How Would You Say It?
Eliciting Lexically Diverse Data for Supervised Semantic Parsing

Abhilasha Ravichander[1]*, Thomas Manzini[1]*, Matthias Grabmair[1]
Graham Neubig[1], Jonathan Francis[12], Eric Nyberg[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Robert Bosch LLC, Corporate Sector Research and Advanced Engineering
{aravicha, tmanzini, mgrabmai, gneubig, ehn}@cs.cmu.edu
jon.francis@us.bosch.com

Abstract

Building dialogue interfaces for real-world scenarios often entails training semantic parsers starting from zero examples. How can we build datasets that better capture the variety of ways users might phrase their queries, and what queries are actually realistic? Wang et al. (2015) proposed a method to build semantic parsing datasets by generating canonical utterances using a grammar and having crowdworkers paraphrase them into natural wording. A limitation of this approach is that it induces bias towards using similar language as the canonical utterances. In this work, we present a methodology that elicits meaningful and lexically diverse queries from users for semantic parsing tasks. Starting from a seed lexicon and a generative grammar, we pair logical forms with mixed text-image representations and ask crowdworkers to paraphrase and confirm the plausibility of the queries that they generated. We use this method to build a semantic parsing dataset from scratch for a dialog agent in a smart-home simulation. We find evidence that this dataset, which we have named SMARTHOME, is demonstrably more lexically diverse and difficult to parse than existing domain-specific semantic parsing datasets.

1 Introduction

Semantic parsing is the task of mapping natural language utterances to their underlying meaning representations. This is an essential component for many tasks that require understanding natural language dialogue (Woods, 1977; Zelle and Mooney, 1996; Berant et al., 2013; Branavan et al., 2009; Azaria et al., 2016; Gulwani and Marron, 2014; Krishnamurthy and Kollar, 2013). Orienting a dialogue-capable intelligent system is accomplished by training its semantic parser with utterances that capture the nuances of the domain. An inherent challenge lies in building datasets that have enough lexical diversity for granting the system robustness against natural language variation in query-based dialogue. With the advent of data-driven methods for semantic parsing (Dong and Lapata, 2016; Jia and Liang, 2016), constructing such realistic and sufficient-sized dialog datasets for specific domains becomes especially important, and is often the bottleneck for applying semantic parsers to new tasks.

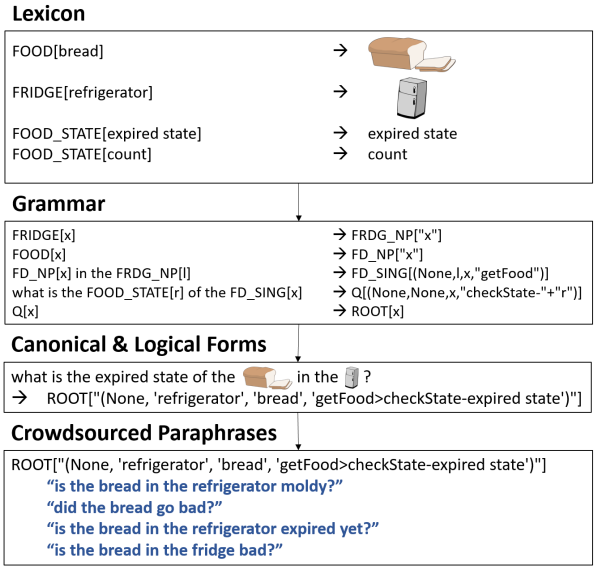

Figure 1: Crowdsourcing pipeline for building semantic parsers for new domains

Wang et al. (2015) propose a methodology for efficient creation of semantic parsing data that starts with the set of target logical forms, and

*The indicated authors contributed equally to this work.

Proceedings of the SIGDIAL 2017 Conference, pages 374–383,
Saarbrücken, Germany, 15-17 August 2017.

generates example natural language utterances for these logical forms. Specifically, the authors of the parser specify a seed lexicon with canonical phrase/predicate pairs for a particular domain, and subsequently a generic grammar constructs canonical utterances paired with logical forms. Because the canonical utterances may be ungrammatical or stilted, they are then paraphrased by crowd workers to be more natural queries in the target language. We argue that this approach has three limitations when constructing semantic parsers for new domains: (1) the seed utterances may induce bias towards the language of the canonical utterance, specifically with regards to lexical choice, (2) the generic grammar suggested cannot be used to generate all the queries we may want to support in a new domain, and (3) there is no check on the correctness or naturalness of the canonical utterances themselves, which may not be logically plausible. This is problematic as even unlikely canonical utterances can be paraphrased fluently.

In this paper, we propose and evaluate a new approach for creating lexically diverse and plausible utterances for semantic parsing (Figure 1.). Firstly, inspired by the use of images in the creation of datasets for paraphrasing (Lin et al., 2014) or for natural language generation (Novikova et al., 2016), we seek to reduce this linguistic bias by using a lexicon consisting of images. Secondly, a generative grammar, which is tailored to the domain, combines these images to form mixed text-image representations. Using these two approaches, we retain many of the advantages of existing approaches such as *ease of supervision* and *completeness* of the dataset, with the added bonus of promoting *lexical diversity* in the natural language utterances, and supporting queries relevant to our domain. Finally, we add a simple step within the crowdsourcing experiment where crowd-workers evaluate the plausibility of the generated canonical utterances. At training time, we conjecture that optionally adding a term to upweight plausible queries might be useful to deploy a semantic parser in real world settings. Encouraging the parser to focus on queries that make sense reduces emphasis on things that a user is unlikely to ask.

We evaluate our method by building a semantic parser from scratch for a dialogue agent in a smart home simulation. The dialogue agent will be capable of answering questions about various sensor activations, and higher-level concepts which map to these activations. Such a task requires understanding the natural language queries of the user, which could be varied and even indirect. For example, in SMARTHOME, *'where can I go to cool off?'* corresponds to the canonical utterance 'which room contains the AC that is in the house?'. Similarly, *'is the temp in the chillspace broke?'* corresponds to 'are the thermometers in the living room malfunctioning?'.

As a result of our analysis, we find that the proposed method of eliciting utterances using image-based representations results in considerably more diverse utterances than in the previous text-based approach. We also find evidence that the SMARTHOME dataset, constructed using this approach, is more diverse than other domain-specific datasets for semantic parsing, such as GEOQUERY or ATIS. We release this dataset to the community[1] as a new benchmark.

2 Example Domain: Smart Home

While our proposed data collection methodology could conceivably be used in a number of domains, for illustrative purposes we choose the domain of a smart home simulation for all our examples. We define a smart home as a home populated with sensors and appliances that are streaming data which can be read. A fully connected dialog agent could reason about and discuss these data streams. Our work attempts to develop a question answering system to support dialogue in this environment.

In the smart home domain, queries could range from complex, such as a user trying to determine the optimal time to start cooking dinner given a party schedule, to simple, asking for a temperature reading. While we believe that many queries could be handled with the methodology that we describe, we have limited the types of queries that can be asked to a reasonable subset, primarily single-turn queries about entity states (for example, *'did I leave the lights in the bedroom on?'* or *'is the dog safe?'*).

3 Approach Overview

Our approach to building a dialog interface for a new domain D, first requires analysis of the domain and identification of the entities involved. This builds on the methodology of Wang et al.

[1] https://github.com/oaqa/resources

(2015), but with three significant additions to elicit diversity and capture domain-relevant queries:

1. An additional step of specifying images for the entities in the domain, and
2. A domain-specific grammar that captures queries relevant to the particular domain.
3. A crowdsourcing methodology that includes crowdworkers annotating canonical utterances for plausibility

After analyzing the domain and the queries we want to support, we construct a seed lexicon and a generative grammar. The generative grammar generates matched pairs of canonical utterances and logical forms. As our seed lexicon contains images, the canonical forms generated are mixed text-image representations. These representations are then shown to workers from Amazon Mechanical Turk[2] to paraphrase in natural language.

3.1 Seed Lexicon

Essential to the goal of reducing lexical bias, is the use of images to describe the entities in the domain. It is beneficial here to choose images which are representative and will be well-understood. The images we used for entities within the SMARTHOME domain are shown in Figure. 3. It is not necessary that all entities be assigned images, in fact it is possible for entities to be named or abstract, and not have any associated images. In these cases, we simply use the natural language description of the image.

We specify a seed lexicon L, consisting of entities e in our domain and associated images (when available) i. Our lexicon consists of a set of rules $\langle e, (i) \rightarrow t[e] \rangle$, where t is a domain type. For our smart home domain, we define possible domain types to be appliances, rooms, food, weather and entities, and their associated subtypes and states (Figure 2.).

3.2 Generative Grammar

Next, we utilize a generative grammar G to produce canonical utterance and logical form pairs (c, z), similar to Wang et al. (2015). Our grammar differs from theirs, in that in our work, the grammar G is not a generic grammar, but is written to generate the kinds of queries we would actually like to support in our domain D. The rules are of the form $\alpha_! \beta_1 \gamma_1 ... \rightarrow t[z]$, where $\alpha\beta\gamma$ are token sequences and t is the domain type. A complete description of our grammar is included in the supplementary material.

3.3 Canonical Utterances and Logical Forms

We generate canonical form - logical form pairs (c, z) exhaustively using the seed lexicon L and grammar G for domain D. This resulted in exactly 948 canonical and logical form pairs in our domain.

The logical formalism we utilize closely corresponds to Python syntax. It consists of functional programs where all questions in our smart-home domain are formulated with the help of a context tree. Each questions is defined as spans over this tree as shown in Figure. 4. The root node of the tree is the environment that we are operating in, and at the surface-level are sensors. These spans are then used to construct a single-line Python statement that is executed against our smart home simulation to retrieve an answer. From this construct, we are able to execute logical forms against the simulation seamlessly after having retrieved them.

3.4 Data Collection Methodology

The next step after forming canonical utterance and logical form pairs, is generating paraphrases for each pair. We use Amazon Mechanical Turk to distribute our data collection task. Over a span of three days, we collected data from nearly 200 Turkers, some of whom participated in the data collection task multiple times.

During the first stage of the task, the Turkers were instructed to paraphrase canonical utterances as naturally as possible, as well as mark the utterances themselves as likely to be asked or not asked. They were also shown a small number of examples, and possible paraphrases. These examples were created using images not present in the lexicon, so as to avoid biasing the Turkers.

In the next stage, the Turkers were asked to enter their paraphrases. Each worker was asked to enter a total of 60 paraphrases over the course of the task. These paraphrases were presented to the worker over 3 pages, with 2 paraphrases per canonical utterance. Turkers were also asked to state if they believed that the question that they were paraphrasing was likely or not. This annotation could subsequently be used for curation, or to bias semantic parsing models towards answers

[2] https://www.mturk.com/mturk/welcome

Lexicon

television | AC | light | humidifier | clock | radio | phone	→ APPLIANCE[television | ...]
refrigerator | stove | dishwasher | toaster | microwave | blender | grill	→ KITCHEN_APPLIANCE[refrigerator | ...]
on| off | malfunctioning | not malfunctioning	→ S_STATE[on | ...]
bedroom | kitchen | livingroom | diningroom| hallway | bathroom | gym | home office	→ ROOM[bedroom | ...]
Bob | Alice | dog | cat	→ ENTITY [Bob | ...]
eggs | bread | milk	→ FOOD[eggs| ...]
expired_state | count	→ FOOD_STATE[expired_state | ...]
rain | sun | wind | snow	→ WEATHER_TYPE[rain | ...]
intensity | duration	→ WEATHER_STATE[intensity | ...]
news | cartoon | comedy	→ TV_PROG_TYPE[news | ...]]
airtime | duration | channel number	→ TV_PROG_STATE[airtime | ...]
safe | hungry | tired	→ ENTITY_STATE[safe | ...]

Figure 2: The lexicon used to generate canonical and logical forms.

Figure 3: Images for terms in the seed lexicon

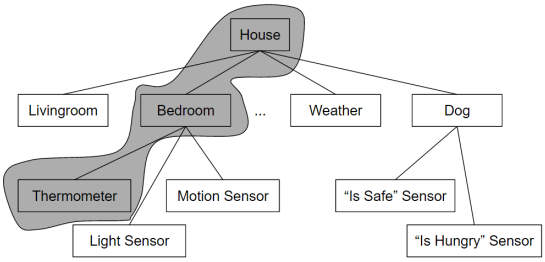

Figure 4: An example of a concept tree that could be used to define the logical form structure.

that users labeled as likely. Most canonical forms had a single image inserted into the text (875 or 92.3%), some had no images inserted into the text (58 or 6.1%), and even fewer had two images inserted into the text (15 or 1.6%). Each logical form was shown to five Turkers for paraphrasing, resulting in approximately ten paraphrases for each logical form.

Finally, we took several post-processing steps to remove improper paraphrases from our dataset. Firstly, a large portion of Turker mistakes arose because of them making real-world assumptions and neglecting to mention locations in their utterances. We automatically shortlisted all paraphrases missing location information. We then manually inspected each of these paraphrases and discarded the ones identified as invalid. In all, this post processing step took less than one day and could have easily been delegated to crowd workers, had it been necessary. Secondly, we automatically pruned all paraphrases in our dataset which were associated with more than one logical form. This left us with 8294 paraphrases.

4 Data Statistics and Analysis

In this section, we describe some statistics of our data set, perform a comparative analysis with the data collection paradigm of existing work, and contrast the statistics of our dataset with other semantic parsing datasets.

4.1 Data Statistics

In its uncurated form, our dataset consists of 10522 paraphrases spread across 948 distinct canonical and logical form pairs. Each pair has a minimum of 10 paraphrases and a maximum of 28 paraphrases. These paraphrases were collected over 195 Turker sessions using the methodology described in the previous section. Following the removal of duplicate paraphrases, and paraphrases missing location information, we are left with 8294 paraphrases over the same 948 logical forms.

4.2 Effect of Data Collection Methodology

We ran an experiment on purely text-based representations as suggested in (Wang et al., 2015) to compare and contrast with our mixed text-image representations. In an effort to subdue domain variance, we utilize our domain-specific grammar

to generate text-based canonical representations. We randomly subsample 100 logical form and canonical utterance pairs from this dataset, and recreate the crowdsourcing experiment suggested by Wang et al. (2015), wherein each canonical utterance is shown to ten Turkers to paraphrase and each Turker receives four canonical utterances to paraphrase. The workers are asked to reformulate the canonical utterance in natural language or state that it is incomprehensible. In this way, we collect 1000 paraphrases associated with the 100 logical forms. For each of these logical forms, we randomly subsample paraphrases from the set gathered using the proposed mixed text-image methodology. We then compare the two and observe the results shown in Table 3. We evaluate the results on three metrics:

Lexical Diversity We estimate the lexical diversity elicited from the two methodologies by comparing the total vocabulary size as well as the type-to-token ratio as shown in Table 1. We find that both the total vocabulary size, as well as the type-to-token ratio of the paraphrases collected using the proposed crowdsourcing methodology is considerably higher than that of an equivalent number of paraphrases collected using the methodology suggested in (Wang et al., 2015).

Lexical Bias We estimate bias by computing the average lexical overlap between the paraphrase generated by the Turker and the canonical utterance they were shown. For the text-image experiment, we consider the equivalent text representation of the canonical utterance, by substituting the images by terms from the lexicon. We find that the proposed crowd sourcing methodology elicits considerably less lexical bias as shown in Table 1.

Relevance We estimate relevance by randomly sampling one paraphrase each for one hundred logical forms using the two methodologies. We then manually annotate them for relevance. Here, relevance is defined as a paraphrase exactly expressing the meaning of the original canonical form.

We performed this analysis on both our final dataset and the the data that was collected in the same manner as described in (Wang et al., 2015). We find that our data set had an estimated relevance of 60% when compared directly with the same random logical forms sampled from the data collected in the manner of (Wang et al., 2015), which had an estimated relevance of 69%.

Randomly sampling from our entire curated dataset, we find that we have an estimated relevance of 66%.

Representation	Vocab Size	TTR	Lexical Overlap
Text (Wang et al., 2015)	291	.044	5.50
Text-Image (ours)	**438**	**.066**	**4.79**

Table 1: Comparison of data creation methodology of (Wang et al., 2015) and this work. 'Vocab size' is the total vocabulary size across an equal number of paraphrase collected for the same logical forms using the two methodologies. TTR represents the word-type:token ratio. Lexical overlap measures the average number of words that are common between the canonical utterances and the paraphrases in the two methodologies.

4.3 Comparison with Other Data Sets

In order to examine the lexical diversity in the original dataset, we examine the ratio of the total number of word types seen in the natural language representations to the total number of token types in the meaning representation. We compare against four publicly accessible datasets:

OVERNIGHT The Overnight dataset (Wang et al., 2015) consists of 26k examples distributed across eight different domains. These examples are obtained by asking crowdworkers to paraphrase slightly ungrammatical natural language realizations of a logical form.

GEO880 Geoquery is a benchmark dataset for semantic parsing (Zettlemoyer and Collins, 2005) which contains 880 queries to a U.S geography database. The dataset is divided into canonical test-train splits with the first 680 examples being used for training and the last 200 examples being used for testing.

ATIS This dataset is another benchmark semantic parsing dataset that contains queries for a flights database, each with an associated meaning representation in lambda calculus. The dataset consists of 5,410 queries and is traditionally divided into 4,480 training instances, 480 development instances and 450 test instances.

Dataset	Example
GEO	how many states border the state with the largest population? answer(A,count(B,(state(B),next_to(B,C),largest(D,(state(C), population(C,D)))),A))
JOBS	what jobs desire a degree but don't use c++? answer(A, (job(A), des_deg(A),+((language(A,C),const(C,'c++')))))
ATIS	what flights from tacoma to orlando on saturday (_lambda $0 e (_and (_flight $0) (_from $0 tacoma : _c i) (_to $0 orlando: _ci) (_day $0 saturday: _da)))
OVERNIGHT	what players made less than three assists over a season (call SW.listValue (call SW.getProperty ((lambda s (call SW.filter (var s) (call SW.ensureNumericProperty (string num_assists)) (string <) (call SW.ensureNumericEntity (number 3 assist)))) (call SW.domain (string player))) (string player)))
SMARTHOME	has the milk gone bad? ROOT["(None, 'refrigerator', 'milk', 'getFood>checkState-expired state')"]

Table 2: Example from datasets GEO, JOBS, ATIS, OVERNIGHT and SMARTHOME

Dataset	NL Types	MR Types	NL/ MR Ratio
GEO	283	148	1.91
ATIS	934	489	1.91
JOBS	387	226	1.71
OVERNIGHT	1422	199	7.14
SMARTHOME (Ours)	1356	83	**16.33**

Table 3: Number of word types in the language compared to number of word types in the logical form. Larger ratio indicates more lexical diversity for the same complexity of the logical form

JOBS The JOBS dataset (Zettlemoyer and Collins, 2005) consists of 640 queries to a job listing database where each query is associated with Prolog-style semantics. This dataset is traditionally divided into 500 examples for training and 140 examples for testing.

An example of the kind of query that can be found in each of these datasets is given in Table 2.

In the analysis, we find that on average SMARTHOME exhibits nearly twice the word type to meaning representation token ratio, as compared to most existing semantic parsing datasets as shown in Table 3.

4.4 Logical Form Plausibility

For each canonical utterance, Turkers were asked to state if the canonical form was 'likely' or 'not likely'. By examining the most polar of these ratings, we see interesting patterns. For example, the canonical form *'what are the readings of the thermometers in the hallway¿* is rated as a highly likely form according to Turkers and does indeed seem like a question that could be asked in the real world. On the other hand, one of the less likely forms according to the Turkers, *'are the televisions in the bathroom on?'*, is indeed not likely, as bathrooms are arguably one of the least likely rooms that one would encounter multiple televisions in. Overall, 752 out of 948 logical forms were identified as very plausible by at least 60% of the Turkers who paraphrased them, indicating they were reasonable questions to ask.

5 Semantic Parsing Experiments

Finally, it is of interest how the data collection methodology influences the realism and difficulty of the semantic parsing task. In this section, we run several baseline models to measure this effect.

5.1 Models

We present three different baselines on our dataset, including a state-of-the-art neural model with an attention-copying mechanism (Jia and Liang, 2016).

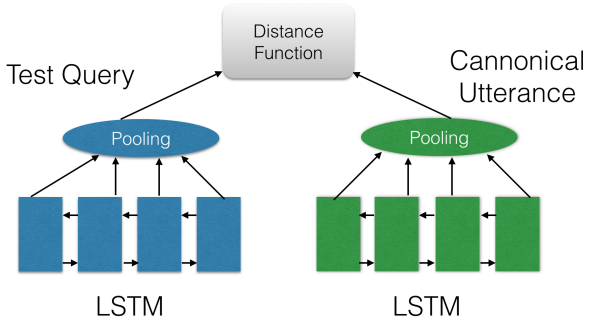

Figure 5: Neural Reranking Model

Jaccard First, we experiment with a simple baseline using Jaccard Similarity which is given by J(A, B) = $\frac{|A \cap B|}{|A \cup B|}$. For each query in the test set, we find the paraphrase in the training set which has the highest Jaccard similarity score with the test query and return its associated logical form.

Neural Reranking Model We next experiment with a neural reranking model for semantic parsing which learns a distribution over the logical forms by means of learning a distribution over their associated paraphrases as a proxy. This model has the added advantage of being independent of the choice of the formal language, and has been used for tasks such as answer selection (Wang and Nyberg, 2015; Tan et al., 2015), but not for semantic parsing. The basic model is shown in Figure. 5. We generate a representation of both the test query and the paraphrasing using a bidirectional-LSTM and use a hinge loss function as specified:

$$L = max(0, M - d(p*, p+) + d(p*, p-))$$

where M is the margin, d is a distance function, p* is the test query, p+ is a paraphrase that has the same meaning representation as p* and p- is a paraphrase that does not. For our experiments, we choose d to be the product of the Euclidean distance and the sigmoid of the cosine distance between the two representations, and M to be 0.05.

We group all the paraphrases by logical form, and create training examples by picking all possible combinations within one grouping as positive samples, and randomly sampling from the remaining top-25 matching paraphrases for negative examples. At test time, we first identify twenty five most likely candidates utilizing a Jaccard-based search engine over the paraphrases in the training data. We then identify the most likely paraphrase from amongst these using the Neural Reranking model.

Neural Semantic Parsing Model We also implement the neural semantic parsing model with an attention-based copying mechanism from (Jia and Liang, 2016). We use the same setting of hyperparameters that gave the best results on GEO, OVERNIGHT and ATIS. Specifically, we run the experiments with 200 hidden units, 100 dimensional word vectors and all the parameters of the network are initialized from the interval [-0.1, 0.1]. We also train the model for 30 epochs starting with a learning rate of 0.1 and halving the learning rate at every 5 epochs from the 15th epoch onwards. We refer the readers to (Jia and Liang, 2016) for further details about the model.

5.2 Results and Discussion

We evaluate these models on independent data in the form of the OVERNIGHT and GEO datasets. We use the standard train-test splits suggested by (Zettlemoyer and Collins, 2005) and (Wang et al., 2015). The full results are presented in Table 4. We observe that the neural semantic parsing model performs relatively poorly on the SMARTHOME dataset compared to OVERNIGHT or GEO. Careful error analysis suggests that most of the errors stem due to the following types of queries in our dataset, which are not present in OVERNIGHT or GEO

- The model not differentiating between the singular and plural forms (For example, *which room in the house can you find the stereo?* maps to the logical form for plural radios instead of the singular)

- The model not recognizing terms which have not been seen in the training data i.e unseen vocabulary (for example, *does bob not have any energy?* does not map to the logical form for checking if Bob is tired, because the model has never seen that to not have energy means being tired for living entities),

- The model not being able to respond to indirect queries in the test set (for example, *how long will the heat have to run?* does not map to the logical form for how long the weather will be cold, or *do i need to change the lights*

System	SMARTHOME(ours)	OVERNIGHT	GEO
Jaccard	18.0%	24.82%	40.7%
Neural Reranker	30.3%	41.91%	60.2%
Seq2Seq (Jia and Liang, 2016)	**42.1%**	75.8%	85.0%

Table 4: Test accuracy results of different systems on the SMARTHOME dataset as compared to OVERNIGHT and GEO

in the living room? does not map to the logical form for the living room lights not working correctly)

- Errors with and between complementary valued variables such as on/off and malfunctioning/not malfunctioning. (For example, *does the tv in the bathroom work?* maps to the logical form for the TV malfunctioning, when it should map to the logical form for the TV not malfunctioning)

We are aware that by accounting for plural nouns, we added a dimension of difficulty for all canonical forms that have a plural/singular sibling which is not present in the datasets which we compare to . We found that 29.7% of the Seq2Seq model's mistakes contained a wrong quantity. Similarly, the smart-home domain includes complementary terms that sometimes form the only difference between two canonical forms (e.g. functioning vs malfunctioning, on vs off). We measure that 43.2% of the Seq2Seq model's errors contain an incorrect complementary term. 9.8% percent contain both a wrong quantity and a wrong complementary term. We conclude that handling plurals and complementary forms makes the task more difficult, particularly as they are often not differentiated well in conversational language. The remaining 36.9% of errors made by the model can largely be attributed to lexical diversity, indirect queries or confusion between entity states.

This work represents a first step in considering lexical diversity as an important criteria while creating semantic parsing datasets. Due to the ambiguity introduced by images (though it is hard to make claims on whether it is ambiguity based only on the interpretation of these images by crowdworkers, or overall difficulty of trying to paraphrase a mixed text-image representation), this could come at the cost of generating slightly less relevant queries. We hope this starts the conversation and inspires further research in finding better ways of introducing lexical diversity.

6 Related Work

Semantic parsing has been used in dialog systems with significant success.(Zhu et al., 2014; Padmakumar et al., 2017; Engel, 2006). Supervised semantic parsing is of special practical interest as while trying to build dialogue systems for new domains, it is important to be able to adapt to domain-specific language. Domains exhibit varied linguistic phenomena and every domain has it's own vocabulary (Kushman and Barzilay, 2013; Matuszek et al., 2012; Tellex et al., 2011; Krishnamurthy and Kollar, 2013; Wang et al., 2015; Quirk et al., 2015). Training a semantic parser for these domains involves understanding the kinds of language used in a domain, however, the cost of supervision of associating natural language with equivalent logical forms is prohibitive.

In an attempt to overcome this overhead of supervision, several approaches have been suggested including learning from denotation-match (Clarke et al., 2010; Liang et al., 2011). As the authors of (Wang et al., 2015) point out, paraphrasing overcomes this overhead by being a considerably lightweight form of supervision. However, methods such as theirs which utilize text induce lexical bias.

Novikova et al. (2016) show that using images reduces this lexical bias for natural language generation tasks. In this work, we unite these strands of research by presenting a methodology where we construct a seed lexicon from images, and use a generative grammar to combine these images into questions, each paired with an associated logical form. These can then be paraphrased by workers from Amazon Mechanical Turk. Our experiment provides evidence that partially replacing canonical form text with images leads

to measurably higher lexical diversity in crowdsourced paraphrases. By contrast to (Wang et al., 2015), we operate only inside a single domain and observe the linguistic patterns specific to the smarthome setting (see Sec 5.2). It remains to be examined whether the observed large increase in diversity can be reproduced in a different domain with different language patterns and colloquialisms. Another immediate research direction, inspired by (Novikova et al., 2016) is replacing more of the canonical form representation with images to further reduce lexical bias and increase variety. This would require the development of a symbol set that is sufficiently expressive while not being overly ambiguous. We anticipate this converging to a tradeoff between the diversity and relevance measures (see Sec 4.2).

7 Conclusion

The primary goal of this paper is to highlight steps to be taken in order to apply semantic parsing in the real world, where systems need robustness against variation in natural language. In this work, we propose a novel crowdsourcing methodology for semantic parsing that elicits lexical diversity in the training data, with the aim of promoting future research in constructing less brittle semantic parsing systems. We utilize combined text-image representations which we believe reduces lexical bias towards language from the lexicon, at the cost of additional ambiguity introduced by the use of images. We find that this crowdsourcing methodology elicits demonstrably more lexical diversity compared to previous crowdsourcing methodologies suggested for creating semantic parsing datasets. The dataset created utilizing this methodology offers unique challenges that result in lower performance of semantic parsing models as compared to standard semantic parsing benchmark datasets. The dataset contains both direct and indirect conversational queries, and we believe that learning to recognize the semantics of such varied queries will open up new directions of research for the community.

Acknowledgments

This research has been supported by funds provided by Robert Bosch LLC to Eric Nyberg's research group as part of a joint project with CMU on the development of a context-aware, dialog-capable personal assistant. The authors are grateful to Bosch Pittsburgh researcher Alessandro Oltramari for valuable insights and productive collaboration. We would also like to thank Alan Black and Carolyn Rose for helpful discussions and Shruti Rijhwani, Rajat Kulshreshtha and Pengcheng Yin for their suggestions on the writing of this paper. We are grateful to all the students from the Language Technologies Institute who helped us test our methodology before releasing it to Turkers including especially Maria Ryskina, Soumya Wadhwa and Evangelia Spiliopoulou. We also thank the anonymous reviewers for helpful and constructive feedback.

References

Amos Azaria, Jayant Krishnamurthy, and Tom M. Mitchell. 2016. Instructable intelligent personal agent. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*. AAAI Press, pages 2681–2689.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*. Association for Computational Linguistics, pages 1533–1544.

S. R. K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 82–90.

James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world's response. In *Proceedings of the fourteenth conference on computational natural language learning*. Association for Computational Linguistics, pages 18–27.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. *CoRR* abs/1601.01280. http://arxiv.org/abs/1601.01280.

Ralf Engel. 2006. Spin: A semantic parser for spoken dialog systems. In *Proceedings of the Fifth Slovenian And First International Language Technology Conference (IS-LTC 2006).*.

Sumit Gulwani and Mark Marron. 2014. Nlyze: Interactive programming by natural language for spreadsheet data analysis and manipulation. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. ACM, New York, NY, USA, SIGMOD '14, pages 803–814.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. volume abs/1606.03622. http://arxiv.org/abs/1606.03622.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics* 1:193–206.

Nate Kushman and Regina Barzilay. 2013. Using semantic unification to generate regular expressions from natural language. North American Chapter of the Association for Computational Linguistics (NAACL).

Percy Liang, Michael I Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 590–599.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423* .

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*. Association for Computational Linguistics, pages 265–273.

Aishwarya Padmakumar, Jesse Thomason, and Raymond J Mooney. 2017. Integrated learning of dialog strategies and semantic parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Chris Quirk, Raymond J Mooney, and Michel Galley. 2015. Language to code: Learning semantic parsers for if-this-then-that recipes. Association for Computational Linguistics.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108* .

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence*.

Di Wang and Eric Nyberg. 2015. Cmu oaqa at trec 2015 liveqa: Discovering the right answer with clues. Technical report, Carnegie Mellon University Pittsburgh United States.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Association for Computational Linguistics*.

William A. Woods. 1977. Lunar rocks in natural English: Explorations in natural language question answering. In *Linguistic Structures Processing*. North Holland, pages 521–569.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*. AAAI Press, AAAI'96, pages 1050–1055.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *In Proceedings of the 21st Conference on Uncertainty in AI*. pages 658–666.

Su Zhu, Lu Chen, Kai Sun, Da Zheng, and Kai Yu. 2014. Semantic parser enhancement for dialogue domain extension with little data. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, pages 336–341.