Contextual Semantic Parsing for Multilingual Task-Oriented Dialogues

Mehrad Moradshahi[1] **Victoria Tsai**[2] **Giovanni Campagna**[1] **Monica S. Lam**[1]

[1] Stanford University, Stanford, CA

[2] Meta, Menlo Park, CA

{mehrad,gcampagn,lam}@cs.stanford.edu

{vitsai}@meta.com

Abstract

Robust state tracking for task-oriented dialogue systems currently remains restricted to a few popular languages. This paper shows that given a large-scale dialogue data set in one language, we can automatically produce an effective semantic parser for other languages using machine translation. We propose automatic translation of dialogue datasets with alignment to ensure faithful translation of slot values and eliminate costly human supervision used in previous benchmarks. We also propose a new contextual semantic parsing model, which encodes the formal slots and values, and only the last agent and user utterances. We show that the succinct representation reduces the compounding effect of translation errors, without harming the accuracy in practice.

We evaluate our approach on several dialogue state tracking benchmarks. On RiSAWOZ, CrossWOZ, CrossWOZ-EN, and MultiWOZ-ZH datasets we improve the state of the art by 11%, 17%, 20%, and 0.3% in joint goal accuracy. We present a comprehensive error analysis for all three datasets showing erroneous annotations can lead to misguided judgments on the quality of the model.

Finally, we present RiSAWOZ English and German datasets, created using our translation methodology. On these datasets, accuracy is within 11% of the original showing that high-accuracy multilingual dialogue datasets are possible without relying on expensive human annotations. We release our datasets and software open source.[1]

1 Introduction

Tremendous effort has gone into the research and development of task-oriented dialogue agents for English and a few other major languages in recent years. A methodology that can transfer the effort to other languages automatically will greatly benefit the large population of speakers of the many other languages in the world.

Underlying an effective TOD agent is *dialogue state tracking*, the task of predicting a formal representation of the conversation sufficient for the dialogue agent to reply, in the form of slots and values. However, DST currently remains restricted to a few popular languages (Razumovskaia et al., 2021). Traditional DST agents require large hand-annotated *Wizard-of-Oz* (Kelley, 1984) datasets for training, which are prohibitively labor-intensive to produce in most languages (Gunasekara et al., 2020). Large, multi-domain WOZ datasets are only available in English and Chinese (Quan et al., 2020; Ye et al., 2021a).

The contributions of this paper are as follows:

1. We propose an automatic technique to build multilingual data sets using machine translation. Machine translation has been shown effective for localizing question-answering agents (Moradshahi et al., 2020). It shows that for open ontology datasets, we need to use an alignment model to properly translate entities in the source language to entities in the target language. This paper shows that alignment is necessary even for closed ontology datasets and dialogues.

Furthermore, we improve alignment to address these challenging issues we discovered unique to dialogues: (1) Translation errors accumulate and can prevent a correct parse for the rest of the dialogue; (2) There are logical dependencies between slot values across different turns; (3) Utterances are generally longer and more complex carrying multiple entities. We found that alignment improves the accuracy on the RiSAWOZ benchmark by 45.6%. This technique eliminates the cost of human post-editing used on all previous translation benchmarks, and can improve machine translation quality on other tasks too.

Using this methodology, we automatically trans-

---

[1] https://github.com/stanford-oval/dialogues

late the RiSAWOZ dataset to English and German, creating RiSAWOZ-EN-auto and RiSAWOZ-DE-auto datasets respectively.

2. We show that the accumulation of translation and annotation errors across turns can be mitigated with a *Contextual Semantic Parsing* (CSP) model for state tracking. We propose a BART-CSP model, a seq-to-seq based on BART, that encodes the belief state, and the last agent and user utterances, rather than the full history of utterances.

BART-CSP improves SOTA on Ri-SAWOZ (Quan et al., 2020) and CrossWOZ (Zhu et al., 2020), two large-scale multi-domain WoZ dialogue datasets, by 10.7% and 17% in Joint Goal Accuracy(JGA). Notably, BART-CSP is more effective on translated data as evident by bigger performance improvement: on RiSAWOZ-EN-auto and RiSAWOZ-DE-auto datasets, automatically translated versions of RiSAWOZ, BART-CSP improves SOTA by 32.4% and 52.5%.

2   Related Work

2.1   Dialogue State Tracking

Dialogue state tracking (DST) refers to the task of predicting a formal state of a dialogue at its current turn, as a set of slot-value pairs at every turn. State-of-the-art approaches apply large transformer networks (Peng et al., 2020; Hosseini-Asl et al., 2020) to encode the full dialogue history in order to predict slot values. Other approaches include question-answering models (Gao et al., 2019), ontology matching in the finite case (Lee et al., 2019), or pointer-generator networks (Wu et al., 2019). Both zero-shot cross-lingual DST transfer (Ponti et al., 2018; Chen et al., 2018) and multilingual knowledge distillation (Hinton et al., 2015; Tan et al., 2019) have been investigated; however, training with translated data is the dominant approach, outperforming zero-shot and few-shot methods.

2.2   Contextual Semantic Parsing

Alternatively to encoding the full dialogue history, previous work has proposed including the state as context (Lei et al., 2018; Heck et al., 2020; Ye et al., 2021b) together with the last agent and user utterance. Recently, Cheng et al. (2020) proposed replacing the agent utterance with a formal representation as well. Existing models rely on custom encoder architectures and loss functions for the state (Heck et al., 2020). Our formulation of CSP

is different since we encode the formal dialogue state directly as text, which simplifies the architecture and makes better use of the pretrained model's understanding of natural text.

Previous work also applied rule-based state trackers that compute the state based on the agent and user dialogue acts (Schulz et al., 2017; Zhong et al., 2018; Zhu et al., 2020). Such techniques cannot handle state changes outside of a state machine defined ahead of time and do not achieve state-of-the-art accuracy on WOZ dialogues.

2.3   Multilingual Dialogues

Several multilingual dialogue benchmarks have been created over the past few years. Dialogue State Tracking Challenge (DSTC) has released several datasets (Kim et al., 2016; Hori et al., 2019; Gunasekara et al., 2020), covering only a few domains and languages. CrossWOZ (Zhu et al., 2020) and RiSAWOZ (Quan et al., 2020) are Chinese datasets collected through crowdsourcing. BiToD (Lin et al., 2021) uses a dialogue simulator to generate dialogues in English and Chinese, then uses crowdsourcing to paraphrase entire dialogues. All these approaches use crowdworkers in one or multiple stages of data collection which is costly and human errors degrade quality. Automatic creation of affordable high-quality dialogue datasets for other languages still remains a challenge (Razumovskaia et al., 2021).

3   Task Setting

We are interested in the *dialogue state tracking* task, in which the goal is to predict a formal representation of a conversation up to a certain point, also known as *belief state*, consisting of the slots that were mentioned in the dialogue and their value. At the beginning of the conversation, the belief state is empty, and it grows as the conversation progresses, accumulating the slots that were mentioned across all turns prior.

Formally, the problem is formulated given a predefined set of *slots* $s_1, s_2, \ldots s_n$ (such as "restaurant name", "restaurant food", etc.). Each slot has one value taken from the *ontology* $v_1, v_2, \ldots, v_n$. The ontology contains the legitimate values for the slot from the database (i.e. the list of restaurant names or restaurant cuisines), as well as the special values "none" indicating the slot was not mentioned, and "dontcare" indicating the slot was explicitly mentioned by the user but the user has
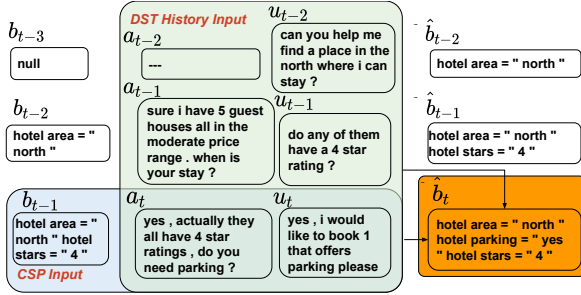
Figure 1: Visualization of the differences between inputs of CSP and traditional DST models. While the latter encodes the full history of the dialogue, CSP only encodes the current state and turn. b, a, and u indicate the dialogue state, agent utterance, and user utterance, respectively. t is the turn number. $\hat{}$ indicates a predicted dialogue state.

no preference.

Therefore, given a partial conversation history composed of turns $x_1, x_2, \ldots, x_t$, where each turn consists of an agent response (a_t) and user utterance (u_t), the task is to predict the belief state: $b_n(x_1, \ldots, x_t) = \{s_1 = v_{1,t}, s_1 = v_{2,t}, \ldots, s_n = v_{n,t}\}$ where n is the number of slots and $v_{i,t}$ is the value of the slot s_i up to turn t of the conversation. Note that the slot could be mentioned at turn t, or in any of the turns before.

4 Reducing Translation Noise

Key to the success of the dialogue state tracking is a precise and consistent annotation of the belief state at every turn. This is challenging in the multilingual setting, where we apply automatic translation to produce new training sets for DST: (1) the translation can be ungrammatical or incorrect, and it can introduce spurious information, and (2) the new annotation (in the target, translated ontology) can diverge from the translation of the sentence, such as referring to the same value in different ways in the target language.

We refer to the error introduced in translation collectively as *translation noise*. We posit that translation noise is the reason for poorer performance in existing translated dialogue datasets, compared to the same model on the dialogue dataset in the source language. In this section, we describe our methods to reduce translation noise.

4.1 Alignment

A major source of translation noise is due to mismatches between the translation of an entity by itself and in sentence. For instance, given a Chinese sentence that refers to an utterance containing the word "aquarium" (水族馆), incorrect translation may result in "ocean museum" in the English utterance, which does not match the slot value "aquarium" in the ontology and annotation anymore. Slot values may also get dropped or transliterated.

Translation with alignment was previously proposed by Moradshahi et al. (2020) to localize open-ontology multilingual semantic parsing datasets. Token alignments, obtained from cross-attention weights of the neural model, are used to track position of entities during translation so they can be replaced afterwards with local entity values. Figure 2 shows the translation and alignment process for an example input.

We show that alignment is useful also for a finite (closed) ontology in a dialogue setting. The dialogue setting is more challenging since the replacement with local entity values must be consistent across turns and dependent slots - slots that their values are logically dependent on each other. For instance, the corresponding price range for a *fast food* restaurant should be *cheap*, or a speaker looking for an attraction to go to with his girlfriend wants a place where *best-for-crowd* = "lover's date". Furthermore, utterances are generally longer and more complex containing multiple entities.

We have made several changes in alignment to address these issues:

1. We use a dictionary constructed from the dataset's ontology for translating the dependent slots to ensure relations are preserved. For all other slots, we randomly replace them with values from the target language ontology similar to previous work.

2. In previous work, quotation marks were used to mark the boundary of entities and to retrieve alignment between tokens in the input and output. We found the translation of quotation marks to be inconsistent. Instead, we omit those marks before translation and purely rely on cross-attention between subwords to compute alignment.

3. We observed alignment does poorly on digits and often misplaces them in the output. We use string matching to retrieve spans for numbers, dates, and time slots if present in the output and omit alignment if successful.

4. Dialogues contain longer utterances with multiple entities per turn. We found breaking
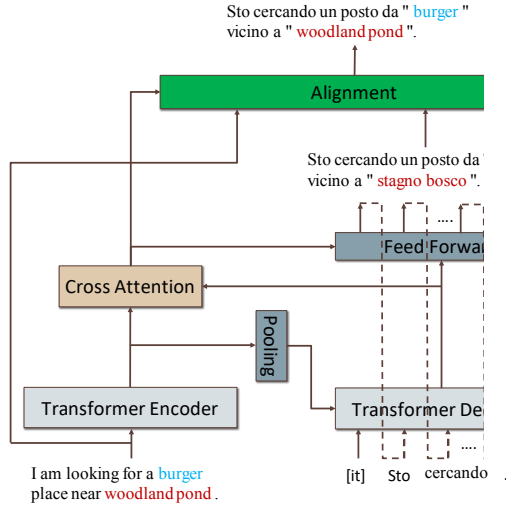
Figure 2: Example of the translation and alignment process. The encoder-decoder cross-attention weights are used to achieve word alignment.

down utterances into individual sentences before translation, significantly improves the quality of outputs when there are fewer entities to align.

To measure the effect of our changes, we translate MultiWOZ English dev set to Chinese with the new and previous alignment methods using MBART-MMT (see Section 5.3 for details). The new approach improves BLEU score by 8 points and JGA by 5% on MultiWOZ-ZH dev set.

4.2 Contextual Semantic Parsing

Previous work on dialogue state tracking encodes all or several turns up to the current one using a neural model, which then predicts the value for each slot. Hence, the input of the model consists exclusively of natural language, and grows as the conversation grows, accumulating any translation noise.

At the same time, we observe that the belief state at turn t, b_t, can be computed from the belief state at turn $t-1$ and the slots mentioned in the utterances at turn t. Hence, we propose to use a *contextual semantic parser* (CSP) for dialogue state tracking that computes $P(b_t|a_t; u_t; b_{t-1})$. The CSP model is applied to the dialogue-state tracking task by iteratively predicting the belief state of all turns, starting from b_0, the initial state consisting of all empty slots.

The CSP formulation condenses the dialogue history into a formal, fixed-length representation. Because the representation does not grow with the
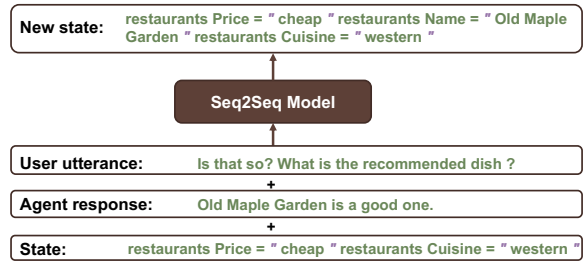

Figure 3: Contextual Semantic Parser model. We use BART for English and MBART for Chinese and German as the seq2seq model.

dialogue, it does not suffer from accumulation of translation noise.

Our CSP model is based on Seq2Seq Transformer models BART (Lewis et al., 2019) for English and MBART (Liu et al., 2020) for all others. Here we refer to them as CSP-BART for simplicity.

The model encodes the belief state as a textual sequence of slot names and slot values. This encoding is concatenated to the agent utterance and user utterance, and fed to the model to predict the belief state at the end of the turn (Fig. 3). Similar to (Yang et al., 2021b), we encode the belief state directly as text, which simplifies the architecture and leverages the pretraining of BART.

5 Experiments

Our experiments are designed to answer the following questions: 1) How well does CSP perform on WOZ datasets compared to DST models that encode the full conversation history? 2) Is our approach effective to reduce the translation noise?

5.1 Datasets

We evaluate our models on the RiSAWOZ (Quan et al., 2020), MultiWOZ (Budzianowski et al., 2018; Eric et al., 2019), and CrossWOZ (Zhu et al., 2020) datasets and their available translated versions: MultiWOZ Chinese (Li et al., 2021), CrossWOZ English (Li et al., 2021). These particular datasets were chosen because they are large Wizard-of-OZ dialogue datasets and therefore more natural and representative of task-oriented dialogues. Additionally, we use our methodology to create the RiSAWOZ English and German datasets.

RiSAWOZ (Quan et al., 2020) is a Chinese WOZ dataset of 11k annotated dialogues and over 150k utterances spanning the 12 domains of attraction, restaurant, hotel, flight, train, weather, movie, TV, computer, car, hospital, and courses. Dialogues are

formed from both single-domain and multi-domain goals, and annotated with dialogue states, dialogue acts, and coreference clusters.

MultiWOZ is an English-language WOZ dataset of 10k single- and multi-domain dialogues spanning the following 7 domains: taxi, restaurant, hotel, attraction, train, police, hospital. Following prior work with this dataset (Lee et al., 2019; Kim et al., 2019), we drop hospital and police from the training set as they are not included in the validation and test set. After the release of MultiWOZ 2.0 (Budzianowski et al., 2018), later iterations (Eric et al., 2019; Zang et al., 2020; Han et al., 2020) corrected some of the misannotations.

CrossWOZ is a Chinese-language WOZ dataset of 6k dialogues and over 102k utterances spanning the same 5 domains as the MultiWOZ validation and test sets: hotel, restaurant, attraction, metro, and taxi. CrossWOZ dialogues are annotated with dialogue states and dialogue acts, and average over 3.24 domains per dialogue, as opposed to the 1.80 of MultiWOZ.

For DSTC-9, Google NMT was used to translate MultiWOZ 2.1 to Chinese and CrossWOZ to English (Gunasekara et al., 2020; Li et al., 2021). To ensure translations of slot values in the dialog are faithful to the ontology, they replace them with their translations from a dictionary before feeding it to the NMT. This approach creates mixed-language sentences, shifting input distribution away from what public NMTs have been trained on, thus reducing quality (Moradshahi et al., 2020). Note that human translators were employed to proofread the translations and check certain slots to ensure values are correctly translated.

Table 1 shows a comparison of statistics for the training split of datasets used or created in this work. All the datasets have a closed ontology: the same entities appear in the train, validation, and test sets.

5.2 Models

We compare BART-CSP results against SOTA for each dataset. Models include the following:

- **TRADE** (Wu et al., 2019) uses a sequence-to-sequence architecture that encodes all utterances in the dialogue. It uses a pointer-generator to output the value of each slot.
- **MLCSG** (Quan and Xiong, 2020) extends TRADE by improving modeling of long contexts through a multi-task learning framework.

- **SOM** (Kim et al., 2019) considers dialogue state as an explicit fixed-sized memory and uses state operation to selectively update slot values at each turn.
- **SUMBT** (Lee et al., 2019) applies a BERT encoder to each utterance on the dialogue and a recurrent network to compute a representation of the whole dialogue, which is then matched against the ontology of each slot.
- **MinTL** (Lin et al., 2020) uses more recent pretrained models such as T5 (Raffel et al., 2019) and BART as the dialogue utterance encoder and builds an end-to-end dialogue model jointly learning dialogue state tracking, policy, and natural language generation tasks.
- **STAR** (Ye et al., 2021b) STAR uses two BERT models for encoding context and slot values. Additionally, they use a slot self-attention mechanism that can learn the slot correlations automatically. They use as input both the previous belief state and history of dialogue.

5.3 Implementation Details

BART-CSP is implemented using the Huggingface (Wolf et al., 2019) and GenieNLP (Campagna et al., 2019) libraries. We use the available open-source code for the other models. Hyperparameters for BART-CSP are discussed below; hyperparameters for other models are taken from the respective papers.

For semantic parsing we used *bart-base* (~139M parameters) for English and *mbart-large-50* (~611M parameters) model for other languages. For translation we used *mbart-large-50-many-to-many-mmt* (~611M parameters) which can translate directly between any pairs of 50 languages it supports. All models use a standard Seq2Seq architecture with a bidirectial encoder and left-to-right autoregressive decoder. All the models are pretrained using text denoising objective. *mbart-large-50-many-to-many-mmt* is additionally finetuned to do translation. BART uses BPE (Gage, 1994) to tokenize the input sentences whereas MBART uses sentence-piece model (Kudo and Richardson, 2018). We used Adam (Kingma and Ba, 2014) as our optimizer with a learning rate of 1×10^{-5} and used transformer warmup schedule (Popel and Bojar, 2018) with a warmup of 80. These hyperparameters were chosen based on a very limited hyperparameter search on the validation set. For the numbers reported in the paper, due to cost, we performed only a single run for each experiment.

	Dataset		
	RiSAWOZ	**CrossWOZ**	**MultiWOZ**
Languages	ZH, EN*, DE*	ZH, EN	EN, ZH
# Domains	12	5	7
# Dialogues	10,000	5,012	8,438
# Turns	134,580	84,692	115,424
# Slots	159	72	25
# Values	4,061	7,871	4,510

Table 1: Statistics of the training datasets (excluding validation and test). *Created using our methodology.

We batch sentences based on their input and approximate output token count for better GPU utilization. We set the total number of tokens to 700 for Mbart and 2000 for Bart models. We use gradient accumulation of 3 for Mbart and 9 for bart models boosting the effective batch size for better training. Our models were trained on NVIDIA V100 GPU using the AWS platform. For a fair comparison, all models were trained for the same number of iterations of 50K.

5.4 Metrics

We evaluate the models using the following two metrics:

- **Joint Goal Accuracy (JGA):** The standard metric of evaluation in DST is *joint goal accuracy*, which measures the average accuracy of predicting all slot assignments to exact match (EM) for any given turn. To compute this metric for CSP, the belief state predicted in previous turn is used as input for the current turn.
- **Gold Joint Goal Accuracy (GJGA):** This metric is similar to JGA but is calculated on a turn by turn basis, with ground-truth belief state used as input. Assuming the belief state correctly captures the state up to current turn of the dialogue, this metric acts as an oracle in evaluation removing the compounding effect of errors from previous turns.

6 Analysis and Results

Table 2 shows the results on test set for BART-CSP and previous SOTA models. For a better comparison we have included some details for each model:

- *Context Encoder*: The neural model used to encode the input.
- *Dialogue History*: If dialogue history is included in the input. A turn is defined as a pair of user utterance and agent response. "Partial" history means only a few turns of dialogue are kept while "Full" history models encode all previous turns.

- *Encodes State*: indicates if the user belief state up to the current turn is included in the input.
- *Predefined Slots or Ontology*: whether the model design or the data processing step needs knowledge of slot names or values.

As shown in Table 2, all previous models encode either a partial or full history. BART-CSP encodes significantly less information as it relies only on the current turn and the latest belief state. This simplifies model design and improves data efficiency for training (Yang et al., 2021a; Kapelonis et al., 2022).

Furthermore, models that rely on predefined ontologies require changes in architecture for new datasets. On the other hand, BART-CSP is a generative model that learns to copy slots from context and can be deployed for a new dataset as is.

On MultiWOZ, we report results for two models: MinTL which uses BART-large, and STAR[2] which uses BART as the context encoder. Our model outperforms MinTL despite using a smaller BART model, and achieves similar performance to STAR despite not having access to the full history. This shows bigger models do not necessarily yield better performance and model architecture and data representation are important too.

6.1 RiSAWOZ

The RiSAWOZ experiments shows that contextual semantic parsing delivers better accuracy than the state of the art on the original data sets; it is even more significant for translated data sets because it is more robust to translation errors.

BART-CSP improves the state of the art by 10.7% on JGA to 76.9% for the original Chinese data set. It provides an even greater improvement on the translated data sets: by 32.4% to 68.6% and by 42.5% to 65.9% on the automatically translated English and German data sets, respectively.

BART-CSP holds two major advantages over models that predict the slot-value pairs from the dialogue history. First, by distilling the belief state into a concise representation, it reduces noise in the input that would otherwise be present in a long dialogue history. Second, by taking the belief state as input, the model becomes more robust to translation errors in utterances from previous turns than

---

[2]SOLOIST (Peng et al., 2021) and SCORE (Yu et al., 2021) achieve better performance than STAR; however, these models use additional dialogue datasets to either pre-train or fine-tune their models. To keep the comparison fair, we only consider prior work that trained on MultiWOZ only.

Dataset	Model	Context Encoder	Dialogue History	Encodes State	Predefined Slots / Ont.	SOTA		BART-CSP	
						GJGA	JGA	GJGA	JGA
RiSAWOZ	MLCSG	Bi-GRU	Full	✗	✓	–	66.2	90.4	**76.9**
RiSAWOZ-EN-auto	MLCSG	Bi-GRU	Full	✗	✓	–	36.2	88.8	**68.6**
(-alignment)	MLCSG	Bi-GRU	Full	✗	✓	–	15.6	65.2	**22.9**
RiSAWOZ-DE-auto	MLCSG	Bi-GRU	Full	✗	✓	–	13.4	86.7	**65.9**
CrossWOZ	TRADE	Bi-GRU	Full	✗	✓	71.3†	36.1	**80.2**	**53.6**
CrossWOZ-EN	SOM	BERT	Partial	✓	✗	–	32.3‡	81.1	**52.3**
MultiWOZ 2.1	MinTL	BART-large	Partial	✓	✗	–	53.6‡	81.2	**53.7**
MultiWOZ 2.1	STAR	BERT	Full	✓	✓	78.7*	**56.7**	**81.2**	53.7
MultiWOZ-ZH 2.1	SUMBT	BERT	Full	✓	✓	–	46.0‡	75.9	**46.3**
MultiWOZ 2.4	STAR	BERT	Full	✓	✓	90.2*	**74.8**	**91.7**	70.4

Table 2: Comparison of results on test set for the BART-CSP model vs state-of-the-art on various datasets. The best result on JGA metric is in bold. GJGA is only bolded if SOTA is available. †RuleDST uses the previous system state and user dialogue acts as input. *STAR-GT uses ground truth previous dialogue state as input. ‡ indicates results are reported from their paper. We reproduced all results for models that source code and pretrained models were available to ensure correct comparison. "-align." denotes an ablation where translation is done without slot-value alignment.

models that accept dialogue histories. Hence it is even more effective on translated datasets.

Alignment is critical to generating a high-quality translated dialogue data set. The English and German semantic parser show only a degradation of 8-11% from the Chinese parser. In an ablation study of direct translation without alignment, the JGA on RiSAWOZ English drops from 68.6% to 22.9%, a difference of 45.7%.

Alignment ensures that entities are translated to the right phrase from the target ontology. For example, the phrase "姑苏区" is translated to "Aguzhou district" in a user utterance when the whole sentence is translated directly, but becomes "Gusu district" in the annotation. With alignment, both are translated identically to "Gusu district." The correspondence of utterance and belief state leads to higher DST performance.

6.2 CrossWOZ

Compared to MultiWOZ and RiSAWOZ, CrossWOZ is a more challenging dataset. Besides having longer conversations with more domains per dialogue, the cross-domain dependency is stronger. For example, the choice of location in one domain will affect the choice of location in a related domain later in the conversation, requiring models to have better contextual understanding. CrossWOZ is not only a smaller, more complex dataset than RiSAWOZ, but also exhibits a higher misannotation rate, to be discussed in Section 7. The current state of the art result was obtained with TRADE, which achieves 36.1% JGA.

The experiments with CrossWOZ also confirm that BART-CSP outperforms prior state-of-the-art models that encode full or partial history. Specifically, it exhibits an improvement of 17.5% in JGA on the original dataset and 20.0% in JGA on the English translated data set.

The GJGA metric for CrossWOZ was obtained by using RuleDST (Zhu et al., 2020), a set of handwritten rules specialized to the dataset to compute the new belief state from the ground truth user and system dialogue acts. BART-CSP outperforms the use of RuleDST in GJGA by 9%, showing that it is not necessary to handcraft these rules.

The translated CrossWOZ-EN data have been manually corrected for slot-value errors. Application of our automatic slot-value alignment technique would have greatly reduced the tedious manual effort required. In both GJGA and JGA, BART-CSP performs within 1% of the original Chinese dataset on English CrossWOZ.

6.3 MultiWOZ

MultiWOZ is a challenging data set because of the well-documented problem of misannotations in the data set (Eric et al., 2019). Misannotations teach the model to mispredict; conversely, correct predictions may be deemed to be incorrect. Thus the current state-of-the-art STAR model can only achieve an accuracy of 56.7%.

While BART-CSP accepts only the belief state as input context, the STAR model accepts both the previous belief state and the dialogue history. The latter offers an opportunity for the model to recover

missing state values from the history, giving a 3% advantage in JGA over BART-CSP. However, we note that once an agent misinterprets a user input, it is not meaningful to measure the accuracy for subsequent turns since the conversation would have diverged from the test data.

On the other hand, parsing history has its own cost: (1) It is less data efficient as you need more data to learn the same task. (2) It requires a more complex model that can find relevant slots among a large number of sentences. BART-CSP outperforms STAR by a 2.5% improvement in GJGA, suggesting that having the dialogue history as input can be detrimental when the past turns of a dialogue have been predicted correctly. On the Chinese translation of MultiWOZ, BART-CSP does slightly better than state of the art, improving JGA by 0.3% to 46.3%.

We also compare BART-CSP performance to MinTL, which is not SOTA, but uses more recent BART and T5 models as encoder. The results show that better performance on DST cannot be achieved by solely relying on better encoders with improved pretraining as both models are outperformed by STAR which uses BERT.

Between MultiWOZ 2.1 and 2.4, BART-CSP results improve by 16.7% on JGA and 10.5% on GJGA, while STAR improves by 18.1% on JGA and 11.5% on GJGA, showing dependence of both BART-CSP and STAR on the quality of annotation. Because MultiWOZ 2.4 only corrects the validations and test sets, CSP is still affected by mis-annotations in the training dataset. The lack of an equally clean training set may be the reason BART-CSP does not exhibit as much improvement across the versions.

7 Data and Error Analysis

A manual inspection revealed the following sources of errors on the CSP model, showing some of the inference limitations and its susceptibility to misannotations.

7.1 Misannotations

A substantial portion of incorrect predictions is due to existing annotation errors in all the datasets. In particular, in a manual review of 200 randomly chosen turns from each dataset, RiSAWOZ exhibits a 10.0% misannotation rate while CrossWOZ and MultiWOZ exhibit 17.9% and 26% misannotation rates, respectively.

Prevalent misannotation error types observed in the three datasets are noted below, with examples in the appendix.

- **Inconsistency:** Annotation inconsistency is a common issue with the Wizard-of-Oz data collection method. Examples of inconsistent annotations include inferred slots and slots that are mentioned by the agent but ignored by the user.
- **Inexact Match:** Typos, i.e. minor mismatches between the utterances and the annotation slot values. Chinese is a homonym-heavy language. It is not unexpected for single-character mismatch typos to occur frequently in the dataset. A second kind of typo is for a character to be entirely missing in an entity name.
- **Missing Slots:** Sometimes, values for some slots are simply just not included in the annotations.

CrossWOZ and MultiWOZ 2.1 are also susceptible to the following:

- **Extra Slots:** The presence of slot names which are not mentioned by either the user or the agent.

The following additional annotations problems are salient in MultiWOZ 2.1:

- **Delayed Annotations:** Slot values that are already confirmed by the user show up at a later turn in the conversation.

In RiSAWOZ, the final parting turn in a dialogue has no annotations, indicating a state reset:

- **Empty Annotation (Hard State Reset):** Some turns are missing annotations altogether.

7.2 Logical Relation Inference

In RiSAWOZ, the model is expected to infer the logical relationships between entities. For instance, the price range for a *fast food* restaurant should be *cheap*; looking for an attraction to go to with a girlfriend implies the interest of a place where *best-for-crowd* = "lover's date"; similarly the desired hotel rating is to be inferred from the utterance rather than explicitly mentioned.

For example:

- 我们一家是外地的，来苏州游玩，你可以帮我找一个在吴中区，中等消费水平的景点吗 (Our family is foreign, we have come to Suzhou to have fun, could you help me find, in the WuZhong area, a medium priced attraction?) The slot value pair 最适合人群= "家庭亲子" (*best-for-crowd* = "family") must be inferred.

- 好的。再看看高新区有什么好吃的火锅店？(Okay. Can you also see if GaoXin District has some good hotpot?)
  The slot 价位= "偏贵" (*price* = "expensive") must be inferred from the fact that the food is hotpot.

We found that most models struggle with this type of inference which requires higher level language understanding and reasoning. It is unclear whether this weakness is inherent to the model or whether it is an artifact of inconsistent annotations. Of the misannotations counted, 33% were missing the slot "Best for crowd."

7.3 Expanded Range Inference

In CrossWOZ, when a user requests, for instance, a restaurant where the cost per person is within a price range, oftentimes the agent cannot find such a restaurant, and responds with a suggestion with a cost outside of that range. In subsequent slots, the price range is expanded to include that cost, typically by rounding to the nearest ten. For example:

- 你好，我想去吃饭。请帮我找一家人均消费是100-150元，有井冈山豆皮这个菜的餐馆。(Hello, I'd like to go eat. Please help me find a place where the cost per person is 100-150 yuan, and has Jingangshan bean curd.) In this turn, the slot value for cost per person is *100-150*. However, following the agent response, the slot changes: 只为您找到一家叫西江美食舫(健德桥店)，但是它家的人均消费是83元。(I could only find a place called XiJiang Gourmet Boat (Jiandeqiao location), but the cost per person is 83 yuan.) Since 83 is outside the original range *100-150*, the slot value is expanded to *80-150* for the remaining turns of the dialogue.

We find that CSP struggles with such turns and will typically mispredict the slot value, assigning the previous range rather than inventing a new one.

8 Conclusion

Given a dialogue dataset in one language, this paper shows how to build contextual semantic parsers for a new language using automatically machine-translated data. We propose an improved alignment approach for dialogues to ensure faithful translation of slot values. This removes the need for costly human-post editing used in all previous benchmarks.

We show that the compounding effects of translation noise across turns can be mitigated with a CSP model for dialogue state tracking. By leveraging pretrained seq2seq models such as BART, training with CSP can outperform state-of-the-art results on RiSAWOZ, CrossWOZ, and MultiWOZ-ZH, and remains competitive on MultiWOZ, despite not encoding any previous conversation turns or having access to a predefined ontology.

We use our methodology to create RiSAWOZ English and German, the first automatically created high-quality translated datasets for dialogue state tracking with no human in the loop. We have implemented our methodology as a toolkit[3] which developers can use to create a new multilingual dialogue dataset as well as a contextual semantic parser for it.

9 Limitations

Organic multilingual dialogue datasets (i.e. created without the use of translation) are scarce, which has limited the scope of our experiments. We would have liked to evaluate the generalization of our approach to other languages. For instance, we partially rely on machine translation models to create datasets. Available translation models for other language pairs, especially from/to low-resource languages have much lower quality, and it would be desirable to measure the effect of that in our experiments.

Our methodology has only been applied to Human-to-Human dialogues annotated with slot-values. Although our approach is independent of data collection technique and formal representation, it should be applied and tested on datasets annotated with representations other than slot-values to study how well it can generalize.

Previous studies (Ding et al., 2021; Hung et al., 2022) utilized human post-editing to guarantee the fluency and accuracy of the translated datasets. However, in order to reduce cost, we have decided not to use manual post-editing in this work. As a result, our findings could be an overestimation of the model's actual performance in real-world scenarios. In future research, we plan to rectify this by manually post-editing the validation and test portions of the datasets.

---

[3]Code can be accessed at `https://github.com/stanford-oval/dialogues`

10 Ethical Considerations

Our translation method replaces the manual work needed to create multilingual dialogue datasets usually done via crowdsourcing. Instead, it requires some computation time which can be an environmental concern. However, in practice, such additional computing is small and much cheaper than the cost of human annotation for the same amount of data. The translation of the data set takes about half an hour on an Nvidia TITAN V GPU. Training takes about 6 hours on an Nvidia V100 GPU. We did not use crowdworkers for this paper. The error analysis was done by the authors.

References

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

Giovanni Campagna, Silei Xu, Mehrad Moradshahi, Richard Socher, and Monica S. Lam. 2019. Genie: A generator of natural language semantic parsers for virtual assistant commands. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pages 394–410, New York, NY, USA. ACM.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xlnbt: A cross-lingual neural belief tracking framework. *arXiv preprint arXiv:1808.06244*.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Bosheng Ding, Junjie Hu, Lidong Bing, Sharifah Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Globalwoz: Globalizing multiwoz to develop multilingual task-oriented dialogue systems.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. *arXiv preprint arXiv:1908.01946*.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Wei Peng, and Minlie Huang. 2020. Multiwoz 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv:2005.02877*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Chiori Hori, Julien Perez, Ryuichiro Higashinaka, Takaaki Hori, Y-Lan Boureau, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, Koichiro Yoshino, and Seokhwan Kim. 2019. Overview of the sixth dialog system technology challenge: Dstc6. *Computer Speech & Language*, 55:1–25.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Multi2woz: A robust multilingual dataset and conversational pretraining for task-oriented dialog. *arXiv preprint arXiv:2205.10400*.

Eleftherios Kapelonis, Efthymios Georgiou, and Alexandros Potamianos. 2022. A multi-task bert model for schema-guided dialogue state tracking. *arXiv preprint arXiv:2207.00828*.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. 2016. The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517. IEEE.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2019. Efficient dialogue state tracking by selectively overwriting memory. *arXiv preprint arXiv:1911.03906*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483, Florence, Italy. Association for Computational Linguistics.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Ryuichi Takanobu, Minlie Huang, and Jianfeng Gao. 2021. Multi-domain task-oriented dialog challenge ii at dstc9. In *AAAI-2021 Dialog System Technology Challenge 9 Workshop*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 pre-proceedings (NeurIPS Datasets and Benchmarks 2021)*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology QA semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. *arXiv preprint arXiv:1809.04163*.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Jun Quan and Deyi Xiong. 2020. Modeling long context for task-oriented dialogue state generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7119–7124, Online. Association for Computational Linguistics.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. Risawoz: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling. *arXiv preprint arXiv:2010.08738*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulić. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *arXiv preprint arXiv:2104.08570*.

Hannes Schulz, Jeremie Zumer, Layla El Asri, and Shikhar Sharma. 2017. A frame tracking model for memory-enhanced dialogue systems. *arXiv preprint arXiv:1706.01690*.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

Puhai Yang, Heyan Huang, and Xian-Ling Mao. 2021a. Comprehensive study: How the context information of different granularity affects dialogue state tracking? *arXiv preprint arXiv:2105.03571*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021b. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021a. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, Shenghui Li, and Emine Yilmaz. 2021b. Slot self-attentive dialogue state tracking. *arXiv preprint arXiv:2101.09374*.

Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, and Ahmed Hassan Awadallah. 2021. Score: Pretraining for context representation in conversational semantic parsing. In *International Conference on Learning Representations*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A Appendix

A.1 Missannotation Examples

See Table 3 below for missannotation examples discussed in Section 7.

Error Type	Dataset	Agent Utterance; User Utterance	Annotation	Correct Annotation
Delayed Annotation	MultiWOZ	great , i can get you a ticket for that train . how many people are riding with you ?; i need to book it for 6 people , can i get the reference number too ? [next turn] can i confirm you want to book this train for 6 people ?; yes , i would like to book the train for 6 people . i need the reference number , please .	train day = " thursday " train departure = " cambridge " train destination = " birmingham new street " train leaveat = " 10:00 " **train book people = " 6 "** train day = " thursday " train departure = " cambridge " train destination = " birmingham new street " train leaveat = " 10:00 "	**train book people = " 6 " train day = " thursday " train departure = " cambridge " train destination = " birmingham new street " train leaveat = " 10:00 "** (already correct)
Extra Slot	MultiWOZ	; hello ! i am planning my trip there and i am trying to find out about an attraction called kettle s yard . what can you tell me about it ?	attraction name = " kettles yard ", **attraction area = " west "**	attraction name = " kettles yard "
Empty Annotation	RiSAWOZ	;你好，刚到苏州，想先我个餐厅吃点东西。有价位便宜的江浙菜餐厅吗 Hello, I just came to Suzhou and am looking for a restaurant to grab a bite. Is there a cheap Jiangzhe restaurant?	**null**	餐厅价位= " 便宜" 餐厅菜系= " 江浙菜" Restaurant Price = "Cheap" Restaurant Cuisine = "Jiangzhe"
Inexact Match	RiSAWOZ	;你好，我这几天在苏州度假，明天准备去狐狸家手工奶酪这家餐厅吃饭，但不是很了解，你能帮我查查那个餐馆附近有没有地铁能直达呢？ ; Hello, I'm vacationing in Suzhou these several days, tomorrow I plan to go to the Fox Family Handmade Cheese restaurant to eat, but I don't really understand, can you help me look up whether there is direct subway access to anywhere near there?	餐厅名称= " 狐狸家手工酸奶酪" Restaurant Name = "Fox Family Handmade **Yogurt** (Yogurt = Sour Cheese)"	餐厅名称= " 狐狸家手工奶酪" Restaurant Name = "Fox Family Handmade Cheese"
Missing Slot	RiSAWOZ	有呀，推荐您去鑫花溪牛肉米粉。; 这家店地址在哪？ Yes, I recommend you go to Xinhuaxi Beef Noodle; What is the address of this place?	餐厅价位= " 中等" 餐厅菜系= " 快餐简餐" Restaurant Price = "Medium" Restuarant Cuisine = "Quick and Easy"	餐厅价位= " 中等" 餐厅名称= " 鑫花溪牛肉米粉" 餐厅菜系= " 快餐简餐" Restaurant Price = "Medium" **Restaurant Name = "Xinhuaxi Beef Noodle"** Restaurant Cuisine = "Quick and Easy"

Table 3: Prevalent annotation error types found in the datasets.