results in Appendix E). Table 4 shows that the vocabulary size tends to increase as the number of taboo words increases. For instance, it increases from 946 to 1345 with 6 on the Larson data, further indicating that crowd workers generate more diverse text using our approach.

Table 2 shows examples where BERT was right on the original test set but wrong on our paraphrase. The new versions generally do not appear significantly different. There are a few exceptions, for instance: "uncle sam" is a more creative though still reasonable phrase that we would want our systems to handle; and "shoes" instead of "tires", which seems unlikely to occur naturally. These stranger cases are relatively rare, but may be worth filtering out with a checking process in future work.

In general, these results indicate that current intent classification and slot-filling evaluation datasets are less than ideal insofar as they do not supply the diversity needed to train robust models. We posit that such datasets are also *too easy* due to this lack of diversity, and do not sufficiently

Intent	Round 2 Taboo Words	Round 2 Examples	Round 3 Taboo Words	Round 3 Examples
routing	routing number help	*what is my bank's rtn* *what are my bank's nine aba digits* *what is the first set of numbers on the bottom of my check*	rtn identifier route	*what are my banks aba digits* *need 9 digit numbers on left side of check for bank* *where is the aba digit listed*
balance	balance have in	*what amount of currency do i own* *please tell me my checking amount* *what is the sum total of my money*	dollars quantity amount	*how can i check my account sum total* *i got how much money* *what is the value of my bank account*
hours	hours open time	*when can i come in to the bank* *the bank shuts down when exactly* *can you check when i can go to the bank*	when early latest	*at what hour does my bank start and end business* *can you check my banks schedule of operations* *what is the earliest i can go to my bank*
phone	number call contact	*how do i reach my bank by phone* *what do i dial to get through to my bank* *how do i phone my bank*	phone reach connect	*how do i ring my bank* *how do i get my bank on the line* *what digits do i press to dial my bank*
checks	checks ordering checkbook	*i need to order new cheques* *could i have a refill for my chequebook as it is empty* *find me a new checkbook, mine's empty*	drafts more slips	*i just wrote my last check can i get others* *can i get some blanks* *i've got to top up my check supply*

Table 5: Example sentences generated by each round of data collection using our taboo crowdsourced paraphrase method along with learned taboo words (accumulated after each round). Restricting crowd workers from using certain taboo words leads to vocabulary and language modifications.

Task	Training Data	Test Data same	Test Data unique	Test Data taboo	Test Data mixed
Intent Classification	same	**99.3**	83.2	83.6	88.6
	unique	98.7	98.4	80.9	92.7
	taboo	99.0	89.7	**97.6**	95.4
	mixed	99.0	**98.8**	**97.6**	**98.5**
Slot Extraction	same	90.9	75.8	77.0	81.0
	unique	90.1	80.4	75.0	81.7
	taboo	90.1	77.2	84.9	84.0
	mixed	**95.8**	**90.9**	**90.6**	**92.3**

Table 6: Model performance for various combinations of training and test data collection methods.

test a model's ability to generalize. These observations are complementary to recent work (Béchet and Raymond, 2018; Niu and Penn, 2019; Larson et al., 2020) that found the ATIS dataset in particular to lack sufficient diversity to evaluate modern slot-filling models.

5.2 Robust Data Collection

Table 6 presents accuracy (top) and F_1 (bottom) for models trained and tested with different data collection methods. As expected, mixed is consistently the best approach. Ignoring mixed, the highest scores are on the diagonals: classifiers trained and tested on data collected using the same method perform the strongest. Looking at the off-diagonals, it seems that taboo and unique are introducing different types of diversity. Both methods see a drop in performance on data collected the other way, and both do well on same's data. However, the drop tends to be larger for models trained on the unique data. This was particularly true for FastText (results in Appendix F), which lacks BERT's large-scale pretraining on external data.

Looking at the data, there are shifts similar to those visible in the examples in Table 1. As shown in Table 5, these included vocabulary changes such as (1) "dial" and "ring" replacing "call", (2) "cheques" replacing "checks" (a spelling substitution), and (3) "digit" instead of "number". They also included use of real-world and domain-specific knowledge, replacing a bank account's "routing number" with "aba digits", "rtn", and "the first set of numbers on the bottom of [a] check". We also looked at the examples in the taboo set broken down by the number of taboo words. We find that the examples sometimes became more unusual as the number of taboo words increased, suggesting two might be enough to introduce diversity without becoming too odd. Finally, we observe that models trained on taboo data are robust to new test sets gathered using taboo, while unique is much less robust (see Appendix F).

6 Conclusion

This paper presents a novel way of guiding data collection away from over-represented areas in the sample space. We show how the approach is a generalization of prior work in crowdsourcing and present a new form of it for dialog data. In experiments on a range of datasets, we show that prior data collection approaches fail to capture diverse examples, leading to brittle models. Finally, we show our approach is complementary to other efforts to increase data diversity, producing higher quality datasets. Collecting data by combining the standard approach, outlier-based collection, and our taboo-based approach produces better training data that in turn leads to more robust models.

Acknowledgements

We thank David Jurgens and the anonymous reviewers for their helpful feedback and discussion.

References

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

Frédéric Béchet and Christian Raymond. 2018. Is ATIS too shallow to go deeper for benchmarking spoken language understanding models? In *Proceedings of InterSpeech 2018*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Giovanni Campagna, Rakesh Ramesh, Silei Xu, Michael Fischer, and Monica S. Lam. 2017. Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

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 (NAACL)*.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Seungyeop Han, Matthai Philipose, and Yun-Cheng Ju. 2013. Nlify: Lightweight spoken natural language interfaces via exhaustive paraphrasing. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop*.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Aaron Jaech, Larry Heck, and Mari Ostendorf. 2016. Domain adaptation of recurrent neural networks for natural language understanding. In *Proceedings of InterSpeech 2016*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding task design trade-offs in crowdsourced paraphrase collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yiping Kang, Yunqi Zhang, Jonathan K. Kummerfeld, Lingjia Tang, and Jason Mars. 2018. Data collection for dialogue system: A startup perspective. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Mathieu Lafourcade and Alain Joubert. 2012. Long tail in weighted lexical networks. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex)*.

Stefan Larson, Eric Guldan, and Kevin Leach. 2020. Data query language and corpus tools for slot-filling and intent classification data. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019a. Outlier detection for improved data quality and diversity in dialog systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

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. 2019b. 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)*.

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)*.

Jingcheng Niu and Gerald Penn. 2019. Rationally reappraising ATIS-based dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lisa Pearl and Mark Steyvers. 2010. Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Abhilasha Ravichander, Thomas Manzini, Matthias Grabmair, Graham Neubig, Jonathan Francis, and Eric Nyberg. 2017. How would you say it? eliciting lexically diverse dialogue for supervised semantic parsing. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3).

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in ATIS? In *Spoken Language Technology Workshop (SLT)*.

David Vickrey, Aaron Bronzan, William Choi, Aman Kumar, Jason Turner-Maier, Arthur Wang, and Daphne Koller. 2008. Online word games for semantic data collection. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics (TACL)*, 7.

William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvit. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT)*.

Mohammad-Ali Yaghoub-Zadeh-Fard, Boualem Benatallah, Fabio Casati, Moshe Chai Barukh, and Shayan Zamanirad. 2020. Dynamic word recommendation to obtain diverse crowdsourced paraphrases of user utterances. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*.

Appendices

A More examples for Section 3

Table 7 displays more examples relevant to the Table 1 discussed in Section 3 of the main paper.

B Data Collection

All data was collected using crowdsourcing. We used the Amazon Mechanical Turk crowdsouring platform. Workers were presented with a prompt which asked them to paraphrase a question or a statement n times (n was 3 in all experiments except the "Robust Data Collection" for intent classification data, where n was 5). An example of a question in a prompt could be "what is my balance?", while a statement could be "tell me how much money I have". Workers were paid $0.05 per paraphrase. We used prompts similar to those shown in Figure 1. For the data collected in the "Challenge Versions of Current Datasets" experiments, we sampled test samples from each dataset's test set, and asked crowd workers to paraphrase these samples. Taboo words were presented as comma-separated lists in prompts. We used a regular expression to prohibit workers from submitting paraphrases that contained taboo words. In the "Challenge Datasets"

Rephrase an original question or statement

Suppose you have an intelligent device such as Amazon Alexa, Apple Siri, or Google Assistant.

Given an original phrase, provide 5 different ways of saying the same phrase.

Original phrase: "how's the weather"

Scenario:

Determine the type of aircraft used on a flight from Cleveland to Dallas that leaves before noon.

Rephrase an original question or statement

Suppose you have an intelligent device such as Amazon Alexa, Apple Siri, or Google Assistant.

Given an original phrase, provide 5 different ways of saying the same phrase.

Original phrase: "what is my routing number"

Don't use the words "routing" or "number" in your responses.

Figure 1: Examples of data collection prompts for rephrase (top, from Larson et al. (2019b)) and scenario (middle, from ATIS) tasks. An example of a rephrase prompt used in the present work with taboo words is shown at bottom.

experiments, each round of data collection introduced 2 new taboo words (except the initial round). In the "Robust Data Collection" experiments, each round of data collection introduced 3 new taboo words for the intent classification experiments (except the initial round), and 2 taboo words for each slot (except the initial round) for the slot-filling experiments.

B.1 Preprocessing

All crowdsourced paraphrases were checked by hand to ensure they were semantically valid. Queries were tokenized on white space. For the slot-filling "Robust Data" experiments, crowd workers were asked to use default slot values in their paraphrases. We used large lists of replacement slot values to replace the default values, so that the slot-filling models would not memorize the default values.

C Datasets used in "Challenge Versions" experiments

This section provides more detail on the datasets investigated in the "Challenge Versions of Current Datasets" experiments.

Newtable: A slot-filling dataset from Jaech et al. (2016) meant for booking restaurants using a virtual assistant. For the "Challenge Versions" experiment, we sampled two slots (`people` and `place`) for the slot filling experiment. We sampled 50 queries to be used as seeds to be paraphrased by crowd workers. All sampled queries contained at least one slot (`people` or `place`).

Taboo	Paraphrases of *What is the capital of Florida?*
-	*- what city is the state capital of florida* *- what is florida's capital* *- florida's capital is what* *- can you name the capital of florida*
florida	*- what is the capital of FL* *- what is the capital of the state that is located directly south of georgia* *- what is the capital of the state where miami is located* *- what is the capital of the sunshine state*
capital	*- what is florida's statehouse city* *- where is the state government of florida headquartered* *- where is the seat of government in florida* *- what city does the governor of florida live in*
what	*- i would like to know the capital of florida* *- can you tell me florida's capital* *- provide the name of the capital of florida* *- tell me the name of florida's capital*

Table 7: Crowdsourced paraphrases with variation in the taboo features workers could not use.

Facebook: An intent classification and slot-filling dataset from Gupta et al. (2018), with intents related to interacting with a task-driven virtual assistant. For the "Challenge Versions" experiment, we sampled 10 intents for the intent classification experiment and two slots (`source` and `destination`) for the slot filling experiment. We sampled 30 queries from each sampled intent to be seeds for the intent classification experiment. We sampled 50 queries containing both `source` and `destination` slots to be seeds for the slot filling experiment.

Snips: An intent classification and slot-filling benchmark from Coucke et al. (2018). For the "Challenge Versions" experiment, we sampled all (seven) intents for the intent classification experiment and two slots (`entity` and `playlist`) from the dataset's `AddToPlaylist` intent. We sampled 50 queries from each intent to be used as crowdsourcing seed, and sampled 50 queries from the `AddToPlaylist` intent for the slot filling experiment (all these sampled contained at least one slot (either `entity` or `playlist`).

Larson: An intent classification benchmark with a wide variety of topic domains and a larg number of intents with limited training data per intent class

	Larson		Snips		ATIS		Facebook		Liu	
# Taboo	FastText	BERT	FastText	BERT	FastText	BERT	FastText	BERT	FastText	BERT
0	83.3	93.7	94.6	96.9	83.7	88.2	72.7	77.4	78.2	93.4
2	59.6	80.4	90.4	94.9	45.2	55.0	62.6	72.6	65.3	84.5
4	42.1	73.8	84.1	93.3	43.3	56.7	51.6	65.6	54.7	85.7
6	29.0	74.5	83.0	93.2	36.1	51.0	46.2	64.9	44.6	76.5

Table 8: Results of testing on paraphrased test sets using taboo paraphrases for the classifier datasets using FastText and BERT. Across almost all datasets and models there is a substantial drop in performance as the number of restricted words increases.

(Larson et al., 2019b). For the "Challenge Versions" experiment, we sampled 40 intents for the intent classification experiment. From each sampled intent, we sampled 10 queries to be used as seeds for crowdsourcing paraphrases.

ATIS: The ATIS corpus (Hemphill et al., 1990) has long been a benchmark for evaluating both slot-filling and intent classification models. We use the dataset split as used by Tur et al. (2010). Intents in ATIS are related to interacting with a flight booking virtual assistant. For the "Challenge Versions" experiment, we sampled six intents for the intent classification experiment and two slots (`to-city` and `from-city`) for the slot filling experiment. For the intent classification experiment, we sampled between 8 and 50 queries to serve as seeds to crowdsourcing paraphrase tasks for the intent classification experiment. For the slot filling experiment, we sampled 50 queries containing both `to-city` and `from-city` to serve as seed phrases for the crowdsourcing paraphrase task.

Liu: We use the dataset from (Liu et al., 2019) as an intent classification benchmark. Intents from this dataset are similar to the Facebook and Snips datasets. For the "Challenge Versions" experiment, we sampled 10 intents for the intent classification experiment. From each intent, we sampled 10 queries to be seeds for crowdsourcing paraphrase tasks.

C.1 A note on ATIS

The ATIS corpus has long been a benchmark for evaluating both slot-filling and intent classification models. While the ATIS dataset was generated in the early 1990s, and hence did not use any modern crowdsourcing platform like Amazon Mechanical Turk to generate data, the corpus was nonetheless collected using a *scenario*-driven data collection scheme using non-expert workers. The ATIS corpus saw human "subjects" recruited to generate natural language queries targeting an automated flight booking system. Subjects were given scenarios with goals (e.g. booking a flight with time or fare constraints). This is essentially the same as the methods used in crowdsourcing today, but with a small set of participants rather than the crowd. An example of such a scenario prompt from the ATIS data collection procedure is shown in Figure 1 (bottom).

C.2 Train-Test Splits for "Challenge Versions of Current Datasets"

For each dataset described above, we generate new test phrases with our taboo paraphrasing method using samples from each dataset's published test set as seed phrases to the crowdsourcing prompts. With the exception of the Liu dataset, all datasets have standard train-test splits: we randomly created an 85-15 train-test split for the Liu dataset.

D Computing and Model Details

The main contribution of our paper is not in model development, but in data collection. However, we discuss the relevant aspects of the models used in the experiments here. We used off-the-shelf BERT, SVM, and FastText models for the intent classification experiments. For BERT, we used the BERT large (uncased) model from `https://github.com/google-research/bert`; when using this model we fine tuned the model to each dataset. We used the sklearn's SVC classifier as our SVM; with the SVM we used bag-of-words feature representations. We used the version of FastText found here: `https://github.com/facebookresearch/fastText`. The slot-filling experiments used a bi-directional LSTM for the evaluation experiments, and a CRF (using token features) for the model to identify taboo words. The LSTM model was adapted from (Finegan-Dollak et al., 2018) and uses pre-trained GloVe word embeddings. The CRF was adapted from `https://sklearn-crfsuite.readthedocs.io/`. All model experiments were run on an Nvidia GPU (in the case

Model	Training	Test Set same	unique	taboo	mixed
SVM Accuracy	same	98.3	77.4	58.9	77.9
	unique	97.8	**97.9**	58.5	85.0
	taboo	97.5	84.1	**94.9**	92.4
	mixed	**98.6**	97.8	94.8	**97.1**
FastText Accuracy	same	98.4	76.8	62.7	78.6
	unique	97.8	**98.1**	62.2	85.9
	taboo	97.6	81.6	94.3	90.8
	mixed	**98.6**	**98.1**	**94.8**	**97.2**
BERT Accuracy	same	**99.3**	83.2	83.6	88.6
	unique	98.7	98.4	80.9	92.7
	taboo	99.0	89.7	**97.6**	95.4
	mixed	99.0	**98.8**	**97.6**	**98.5**

Table 9: Classifier model accuracy when training and testing on data collected by each data collection method. Models not trained on the taboo data perform poorly on the taboo data, indicating that data collected by taboo is challenging.

Domain	Training	Test Set same	unique	taboo	mixed
flights F_1	same	96.4	83.3	67.4	83.2
	unique	94.7	92.2	68.4	86.0
	taboo	94.7	86.1	82.7	88.2
	mixed	**98.0**	**94.1**	**85.7**	**93.0**
transfer F_1	same	97.9	91.8	70.8	87.8
	unique	97.7	95.7	70.2	88.9
	taboo	96.0	90.2	83.4	90.2
	mixed	**98.6**	**96.6**	**84.8**	**93.8**
restaurant F_1	same	90.9	75.8	77.0	81.0
	unique	90.1	80.4	75.0	81.7
	taboo	90.1	77.2	84.9	84.0
	mixed	**95.8**	**90.9**	**90.6**	**92.3**

Table 10: Slot-filling F_1 performance on various slot-filling datasets. The performance of models trained on data collected using the same and unique methods drop substantially when tested on data collected using the taboo method. Models trained on the mixed datasets produce the best performance overall.

of BERT and the bi-LSTM) or using in Intel i7 CPU (all other models).

As our paper does not introduce a new model, we do not compare average runtimes for each approach, nor do we compare number of parameters in each model, as each of the models we use in our experiments are well-established.

E FastText Results for "Challenge Versions" Experiments

Table 8 shows side-by-side comparison of FastText and BERT for the "Challenge Versions of Current Datasets" intent classification experiments. The performance drop for FastText is much more severe than BERT, falling to as low as 29.0 accuracy on the Larson dataset.

F Additional "Robust Data Collection" Results

Tables 9 and 10 show additional results (SVM and FastText) for the intent classification experiment and on additional datasets for the slot-filling experiment. Table 11 shows the performance of a BERT classifier when trained and tested on data collected using the same method. This experiment mimics the setup of the "Challenge Version" experiment. When trained and tested on data collected using the taboo method, BERT stays robust even when the number of taboo words for the test set is increased. However the performance of the classifier trained and tested on the data collected using the same and unique methods suffers when the number of taboo words for the test set is increased.

# Taboo	same	unique	taboo
0	98.1	98.6	97.6
2	79.0	82.1	97.3
4	80.1	80.5	96.4
6	78.5	76.7	96.1

Table 11: Classifier model accuracy when the training and testing data are collected using the same method, keeping seed prompts constant, but with a varying number of taboo words used for the testing set. The classifier used here is BERT. We observe that taboo yields a model that is robust to the taboo data collection method that was able to "break" the models trained on the published datasets in the "Challenge Versions" experiments in Section 5.1. The unique and same approaches are much less robust.

F.1 Dataset Statistics for "Robust Data Collection" Experiments

The sizes of the datasets used in the "Robust Data Collection" Experiments are presented here. For intent classification, same had 6091 samples, unique had 5999 samples, and taboo had 6097 samples. For the slot filling experiments, flights-same had 639 samples, flights-unique had 648 samples, and flights-taboo had 586 samples. The transfer-same had 601, transfer-unique had 618, and transfer-taboo had 529 samples. The restaurant-same had 632, restaurant-unique had 629, and restaurant-taboo had 649 samples.