

Towards Zero-Shot Frame Semantic Parsing with Task Agnostic Ontologies and Simple Labels

Danilo Ribeiro^{1,2,*}, Omid Abdar^{1,*}, Jack Goetz¹, Mike Ross¹, Annie Dong¹,
Kenneth Forbus², Ahmed Mohamed¹

¹Meta AI, {jrgoetz, mikeross, asyd, ahmedkm}@meta.com,

²Northwestern University, {dnribeiro, forbus}@u.northwestern.edu,

Abstract

Frame semantic parsing is an important component of task-oriented dialogue systems. Current models rely on a significant amount training data to successfully identify the intent and slots in the user’s input utterance. This creates a significant barrier for adding new domains to virtual assistant capabilities, as creation of this data requires highly specialized NLP expertise. In this work we propose OpenFSP, a framework that allows for easy creation of new domains from a handful of simple labels that can be generated without specific NLP knowledge. Our approach relies on creating a small, but expressive, set of domain agnostic slot types that enables easy annotation of new domains. Given such annotation, a matching algorithm relying on sentence encoders predicts the intent and slots for domains defined by end-users. Experiments on the TopV2 dataset shows that our model trained on these *simple labels* have strong performance against supervised baselines.

1 Introduction

Frame semantic parsing is an important sub-problem with many applications, and in particular is critical for task-oriented dialogue assistants to identify the desired action (intent) and specific details (slots) (Coucke et al., 2018; Gupta et al., 2019). This is typically modeled as a semantic parsing problem (usually solved via a combination of ML and rules) with custom ontology that reflects capabilities of the system. Creation of this custom ontology, and annotation of consistent data is highly non-trivial, and typically requires specialized skills (Ahmad et al., 2021). This limits extension of the ontology and generation of parsing data to a small group of experts.

On the other hand, intent-slot concepts map well to functions and arguments of an API call,

*These authors contributed equally.

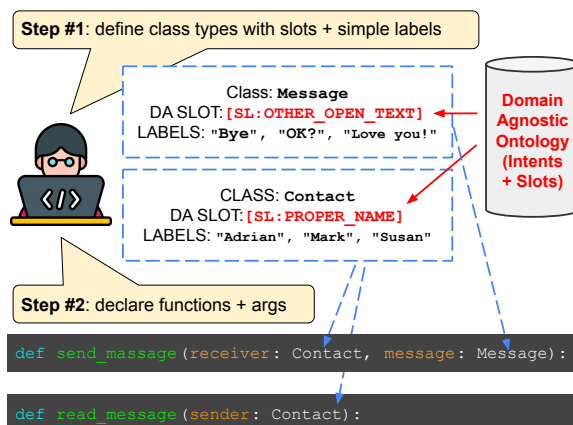


Figure 1: Illustration of proposed OpenFSP framework with a domain agnostic ontology and simple labels provided by the software developer (natural language textual examples). OpenFSP can facilitate the development of new domains and automatic construction of new ontologies by decoding functions or API specifications.

a paradigm well understood by software developers. Therefore, making extension to the parser is the primary blocker to enabling support for new capabilities (domains) within a task-oriented assistant system. As shown in Figure 1, our goal is to enable non-NLP experts to define allowed intent-slot combinations, and provide a small amount of non-NLP specialized labels, which we call *simple labels*. These data enable the creation of a parser for those intent-slot combinations. This new problem definition lies somewhere between zero-shot and few-shot learning. It requires zero fully annotated semantic parse examples, but does require some human produced labels.

To this end, we develop a framework called *Open Frame Semantic Parser* (OpenFSP). OpenFSP takes as input the developer’s defined functions and their annotations to augment an existing assistant system for new tasks. Underlying OpenFSP is a two module model consisting of a *general semantic parser* and a *matching module*. The gen-

eral semantic parser can identify the intent and slots according to the pre-defined domain agnostic ontology, while the matching module will take this intermediate representation and match them to the specific function and arguments defined in the domain specific ontology.

In summary, our contributions are: (1) we formalize a new framework, namely OpenFSP, that allows for easy development of new domains without much expertise knowledge from software developers (2) we define a general-purpose domain agnostic ontology by analysing the semantic similarity of slots from TOPV2 (Chen et al., 2020b), a well established task-oriented semantic parsing dataset (3) we propose an approach consisting of a parser and a matching module that can outperform strong baselines in the *simple labels* setting.

2 Related Work

Data-Efficient Semantic Parsing In one of the first attempts to use data-efficient methods to perform frame semantic parsing, (Bapna et al., 2017) applied recurrent neural networks to perform semi-supervised intent classification and slot filling (IC/SF) while leveraging natural language descriptions of slot labels. The work of Krone et al. (2020) more clearly formalized few-shot learning for IC/SF, while providing a few-shot split of three public datasets. Wilson et al. (2019) implemented and deployed a kiosk virtual assistant, which could handle multiple modalities by learning from a few examples using analogical learning.

Multiple low resource IC/SF approaches were proposed (Chen et al., 2020b; Desai et al., 2021; Yin et al., 2022; Basu et al., 2022). All of these approaches either rely on a non-trivial amount of training data (hundreds to thousands of examples), or use a fixed set of intents and slots, making it harder to adapt to new domains. Our matching module shares some similarities with retrieval based systems (Yu et al., 2021; Shrivastava et al., 2022), however these methods learn from standard utterance and semantic frames, instead of simple textual labels for each slot and intent. This is an important differentiator, as the annotation of even a small number of semantic frames requires overcoming a significant knowledge barrier.

Sentence Encoders for Language Understanding One key aspect of our matching module is to encode the textual spans using sentence encoders. These models have the advantage of working well

in low resource settings, and have been used in many applications including natural language inference (Conneau et al., 2017), semantic textual similarity (Reimers and Gurevych, 2019), dense passage retrieval (Karpukhin et al., 2020; Izacard and Grave, 2021), natural language explanations (Neves Ribeiro et al., 2022), and many others. The idea of using sentence embeddings for text classification has been previously explored (Perone et al., 2018; Piao, 2021). Most notably, the work of Tunstall et al. (2022) proposed SETFIT, a prompt-free model that can learn text classification tasks from a handful of examples. However, none of these works directly applied to semantic parsing which require multiple consistent predictions from the input utterance.

Task-Oriented Ontologies Task-oriented dialogue systems are natural language interfaces for a system that parses user utterances. A common approach in these systems is the use of ontologies to represent domain information (Wessel et al., 2019). In general, such ontologies can be created manually using rule-based techniques, which is a highly accurate but time consuming and expensive approach that usually requires domain experts (Meditkos et al., 2020), or via ontology induction approaches using machine learning (Poon and Domingos, 2010). While these approaches both involve trade-offs, curated ontologies can also be created via simplification of an existing ontology based on custom needs (Kutiyanawala et al., 2018; Laadidi and Bahaj, 2018). On the other hand, our work simplifies the ontological creation by abstracting the existing functions and arguments defined by the application itself.

3 Problem Definition

The standard frame semantic parsing has two main tasks which maps the input utterance x with tokens x_1, \dots, x_n to some structured output frame y . For the *slot-filling* task, the output frame F consists of a set of m non-overlapping spans and their respective labels $F = \{(s_i, e_i, l_i)\}_{i=1}^m$ indicating that subsequence $x_{[s_i:e_i]}$ has label $l_i \in \mathcal{L}$, where \mathcal{L} is the set of possible labels (e.g., the text span “noon tomorrow” has the label `SL:DATE_TIME`). The *intent classification* assigns a label to the whole utterance. For simplicity, we assume that the intent can be thought of as another slot filling with $s_i = 0$ and $e_i = n$, with l_i as the intent type.

To simplify annotation efforts, we define an on-

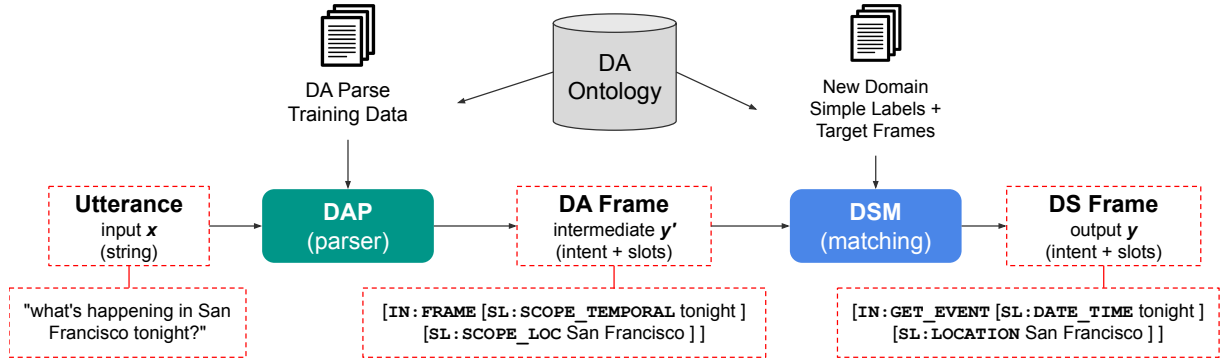


Figure 2: System overview with two components, namely the *domain agnostic parser* (DAP) and the *domain specific matching* (DSM).

ontology with a set of domain-agnostic labels \mathcal{L}_A , where $|\mathcal{L}_A| \ll |\mathcal{L}|$. These domain-agnostic labels can be interpreted as generic “label types”, with an existing many-to-one mapping $\psi : \mathcal{L} \rightarrow \mathcal{L}_A$ between the two sets. This mapping is further described in Appendix A.1. In our proposed simple label setting, training data for new domains will not include x , but only a subsequence of the tokens defined by the frame spans. The number of examples for each slot will be relatively small, with 5 to 50 examples per slot type. The set of all possible domain specific target frames F^* (i.e., defined functions and their arguments) is assumed to be known a priori.

4 Approach

Our model is comprised of two main components, the *domain agnostic parser* (DAP) and the *domain specific matching* (DSM). The system overview is shown in Figure 2. The DAP is trained to take the input utterance x and output a domain agnostic frame F_A , where span labels belong to \mathcal{L}_A . Afterwards the DSM module will score the potential domain specific frames from F^* according to their similarity to the domain agnostic frame F_A .

The DAP module can be any frame semantic parser. Ours is built on a span-pointer network (Shrivastava et al., 2021), which is a non-autoregressive parser that replaces the decoding from text generation with span predictions. It is trained on domain agnostic data obtained from existing domain specific data using the slot label function ψ .

The DSM module has to learn a similarity function between a text span and its slot label (same applies to intents). Since the slot label scoring function has to be done in a few-shot setting, the

DSM module uses a sentence encoder ϕ (Reimers and Gurevych, 2019) with a classification head on top. The module modifies a pre-trained transformer language model fine tuned to output semantically meaningful sentence embeddings.

The sentence encoder ϕ is tailored to work on generic text, minimizing the distance between semantically similar sentences while maximizing the distance of dissimilar sentences. We use a classification head H over the produced sentence embedding of a given text span x . Therefore, the probability score of a label l_i can be computed as follows:

$$P(l_i | \mathbf{x}) = \frac{\exp(H(\phi(\mathbf{x})))_i}{\sum_{j=1}^{|\mathcal{L}|} \exp(H(\phi(\mathbf{x})))_j} \quad (1)$$

Note that the input training data does not contain fully annotated domain specific frames (only text and their intent-slot labels). For this reason, the DSM module has to aggregate these individual intent and slot probability scores to predict the full frame scores. The similarity score for a target domain specific frame $F = \{(s_i, e_i, l_i)\}_{i=1}^{|F|}$ given a query domain agnostic parse $F_A = \{(s_i^A, e_i^A, l_i^A)\}_{i=1}^{|F_A|}$ is given by:

$$\text{sim}(F, F_A) = \max_{F' \in \mathfrak{S}(F)} \frac{1}{|F'|} \sum_{i=1}^{|F'|} P(l_i | \mathbf{x}_{[s_i^A : e_i^A]}) \quad (2)$$

Where $\mathfrak{S}(F)$ is the set of all slot permutations of F . To make the final prediction, we also check that the typing between domain agnostic and domain specific types match. Therefore DSM selects the best frame $F \in F^*$ using the formula:

$$\arg \max_{F \in F^*} \begin{cases} 0 & \text{if } \exists i : \psi(l_i) \neq l_i^A \\ \text{sim}(F, F_A) & \text{otherwise} \end{cases} \quad (3)$$

Note that for each new domain, only a very small set of parameters are trained (namely, the classification head H), which makes this approach computationally efficient.

4.1 Data

We conduct experiments using the TopV2 (Chen et al., 2020b), a multi-domain dataset containing intent and slot annotations for task-oriented dialogue systems. We select a subset of the data points, ignoring the ones containing more than one intent (even though our method would also work for nested frames) or utterances with IN:UNSUPPORTED intent label. The final dataset contains a total of eight different domains (namely: alarm, event, messaging, music, navigation, reminder, timer, and weather). The dataset is then split into train (104278), evaluation (14509) and test (32654) sets.

5 Experiments

5.1 Evaluation Setup

We conduct experiments to evaluate how well a model can adapt to a new unseen domain by leveraging the “simple labels” for this new domain. To simulate this adaptation process we perform multiple test rounds, where each of the eight TopV2 domains are treated as unseen domains. We use *frame accuracy* as a metric, which refers to the ratio of examples where the system correctly predicts all the frame’s spans and labels. The results for each left-out domain are averaged to obtain the final metrics. Each label $l_i \in \mathcal{L}$ of the unseen domain will be assigned a few textual examples that will be used for training. In our experiments we use 5, 10 and 50 examples per unseen domain’s label.

5.2 Baselines

The first two methods are used as baselines and can be seen as soft upper bounds. They do not follow the “simple labels” settings, and use the full training data for the new domains instead. Other two methods are used to further evaluate the different implementation choices of the model, as ablation studies.

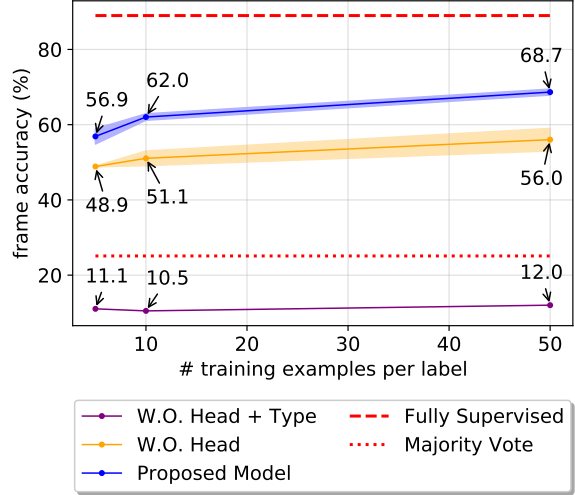


Figure 3: Main experiment results with different number of examples per label.

Majority Vote This simple method always selects the most commonly occurring intent and slots for a given domain. It uses the DAP module to generate F_A , and assigns the domain specific labels according to the number of occurrences of $F \in F^*$ (such that $|F| = |F_A|$) in the training data.

Fully Supervised Uses the same architecture as the DAP (semantic parser) module, but it is trained using the full training data. Despite not being at all a fair comparison with the *simple label* settings, we include these baseline results as a way to visualize the best-case scenario.

W.O. Head This method does not use a classification head H , instead, it predicts the class label by selecting the example with highest semantic similarity with the input text using the cosine similarity score.

W.O. Head + Type This method not only uses the classification head, but also disregards the type constraint in Equation 3 such that the best frame is always $\arg \max_{F \in F^*} (\text{sim}(F, F_A))$.

5.3 Implementation Details

The span-pointer architecture used by the DAP module is a encoder-decoder model based on RoBERTa (Liu et al., 2019), with the encoder containing 12 layers and the decoder containing 1 layer. We train the model for 85 epochs using a learning rate of $1.67 \cdot 10^{-5}$. The sentence encoder used in the DSM module is built on top of a 36 layer XLM-R model (Conneau et al., 2019) fine-tuned

Eval. Setting	messa.	alarm	music	event	navig.	remind.	timer	weath.	avg.
Standard	86.7	67.6	55.5	73.9	60.6	58.5	72.7	79.6	69.4
+ Golden Parse	93.9	50.1	60.7	88.7	60.7	78.7	78.7	89.1	75.1
+ Recall@3	92.9	78.8	65.0	82.2	75.0	68.9	78.9	90.2	79.0
+ Intent Acc.	96.9	82.3	76.6	97.9	78.6	61.5	80.4	85.7	82.5

Table 1: Break down of results by domain for different evaluation settings. The results shown correspond to a single run of our proposed model with 50 labels per example.

to capture general sentence similarity. When fine-tuning the models we used a machine containing two NVIDIA Tesla P100 graphics processing units. Note that the underlying models used are relatively small when compared to current large language models (Sanh et al., 2019) and would be suitable for “on-the-edge” device computation.

5.4 Results

We perform three test runs with different random seeds for each evaluation setting, which influences both the model initialization and the set of training examples per label. The main results are shown in Figure 3. The baselines using the full training data (i.e., *Fully Supervised* and *Majority Vote*) are shown as dashed lines, and their results do not change according to the x-axis. The remaining results show the mean (and standard deviation as the shaded region) among the three test runs.

The results show that our proposed model outperforms most of the baselines, with numbers comparable to the *fully supervised* baseline (77.9% of its frame accuracy with 50 examples per label), even though it relies on a much smaller and simplified version of the training data.

There are a few other takeaways. First, we notice that increasing the number of “simple label” training examples significantly improves the frame accuracy. However, only five training examples is enough to produce decent results. More examples also seem to increase the variance of the models without a classification head. Second, the type filtering from Equation 3 is one key aspect of why the system can perform so well with so few example. Because of the generic nature of the domain agnostic ontology, filtering out invalid frames greatly reduces the size of the target space F^* , with a size reduction of 96% for certain domains such as `messaging`.

5.4.1 Results Break Down

To obtain further insights on the model, we show the results broken down by domain in Table 1. We used our proposed model trained on 50 examples per label, and different evaluation settings described as follows. The *Standard* settings are the same as the ones displayed in Figure 3. The *Golden Parse* assumes that the DAP module outputs only correct domain agnostic parses (i.e., the best-case scenario for the parser). The *Recall@3* results uses the same model as the *Standard* setting, but checks if the correct answer is in the top-3 scored matches (instead of top-1) from DSM. Finally, the *Intent Acc.* setting evaluates if the model correctly predicts the intent of the given input utterance. These results help us answer the following questions:

How much error from the parser gets propagated? We can notice from the *Golden Parse* results that there is a reasonable improvement (8.2% increase) in accuracy when using the gold test frames. This means that incorrect parses from the DAP module certainly propagates forward and improving the DAP module could certainly benefit the system as a whole.

Is the matching module nearly missing the right answer? When looking *Recall@3* results, we notice a significantly larger improvement in results (13.8% increase). With an average mean reciprocal ranking among all domains of 74.9. Having a high frame score values in the top-3 is significant considering that on average the size among all domains for the target space F^* is around 279.8 frames.

5.5 Error Analysis and Future Work

To understand the mistakes made by the OpenFSP system we perform some error analysis and suggest some possible improvement avenues for future work. For this analysis we use the development set and randomly sampled 100 output frames from different domains. We manually categorize the errors

as follows.

Parsing Errors We notice that 45% of the errors are due to parsing errors. This includes cases when the DAP module predicts an incorrect number of slots (~93% among parsing errors) or when the number of slots are correct, but some of the slots have the incorrect labels (~7% among parsing errors). A future direction could be to use the DAP and DSM modules to over-generate valid frames and rank (Varges, 2006; Zhang and Wan, 2022), which could circumvent parsing errors.

Intent Classification Error Another common error was the mislabeling of the utterance’s intent, corresponding to 32% of the manually categorized examples. This kind of error would often happen between semantically similar intents (e.g., IN:PREVIOUS_TRACK_MUSIC and IN:REPLAY_MUSIC in the music domain) and with examples from TopV2 that were labeled as unsupported (e.g., IN:UNSUPPORTED_WEATHER and IN:UNSUPPORTED_MUSIC) that often have out of scope questions that are harder to classify (e.g., “What is the hottest temperature this month”). One possible future direction would be to use Contrastive Learning (Chen et al., 2020a; Basu et al., 2022) that could improve the classification boundary of similar examples.

Slot Classification Error The last 23% of the errors were due to slot type misclassification. Again, semantically similar slot types are more challenging to classify. For instance, in the alarm domain SL:DATE_TIME_RECURRING, SL:DATE_TIME, SL:PERIOD and SL:DURATION were particularly hard to classify since they were all part of the same domain agnostic type SL:SCOPE_TEMPORAL. Another common issue was identifying proper names (~21% of the slot classification errors), including artist, event, album, and playlist names. A future direction would be to integrate a named entity recognition module to help classify slots involving proper names.

6 Conclusion

In this work we propose OpenFSP, a framework designed to simplify the process of adapting an existing task-oriented dialogue system to new domains. This framework enables non-experts to automatically build new domain ontologies from well

defined software engineering concepts such as functions and arguments. We define a general-purpose domain agnostic ontology, that when combined with textual examples of new slots and intents (which we call *simple labels*), provides sufficient data to adapt the system to a new domain.

Finally, we propose a two-module system that can use these simple labels to reasonably parse input utterances into the domain specific frames. Our experiments show that the proposed model outperforms strong baselines and is able to obtain results comparable with a fully supervised model (achieving 77.9% of its semantic frame accuracy). We hope that our work will facilitate the development of new assistant capabilities, allowing end-users to interact with more software applications through natural language.

Limitations

Domain-agnostic (DA) slots represent the overall semantic space covered by underlying TopV2 slots. Given their coarse-grain nature, DA slots are likely to be distributed more or less evenly across all domains. This assumption is key when training the parser on data that does not contain a particular target domain. In addition, we find that our approach is also sensitive to the types of linguistic structures accounted for by each DA slot and works best when these structures are consistent across domains.

More specifically, we conducted a series of leave-one-out (LOO) experiments where a separate parser was learned for each domain using training data from all other domains and then tested on test data from the domain in question exclusively. Error analysis of 100 randomly selected predictions in our LOO model for the “alarm” domain revealed that 32% of the errors were utterances such as “I want alarms set for next Monday at 6.00am and 7.00am” where the model predicted [SL:SCOPE_TEMPORAL for next monday at 6.00am] and [SL:SCOPE_TEMPORAL 7.00am]] whereas a compound [SL:SCOPE_TEMPORAL for next monday at 6.00am and 7.00am]] slot was expected. Upon closer inspection, we observed that these [SL:SCOPE_TEMPORAL X and Y] compound nominal constructions appear predominantly (83.3%) in the alarm domain, hence causing inaccurate predictions in a LOO model for this domain.

For these types of errors to be mitigated in our approach, [SL:SCOPE_TEMPORAL X and

Y] constructions would need to either be more evenly distributed across all domains, or re-annotated as [[SL:SCOPE_TEMPORAL X] and [SL:SCOPE_TEMPORAL Y]] in the alarm domain to improve homogeneity in the data.

Ethics Statement

No private data or non-public information was used in this work.

References

- Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. [Intent classification and slot filling for privacy policies](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4402–4417, Online. Association for Computational Linguistics.
- Ankur Bapna, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. In *Interspeech*.
- Samyadeep Basu, Amr Sharaf, Karine Ip Kiun Chong, Alex Fischer, Vishal Rohra, Michael Amoake, Hazem El-Hammamy, Ehi Nosakhare, Vijay Ramani, and Benjamin Han. 2022. [Strategies to improve few-shot learning for intent classification and slot-filling](#). In *Proceedings of the Workshop on Structured and Unstructured Knowledge Integration (SUKI)*, pages 17–25, Seattle, USA. Association for Computational Linguistics.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.
- Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020b. [Low-resource domain adaptation for compositional task-oriented semantic parsing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. [Supervised learning of universal sentence representations from natural language inference data](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.
- Shrey Desai, Akshat Shrivastava, Alexander Zotov, and Ahmed Aly. 2021. Low-resource task-oriented semantic parsing via intrinsic modeling. *arXiv preprint arXiv:2104.07224*.
- Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. [Simple, fast, accurate intent classification and slot labeling for goal-oriented dialogue systems](#). In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 46–55, Stockholm, Sweden. Association for Computational Linguistics.
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Jason Krone, Yi Zhang, and Mona Diab. 2020. [Learning to classify intents and slot labels given a handful of examples](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 96–108, Online. Association for Computational Linguistics.
- Aliasgar Kutianawala, Prateek Verma, and Zheng Yan. 2018. Towards a simplified ontology for better e-commerce search. *ArXiv*, abs/1807.02039.
- Yassine Laadidi and Mohamed Bahaj. 2018. [Simplification of owl ontology sources for data warehousing](#). In *Proceedings of the 2018 International Conference on Software Engineering and Information Management, ICSIM2018*, page 77–81, New York, NY, USA. Association for Computing Machinery.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

- Georgios Meditskos, Efstratios Kontopoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2020. Converser: Ontology-driven conversational awareness and context understanding in multimodal dialogue systems. *Expert Systems*, 37(1):e12378.
- Danilo Neves Ribeiro, Shen Wang, Xiaofei Ma, Rui Dong, Xiaokai Wei, Henghui Zhu, Xinchu Chen, Peng Xu, Zhiheng Huang, Andrew Arnold, and Dan Roth. 2022. [Entailment tree explanations via iterative retrieval-generation reasoner](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 465–475, Seattle, United States. Association for Computational Linguistics.
- Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.
- Guangyuan Piao. 2021. Scholarly text classification with sentence bert and entity embeddings. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25*, pages 79–87. Springer.
- Hoifung Poon and Pedro Domingos. 2010. Unsupervised ontology induction from text. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pages 296–305.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.
- Akshat Shrivastava, Pierce Chuang, Arun Babu, Shrey Desai, Abhinav Arora, Alexander Zotov, and Ahmed Aly. 2021. [Span pointer networks for non-autoregressive task-oriented semantic parsing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1873–1886, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Akshat Shrivastava, Shrey Desai, Anchit Gupta, Ali Elkahky, Aleksandr Livshits, Alexander Zotov, and Ahmed Aly. 2022. Retrieve-and-fill for scenario-based task-oriented semantic parsing. *arXiv preprint arXiv:2202.00901*.
- Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.
- Sebastian Varges. 2006. Overgeneration and ranking for spoken dialogue systems. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 20–22.
- Michael Wessel, Girish Acharya, James Carpenter, and Min Yin. 2019. Ontovpa—an ontology-based dialogue management system for virtual personal assistants. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, pages 219–233. Springer.
- Jason R Wilson, Kezhen Chen, Maxwell Crouse, Constantine Nakos, Danilo Neves Ribeiro, Irina Rabkina, and Kenneth D Forbus. 2019. Analogical question answering in a multimodal information kiosk. In *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*.
- Pengcheng Yin, John Wieting, Avirup Sil, and Graham Neubig. 2022. [On the ingredients of an effective zero-shot semantic parser](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1455–1474, Dublin, Ireland. Association for Computational Linguistics.
- Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. [Few-shot intent classification and slot filling with retrieved examples](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 734–749, Online. Association for Computational Linguistics.
- Yunxiang Zhang and Xiaojun Wan. 2022. [MOVER: Mask, over-generate and rank for hyperbole generation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6018–6030, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Domain Agnostic Ontology

To create our domain-agnostic ontology, we manually reviewed all existing slots in the `TopV2` ontology and categorized the semantic nature of each slot. We then conflated semantically-similar `TopV2` slots under overarching, domain-agnostic terms that cover the overall semantic space of the underlying `TopV2` slots. For instance, we categorize slots that indicate the user is requesting a "to-do item", "reminder," and "alarm" as roughly the overall "deliverable" item that is being requested, hence conflating the domain-specific slots `SL:TODO`, `SL:METHOD_TIMER`, and `SL:ALARM_NAME` under the domain-agnostic slot `SL:DELIVERABLE`. Table 2 contains the mapping between the `TopV2` ontology and the domain agnostic ontology used in this work.

Domain-agnostic (DA) slots	Domain-specific (DS) slots
SL:DELIVERABLE	SL:TYPE_REACTION, SL:TODO, SL:TODO_NEW SL:METHOD_TIMER, SL:TIMER_NAME, SL:ALARM_NAME
SL:RECIPIENT	SL:RECIPIENT, SL:PERSON_REMINDED_ADDED, SL:PERSON_REMINDED_REMOVED, SL:PERSON_REMINDED, SL:ATTENDEE_REMOVED, SL:ATTENDEE_ADDED
SL:SCOPE_TEMPORAL	SL:DATE_TIME, SL:DATE_TIME_RECURRING, SL:DURATION, SL:PERIOD, SL:RECURRING_DATE_TIME, SL:TIME_ZONE, SL:DATE_TIME_DEPARTURE, SL:DATE_TIME_ARRIVAL, SL:FREQUENCY, SL:RECURRING_DATE_TIME_NEW, SL:DATE_TIME_NEW, SL:SCOPE_TEMPORAL_RECURRING
SL:SCOPE_LOC	SL:LOCATION, SL:POINT_ON_MAP, SL:LOCATION_HOME, SL:LOCATION_USER, SL:LOCATION_MODIFIER, SL:WAYPOINT_ADDED, SL:LOCATION_WORK
SL:SCOPE_DISAM	SL:ORDINAL, SL:TYPE_CONTENT, SL:GROUP, SL:RESOURCE, SL:CONTENT_EMOJI, SL:TYPE_CONTACT, SL:MUTUAL_EMPLOYER, SL:MUTUAL_SCHOOL, SL:TYPE_INFO, SL:MUTUAL_LOCATION, SL:CONTACT_RELATED, SL:MUSIC_GENRE, SL:UNIT_DISTANCE, SL:WEATHER_TEMPERATURE_UNIT, SL:MEASUREMENT_UNIT, SL:METHOD_RETRIEVAL_REMINDER
SL:OTHER_OPEN_TEXT	SL:CATEGORY_EVENT, SL:SEARCH_RADIUS, SL:ATTRIBUTE_EVENT, SL:CATEGORY_LOCATION, SL:NAME_EVENT, SL:ATTENDEE, SL:ATTENDEE_EVENT, SL:TYPE_RELATION, SL:ORGANIZER_EVENT, SL:TAG_MESSAGE, SL:CONTENT_EXACT, SL:MUSIC_TYPE, SL:MUSIC_TRACK_TITLE, SL:MUSIC_ALBUM_TITLE, SL:MUSIC_PLAYLIST_TITLE, SL:MUSIC_RADIO_ID, SL:METHOD_TRAVEL, SL:JOB, SL:WEATHER_ATTRIBUTE, SL:OBSTRUCTION_AVOID, SL:ROAD_CONDITION_AVOID, SL:ROAD_CONDITION
SL:NUMS	SL:AMOUNT, SL:AGE
SL:PROPER_NAME	SL:NAME_EVENT, SL:CONTACT, SL:ORGANIZER_EVENT, SL:SENDER, SL:MUSIC_TRACK_TITLE, SL:MUSIC_PROVIDER_NAME, SL:MUSIC_ALBUM_TITLE, SL:MUSIC_ARTIST_NAME, SL:SOURCE, SL:DESTINATION, SL:PATH, SL:PATH_AVOID, SL:WAYPOINT_AVOID, SL:LOCATION_CURRENT, SL:PATH_AVOID, SL:WAYPOINT_AVOID, SL:LOCATION_CURRENT, SL:WAYPOINT, SL:ATTENDEE, SL:NAME_APP

Table 2: Mapping between TopV2 ontology and our proposed domain agnostic ontology.