Semantic Parsing of Brief and Multi-Intent Natural Language Utterances

Logan Lebanoff[1] **Charles Newton**[1] **Victor Hung**[1]
Beth Atkinson[2] **John Killilea**[2] **Fei Liu**[3]

[1]Soar Technology, Inc. [2]U.S. Navy [3]University of Central Florida

{logan.lebanoff, charles.newton, victor.hung}@soartech.com
{beth.atkinson, john.killilea}@navy.mil
feiliu@cs.ucf.edu

Abstract

Many military communication domains involve rapidly conveying situation awareness with few words. Converting natural language utterances to logical forms in these domains is challenging, as these utterances are brief and contain multiple intents. In this paper, we present a first effort toward building a weakly-supervised semantic parser to transform brief, multi-intent natural utterances into logical forms. Our findings suggest a new "projection and reduction" method that iteratively performs projection from natural to canonical utterances followed by reduction of natural utterances is the most effective. We conduct extensive experiments on two military and a general-domain dataset and provide a new baseline for future research toward accurate parsing of multi-intent utterances.

1 Introduction

Semantic parsing to map a natural language utterance to its logical form is regarded as a challenging task partly due to a lack of annotated data (Berant and Liang, 2014; Yin et al., 2018; Gardner et al., 2018). A promising avenue of research is to generate a set of candidate logical forms paired with their canonical realizations in natural language. Then, the canonical utterance that best matches the input is identified by a model, and its logical form is used as output (Berant and Liang, 2014). A paraphrase/sequence-to-sequence model may additionally be used to translate a canonical utterance to a logical form (Wang et al., 2015; Herzig and Berant, 2019; Cao et al., 2020; Marzoev et al., 2020). While the results are promising, most existing works do not handle natural language utterances with multiple intents. We refer to an *intent* as a goal intended by a user's utterance. Multi-intent utterances allow people to communicate core aspects of a situation in a consistent and timely manner, as illustrated in Figure 1.

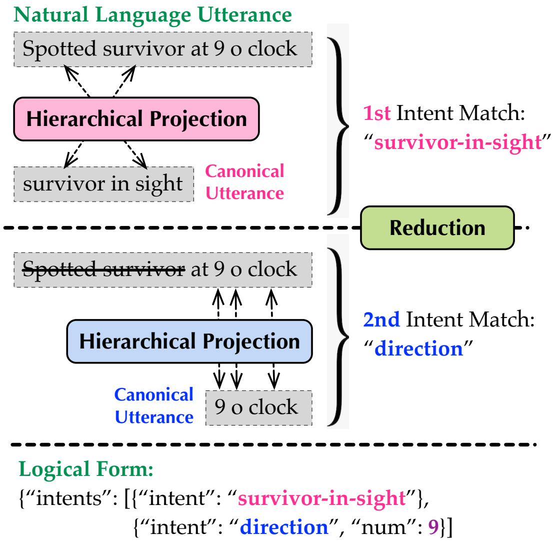
Natural Language Utterance
Spotted survivor at 9 o clock
Hierarchical Projection
survivor in sight
Canonical Utterance
1st Intent Match: "survivor-in-sight"
Reduction
~~Spotted survivor~~ at 9 o clock
Hierarchical Projection
Canonical Utterance
9 o clock
2nd Intent Match: "direction"
Logical Form:
{"intents": [{"intent": "survivor-in-sight"},
{"intent": "direction", "num": 9}]

Figure 1: Example generated semantic parse. A parser must (1) understand paraphrases of canonical utterances and (2) parse multiple intents in one utterance.

Multi-intent semantic parsing is especially suitable for military domains where emphasis is placed on communication skills, terminology, and brevity (Weinstein, 1990). While communication protocols are often published, variations are allowed given the current situation. An area of interest is Intelligence, Surveillance, and Reconnaissance (ISR) domains where contact reports (e.g., "*Arriving at home base and ready to descend*") often contain multiple intents, and a system must determine the number of intents, interpret the natural language and predict the exact logical forms for every intent, which can be highly challenging.

We investigate new methods for semantic parsing of utterances with multiple intents. Importantly, and distinguishing our work from earlier literature (Iyer et al., 2017; Zhong et al., 2017; Yu et al., 2018; Dong and Lapata, 2018; Zeng et al., 2020), our domain areas have no supervised training data, nor can pseudo-language utterances be created through crowdsourcing due to their sensi-

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 255–262
April 20, 2021. ©2021 Association for Computational Linguistics

tive nature and requirement of expert knowledge. We thus operate in a weakly-supervised setting by assuming only access to a grammar that generates canonical utterances and logical forms. Obtaining a comprehensive collection of natural utterances for military applications is difficult; it can be easier to create a grammar that generates canonical utterances for the application. In addition, there are scenarios where there is insufficient time or funding to obtain supervised data, e.g. quickly building a virtual assistant for a new mobile app. Our goal is distinct from related efforts in dialog systems (Gupta et al., 2018; Vanzo et al., 2019; Lee et al., 2019; Ham et al., 2020); the parser does not have additional context or interaction but focuses on modeling complex compositional intents. We build on methods that project natural utterances to the canonical space (Marzoev et al., 2020) and investigate novel adaptations for handling multi-intent utterances. Our contributions are as follows.

- We present a first effort at parsing brief, multi-intent utterances into logical forms; this work sheds light on parsing of airborne communications for which parallel resources are limited.
- We perform experiments on two military communications datasets and a general-domain dataset. Our findings suggest that a new approach that iteratively projects the natural language utterance to a canonical utterance, followed by a reduction step can achieve the best performance.

2 Hierarchical Projection

Let X and Y be the set of all natural language utterances and logical forms (LF), respectively. Given a natural language utterance $x \subseteq X$, we wish to produce $y \subseteq Y$. We assume only access to a grammar G that defines a set of n production rules whose union forms a canonical set of utterances Z. A grammar is assumed to be in the form $G = R_1|\ldots|R_n$, where each production rule $R_i \rightarrow (\alpha, \tau)$ is defined as rule expansion α and a tag τ. Tags define the semantic content associated with a rule, which can be used to build LFs. Figure 2 shows an example grammar. Canonical utterances found in Z do not cover the full range of variation available in natural language. A viable option, described below, is to develop a projection function π which maps X directly into Z and obtain an appropriate y through G.

We follow Marzoev et al. (2020) and use a pre-trained language model (LM) to obtain semantic representations of utterances in $\mathbb{R}^d$. A distance function δ is used to compute the closest canonical utterance in vector space to the natural language utterance. The projection function is defined as:

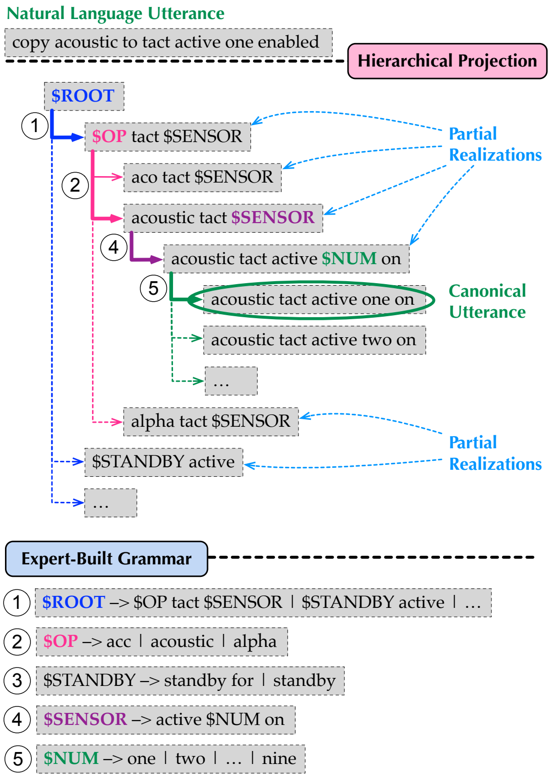
Natural Language Utterance
copy acoustic to tact active one enabled
Hierarchical Projection
$ROOT
1
$OP tact $SENSOR
2
aco tact $SENSOR
acoustic tact $SENSOR
4
acoustic tact active $NUM on
5
acoustic tact active one on
acoustic tact active two on
...
alpha tact $SENSOR
$STANDBY active
...
Partial Realizations
Canonical Utterance
Partial Realizations
Expert-Built Grammar
1 $ROOT –> $OP tact $SENSOR | $STANDBY active | ...
2 $OP –> acc | acoustic | alpha
3 $STANDBY –> standby for | standby
4 $SENSOR –> active $NUM on
5 $NUM –> one | two | ... | nine

Figure 2: (Top) Method to project a natural language utterance to a canonical utterance; the logical form can then be inferred directly. (Bottom) Grammar from our ISR data.

$$\pi(x) = \arg\min_{z \in Z} \; \delta(\mathrm{LM}(x), \mathrm{LM}(z)) \qquad (1)$$

$\mathrm{LM}(\cdot)$ is calculated as the average of BERT-Base (Devlin et al., 2018) representations, and δ is cosine similarity. Computing the $\arg\min$ requires $O(Z)$ operations which can be intractable for many grammars. To handle this, we use a hierarchical projection method by performing a search through the grammar (Marzoev et al., 2020). The $\$root$ is expanded by taking one step in the grammar to yield several partial instantiations z', which takes the place of z in Eq. 1. We refer to a partial instantiation z' as a canonical utterance that still contains non-terminals. The z' closest to x is chosen in the next search iteration. Non-terminals in z' are expanded until only terminals remain (Figure 2).

3 Adaptation

3.1 Non-Terminal Averaging

Vector representations of partial instantiations introduce difficulty as non-terminals are not well-

	Single-Intent		Multi-Intent		Avg
Dataset	**Can.**	**NL**	**Can.**	**NL**	**Len**
ISR	445	790	20,000	600	7.4
HELI	45	170	8,000	2,000	3.0
OVERNIGHT	302	2,416	20,000	2,000	10.8

Table 1: Number of canonical (*Can.*) and natural language utterances (*NL*) and average length of utterances. Each *Can.* and *NL* utterance is paired with a gold standard LF. *Can.* pairs are used for training and *NL* pairs are for evaluation. OVERNIGHT numbers are averaged over its eight subdomains.

understood by pre-trained LMs. This can be somewhat resolved by replacing non-terminals with the [MASK] token (Marzoev et al., 2020). It conveys to the LM that a word or phrase should exist at that position, but it's not clear yet what exactly belongs there, and allows the LM to form representations for the other tokens with the knowledge that something will exist there. However, partial instantiations may contain few or no terminals at all, meaning LM input will be dominated by [MASK] tokens. For example, for the given partial instantiation – *$TypeNP whose $RelNP is $EntityNP* – it is not clear what values may be used for the non-terminals, and the resulting utterance representation will not be useful.

We introduce a strategy to mitigate this issue that we call non-terminal averaging. We observe that a non-terminal is restricted to certain values defined by the grammar. We obtain a representation of the non-terminal by averaging over the representations of these possible values, which gives a much better representation than the [MASK] token. This is important when projecting over multiple intents, as discussed in the next section.

3.2 Multi-Intent Projection

We explore two methods for parsing utterances with multiple intents.

Meta-Grammar A simple method for handling multiple intents is to create a meta-grammar based on the original grammar. The $\$root$ is renamed to $\$subroot$ while keeping all other rules unchanged. A new $\$root$ is created with the rule $\$root \rightarrow \$subroot \mid \$subroot\ \$subroot \mid \ldots$ It encapsulates all combinations of multiple intents, where each combination is a concatenation of ≥ 1 intents.

Reduction Another approach is to first greedily project to the closest canonical utterance, remove all tokens in the input utterance that appear in the canonical utterance, and repeat to find another similar canonical utterance. This iterative process of projection and reduction is repeated until no tokens remain or a continuation threshold is met. Figure 1 presents an example.

To more accurately perform token removal for our *Reduction* method, we compare BERT representations of tokens rather than comparing exact string matches between tokens, similar to BERTScore (Zhang et al., 2019). If the cosine similarity scores between two tokens meet a certain similarity threshold, then those two tokens are treated as equivalent, and the token will then be removed. This technique can better handle slight variations in word choice (e.g. "survivor" and "survivors", or "spotted" and "in sight"). We used 0.5 as the similarity threshold, but the model is not very sensitive to this value.

4 Data

Most semantic parsers are given access to (natural language utterance, LF) pairs during training. Our setting, however, assumes no access to these pairs and are only given a grammar to generate canonical utterances and their LFs. We use two proprietary military communication datasets and a general-purpose dataset OVERNIGHT (see Table 1).

ISR Intelligence, surveillance, and reconnaissance (ISR) subject matter experts were consulted to develop a corpus of known utterances that an intelligence operator would say during a mission. The LFs appear in JSON format containing an intent and slots to be filled. The utterances are consolidated into a grammar with a relatively deep structure where an intent may contain slots for nested intents, making it closer to semantic parsing datasets like TOP (Gupta et al., 2018). It consists of domain-specific words and acronyms outside of ordinary vernacular, making this dataset particularly challenging. The grammar contains 37 rules, 36 non-terminals, and approximately 60 terminals.

HELI Short commands were collected from helicopter communications and consolidated into a similar grammar to ISR. The grammar has a shallow structure and does not contain many nested intents, but each utterance is short (1–5 tokens), which has its own challenges. The grammar contains 48 rules, 47 non-terminals, and approximately 60 terminals. Example in Fig. 1.

Natural language utterances in both datasets are wholly defined by its grammar. For evaluation,

		ISR	HELI
Single Intent	seq2seq (Lewis et al., 2020)	34.4	58.2
	proj (Marzoev et al., 2020)	82.5	**74.1**
	NT-Avg	**85.9**	**74.1**
Multi Intent	seq2seq (Lewis et al., 2020)	48.1	**45.5**
	NT-Avg + MetaGrammar	16.3	25.2
	NT-Avg + Reduction	**49.1**	38.8

Table 2: Logical form accuracies for internal ISR and HELI datasets

we expand to a set of paraphrased canonical utterances using an English-to-$X \rightarrow X$-to-English procedure similar to those used for augmentation in paraphrase datasets (Wieting and Gimpel, 2018; Hu et al., 2019).

OVERNIGHT (Wang et al., 2015) is a semantic parsing dataset over eight domains, including sports, restaurants, and social media. Each domain contains a grammar to generate canonical utterances and LFs, as well as natural language paraphrases. As we are interested in weakly-supervised parsing, we ignore natural language utterances in training and only use those in the test set for evaluation. The datasets we use contain grammars and natural language data for utterances with a single intent, but they lack multi-intent data. We create simulated multi-intent utterances by concatenating natural language utterances together, with target LFs as concatenations of the utterances' LFs. We enforce a limit of three intents to keep task difficulty manageable.

5 Results

We present several baselines in our experiments. We train a sequence-to-sequence (*seq2seq*) model on sequence pairs of the form (canonical utterance, LF). At evaluation, the model is given a natural language utterance associated with a canonical utterance and evaluated based on the original LF. We make use of pre-trained BART (Lewis et al., 2020) by fine-tuning on task-specific data. *Proj* is the technique of projecting a natural language utterance to a canonical utterance in the grammar, described in Section 2. *NT-Avg* is the proposed method of averaging the representations of a non-terminal's possible values. For the single-intent OVERNIGHT datasets, we display baseline results presented in Marzoev et al. (2020). Finally, we experiment with two methods on top of NT-Avg for multi-intent parsing – *Meta-Grammar* and *Reduction*.

Table 2 presents LF exact match accuracies for our internal datasets in both single-intent and multi-intent settings. We observe that projection techniques outperform seq2seq methods for single-intent, consistent with prior work (Marzoev et al., 2020). Our proposed method (NT-Avg) achieves a sizeable improvement in ISR, but equal performance on HELI. This disparity may be due to HELI's shallow grammar, demonstrating that non-terminal averaging provides gains on domains with deep, hierarchical grammars but less on simple grammars. For multi-intent, Reduction outperforms MetaGrammar by a wide margin. MetaGrammar must simultaneously predict the number, type, and location of intents. Reduction iteratively simplifies the process by searching for one intent at a time. We also observe that seq2seq achieves much stronger performance for multi-intent with similar accuracy to Reduction on ISR and achieving much higher accuracy on HELI. We believe the improvement is due to the larger amount of data available to train seq2seq models, since we can concatenate multiple single-intent canonical utterances together to form large simulated training sets.

The OVERNIGHT dataset contains a more complex grammar and longer utterances and LFs compared to our internal datasets (Table 3). NT-Avg outperforms other approaches on single-intent utterances, similar to results on ISR and HELI.

All systems evaluated on the multi-intent split of OVERNIGHT struggle to perform well. A system must be able to determine the number of intents in an utterance, interpret the natural language in each intent, and predict LFs that exactly match the LFs for every intent. Accuracies range between 0% and 2% (see supplementary). This demonstrates that it is non-trivial to transfer parsing systems from the single-intent setting to multi-intent. To tease out performance differences between systems, we instead evaluate a system prediction to be correct if *at least one* predicted LF has an exact match with any one of the gold standard LFs.

For multi-intent utterances, Reduction achieves the highest accuracies. We believe the long structure of the LFs in OVERNIGHT provide a challenge for current seq2seq models to generate accurately. Meanwhile, grammar-based approaches can easily side-step this issue by producing LFs directly from the grammar, evidenced by the higher accuracies.

An additional phenomenon appearing in all datasets is lower layers of BERT used for projection perform better than higher layers (Figure 3). How-

		Bas	Blo	Cal	Hou	Pub	Rec	Res	Soc	Avg
Single Intent	seq2seq+BERT (Marzoev et al., 2020)	**60.0**	21.0	31.0	31.0	34.0	36.0	36.0	**31.0**	35.0
	proj (Marzoev et al., 2020)	47.0	27.0	32.0	36.0	34.0	49.0	43.0	28.0	37.0
	NT-Avg	32.7	**37.0**	**42.2**	**48.6**	**51.5**	**52.3**	**56.3**	30.0	**43.8**
Multi Intent	seq2seq (Lewis et al., 2020)	18.0	16.0	11.5	7.5	23.0	21.5	27.5	7.0	16.5
	NT-Avg + MetaGrammar	18.0	11.4	22.9	16.9	21.1	40.0	23.6	22.3	22.0
	NT-Avg + Reduction	**31.7**	**24.2**	**38.8**	**36.4**	**41.5**	**57.7**	**42.0**	**23.4**	**36.9**

Table 3: Logical form accuracies against OVERNIGHT datasets. Partial accuracies are reported for multi-intent data.

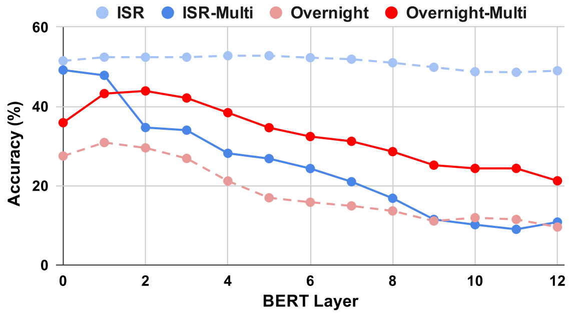

Figure 3: LF accuracies for NT-Avg system by varying BERT layer. Lower layers result in higher accuracies, especially in the multi-intent setting.

ever, we notice that layers 0-1 achieve higher accuracies in ISR and HELI, while layers 1-3 achieve higher accuracies in OVERNIGHT. We believe this is due to the role that context plays in each domain. In terse military domains, words often carry unambiguous meaning and require little context to understand. In traditional domains, context is required to interpret the meaning of a word.

6 Conclusion

We tackle multi-intent semantic parsing using weakly-supervised methods. Our results show that an iterative approach of projecting the natural utterance to a canonical utterance followed by a token reduction step achieves the best performance. Potential further improvement could be achieved by fine-tuning the BERT model on free text in the desired domain (e.g. military training materials) to create better utterance embeddings. Future research includes parsing more complex multi-intent utterances, borrowing ideas from dialogue systems and capturing dependencies between intents (Gangadharaiah and Narayanaswamy, 2019).

7 Ethics and Broader Impacts

Military Applications It is vital for military personnel to use precise language in the field to minimize confusion. This work is part of an effort to train operators of specialized military equipment to accurately communicate in search-and-rescue team and aircraft management operations. Improvement in these occupations leads to better airspace safety and rescue outcomes.

Broader Impacts This work has a larger societal impact outside of military domains. For example, natural language understanding systems in healthcare require the use of audio data or transcripts of patient interactions, and the collection of this sensitive data has major ethical considerations. Our technology is flexible enough to be used in these specialized domains without the need for training on sensitive data and thus has a positive impact in the healthcare field. Potential misuses of this technology, however, could lead to decreased privacy for individuals whose voice is recognized.

Environmental Impact As stated in the paper, our models do not require any training, which greatly reduces the number of computations and thus lessens the environmental impact of natural language technology. Instead our models are based on pre-trained language models used in an unsupervised manner, so the only computation time comes from inference and experiments.

Acknowledgments

This research is based upon work supported by the Naval Air Warfare Center Training Systems Division and the Department of the Navy's Small Business Innovation Research (SBIR) Program, contract N68335-19-C-0052. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Department of the Navy or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland. Association for Computational Linguistics.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6806–6817, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. 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 564–569, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Pradeep Dasigi, Srinivasan Iyer, Alane Suhr, and Luke Zettlemoyer. 2018. Neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–18, Melbourne, Australia. Association for Computational Linguistics.

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*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Jonathan Herzig and Jonathan Berant. 2019. Don't paraphrase, detect! rapid and effective data collection for semantic parsing. 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 3810–3820, Hong Kong, China. Association for Computational Linguistics.

J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.

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 (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada. Association for Computational Linguistics.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. ConvLab: Multi-domain end-to-end dialog system platform. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 64–69, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Alana Marzoev, Samuel Madden, M Frans Kaashoek, Michael Cafarella, and Jacob Andreas. 2020. Unnatural language processing: Bridging the gap between synthetic and natural language data. *arXiv preprint arXiv:2004.13645*.

Andrea Vanzo, Emanuele Bastianelli, and Oliver Lemon. 2019. Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 254–263, Stockholm, Sweden. Association for Computational Linguistics.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342, Beijing, China. Association for Computational Linguistics.

Clifford J. Weinstein. 1990. Opportunities for advanced speech processing in military computer-based systems*. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. Structvae: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–765.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Jichuan Zeng, Xi Victoria Lin, Caiming Xiong, Richard Socher, Michael R Lyu, Irwin King, and Steven CH Hoi. 2020. Photon: A robust cross-domain text-to-sql system. *arXiv preprint arXiv:2007.15280*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

	Bas	Blo	Cal	Hou	Pub	Rec	Res	Soc	Avg
Multi-Intent									
seq2seq (Lewis et al., 2020)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NT-Avg + MetaGrammar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NT-Avg + Reduction	**0.35**	**0.35**	**0.35**	**1.30**	**1.30**	**1.95**	**1.75**	**0.30**	**0.96**

Table 4: Exact match logical form accuracies against OVERNIGHT multi-intent datasets

A Model Details

We use BART-base (Lewis et al., 2020) to closely match the number of parameters and amount of pre-training data used by BERT-base (Devlin et al., 2018), which is used for the projection approaches. BART-base uses the Transformer encoder-decoder architectures with 6 layers in the encoder and decoder, 12 attention heads in the encoder and decoder, and hidden size of 768. We train with a batch size of 4, optimized with Adam, a learning rate of 4e-5. The model converged after an average of five epochs for the OVERNIGHT single-intent datasets and one epoch for multi-intent. The model took longer to converge on the ISR and HELI datasets taking 20 epochs and 40 epochs, respectively. This is likely because of the unfamiliar military terms and terse utterances. A beam size of 10 is used for all projection techniques (including the proposed approaches).

Our results for *proj* differ from those presented in (Marzoev et al., 2020) because we use the hierarchical projection method, which forces a search through the grammar to find the closest canonical utterances. Marzoev et al. (2020) use a linear projection method, which instead compares to all canonical utterances directly, which generally performs better but is not tractable for complex grammars.

B Full Logical Form Results

All semantic parsing systems that we evaluated OVERNIGHT struggle to perform well on parsing multi-intent utterances. It can be difficult to simultaneously determine the number of intents in an utterance, interpret the natural language in each intent, and predict LFs that exactly match the LFs for every intent. Table 4 presents the exact match logical form accuracies for OVERNIGHT multi-intent.