IndoToD: A Multi-Domain Indonesian Benchmark For End-to-End Task-Oriented Dialogue Systems

Muhammad Dehan Al Kautsar[1], Rahmah Khoirussyifa' Nurdini[1], Samuel Cahyawijaya[3],
Genta Indra Winata[2], Ayu Purwarianti[1]
[1]Institut Teknologi Bandung   [2]Bloomberg
[3]The Hong Kong University of Science and Technology
dehanalkautsar3@gmail.com

Abstract

Task-oriented dialogue (ToD) systems have been mostly created for high-resource languages, such as English and Chinese. However, there is a need to develop ToD systems for other regional or local languages to broaden their ability to comprehend the dialogue contexts in various languages. This paper introduces IndoToD, an end-to-end multi-domain ToD benchmark in Indonesian. We extend two English ToD datasets to Indonesian, comprising four different domains by delexicalization to efficiently reduce the size of annotations. To ensure a high-quality data collection, we hire native speakers to manually translate the dialogues. Along with the original English datasets, these new Indonesian datasets serve as an effective benchmark for evaluating Indonesian and English ToD systems as well as exploring the potential benefits of cross-lingual and bilingual transfer learning approaches.

1 Introduction

Task-oriented dialogue (ToD) systems are conversational agents designed to interact with users and assist them in various domains, such as restaurant (Bordes et al., 2016; Wen et al., 2017), public transport (Budzianowski et al., 2018; Lin et al., 2021b), and in-car assistance (Eric et al., 2017). This system also serves as the base for many commercial products that operate using the dialogue systems approach because of its ability to operate and understand the dialogue context without using hand-crafted rules (Lin et al., 2021b). Despite the recent growing interest in developing end-to-end ToD systems due to their simplicity, ToD systems are mostly created using monolingual datasets in high-resource languages such as English and Chinese. Moreover, building a ToD system can be challenging due to the limited availability of datasets for training and evaluating the system, which has been identified as the most critical factor in preventing the creation of bilingual/multilingual ToD

systems (Wen et al., 2017; Razumovskaia et al., 2021).

Developing ToD systems in additional underrepresented languages is essential to expand their capabilities to understand dialogue contexts in diverse languages (Kanakagiri and Radhakrishnan, 2021). One of them is Indonesian, a language spoken by many people worldwide yet still considered an underrepresented language in the end-to-end ToD system. Indonesian is ranked as the fourth most internet users in the world based on the latest data per country (Aji et al., 2022), with around 212 million internet users.[1] However, the language itself is still categorized as an underrepresented language in the natural language processing (NLP) community because of problems such as scattered datasets and minimum community engagement (Wilie et al., 2020; Cahyawijaya et al., 2021, 2022). To the best of our knowledge, there is only one publicly available end-to-end Indonesian ToD dataset[2] which is COD (Majewska et al., 2023), a multilingual ToD dataset that is solely used for evaluation – only test set available. It has a very limited number of samples for Indonesian with 194 dialogues across 11 domains (∼18 dialogues per domain) and no training data provided. This emphasizes the need to create larger end-to-end Indonesian ToD datasets to expand the capabilities to build and evaluate localized Indonesian ToD systems.

To address the aforementioned issues, we propose **IndoToD**, an end-to-end multi-domain ToD benchmark in Indonesian. IndoToD comprises the collection of two parallel Indonesian end-to-end ToD datasets covering four different domains by manually translating two existing English datasets: CamRest676 (Wen et al., 2017) and SMD (Eric et al., 2017) using delexicalization and lexicalization processes, as well as the evaluation of existing end-to-end ToD frameworks in various settings

---

[1]https://www.internetworldstats.com/stats3.htm
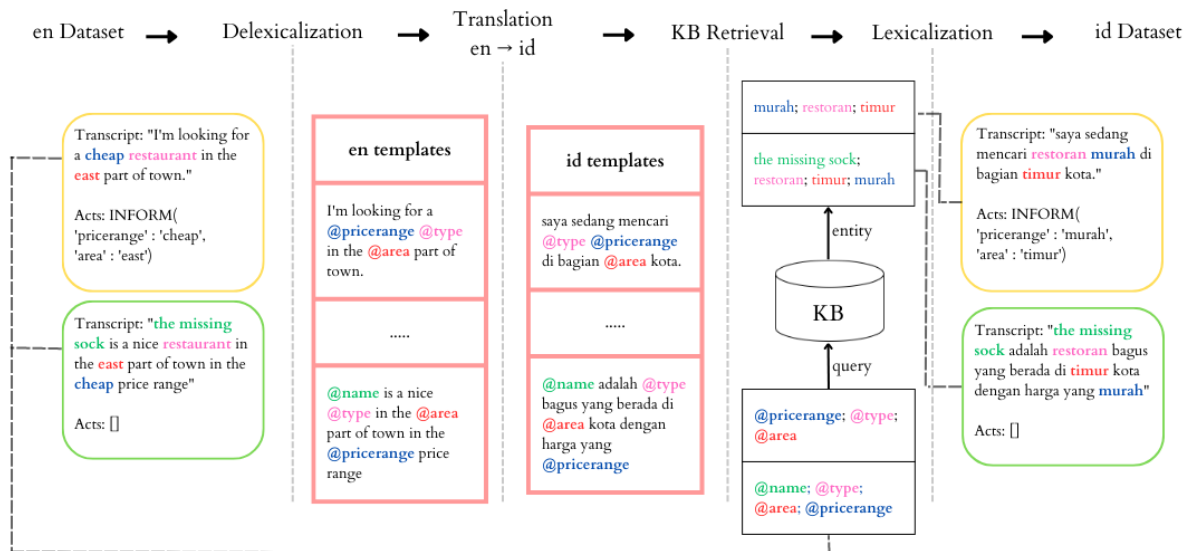[2]https://indonlp.github.io/nusa-catalogue/

Figure 1: Illustration of the dialogues collection pipeline: a) Delexicalize dialogue conversations to create English templates; b) Translate English conversation templates to Indonesian; c) Retrieve KB to retrieve corresponding entities; d) Lexicalize the dialogue conversations to collect Indonesian dialogues.

such as monolingual, cross-lingual, and bilingual. IndoToD provides more dialogue samples compared to (Majewska et al., 2023) that can be utilized for training and evaluation.

This paper's contributions are summarized three-fold as follows:

- We introduce IndoToD, the multi-domain benchmark for Indonesian ToD systems. The benchmark comprises two datasets: IndoCam-Rest and IndoSMD, with four different domains that serve as resources for training and evaluation.

- We establish baselines on monolingual, bilingual, and cross-lingual settings on existing end-to-end ToD frameworks.

- We analyze the effectiveness of training bilingual datasets to leverage more training data and handle tasks in both languages in building ToD systems.

2 IndoToD Benchmark

IndoToD benchmark[3] is created to develop the Indonesian ToD system. The benchmark covers four different domains, i.e., `restaurant search`, `point-of-interest (POI) navigation`, `calendar scheduling`, `and weather information`. IndoToD extends

---
[3] https://github.com/dehanalkautsar/IndoToD

the existing English datasets, and we follow a streamlined process to conduct dialogue collection. The dialogues are multi-turn conversations that involve two speakers (i.e., user U and system S). For each conversation, there is a knowledge base (KB) for the system to generate the correct entity for the user.

2.1 Dataset Collection

We use CamRest676 (Wen et al., 2017) and SMD (Eric et al., 2017) as the original English datasets that will go through several stages before becoming datasets that can be used in various Indonesian Task-oriented Dialogue (ToD) system experiments. CamRest676 is a ToD dataset that focuses on restaurant search queries collected via Wizard-of-Oz (WoZ) framework (Kelley, 1984). It consists of a collection of dialogues between the user and the system, where each dialogue has a task-specific goal (e.g., finding a restaurant). In this dataset, the user acts as a client who requests restaurant information and the system acts as an information provider that guides and answers the user's request.

Meanwhile, SMD is an in-car assistant multi-domain ToD dataset. It covers several domains such as POI navigation, calendar scheduling, and weather information. This dataset was also created by using the same WoZ framework to get a high-quality dataset that imitates a conversation between two individuals in a way that resembles a

86

Statistics	COD-id	IndoCamRest (ours)	IndoSMD (ours)
# domains	11: Flights, Weather, Alarm, Ride Sharing, Movies, Music, Media, Homes, Banks, Travel, Payment	1: Restaurant Search	3: POI navigation, Calendar Scheduling, Weather Information
# dialogues	194	676	323
# dialogues/domain	17.64	676	107.67
# distinct entities	N/A	110	325
# turns/dialogue	12.83	4.06	2.63
# tokens/utterance	N/A	9.54	8.14
# vocab (lexicalized)	N/A	1,103	883
# vocab (delexicalized)	N/A	829	577

Table 1: Comparison of our IndoToD benchmark with the Indonesian subset of COD (Majewska et al., 2023).

real-life interaction between a driver and an in-car assistant. In the WoZ framework, the source of conversation is created through a human-to-human dialogue conversation which is collected through crowd-sourcing. Because of that, the conversations between the user and the system are more natural.

These datasets then will be used as a reference for generating a ToD dataset that mimics natural conversation between two speakers and includes relevant knowledge base information for generating accurate responses. Both two datasets provide a diverse variety of dialogues since they have a different set of domains that complement the IndoToD benchmark.

2.2 Dataset Construction

The datasets have been created through several steps, as shown in Figure 1. First, existing English datasets are delexicalized into several user and system template sentences using Madotto et al. (2020) approach. Then, these template sentences are translated by native Indonesian annotators into Indonesian sentences. Lastly, new Indonesian datasets are built through the KB retrieval and lexicalization process.

Delexicalization Initially, the datasets are preprocessed by delexicalizing the entities. The process is carried out using the source code implemented by Madotto et al. (2020) to remove the entities in the dataset's dialogue. The aim of this step is to reduce the number of dialogues that require annotation (translation) by replacing all entities with a common, pre-defined value based on the type of entity. This pre-processing method is effective because the number of sentences that need to be

translated is reduced by 30% and 34% for CamRest676 and SMD, respectively. The output of this step is a list of pre-processed, entity-less sentences for each dataset.

Translation The list of pre-processed sentences is translated into Indonesian by annotators. In total, there are 3,834 and 1,126 sentences translated from CamRest676 and SMD, respectively. This process is followed by a cross-validation process where an annotator has to check the work of other annotators to maintain the dataset's quality. Note that for the SMD dataset, we only use 11% (323 dialogues) of its original dialogues due to the limitations on the annotator side.

KB Retrieval and Lexicalization After the translation process, to build each dialogue, a collection of Indonesian sentences and their corresponding entities are needed to begin the lexicalization process. We retrieve the Indonesian entities by querying the KB for the required slots to construct the dialogue. The lexicalization process is conducted by using Indonesian entities and sentences that have been retrieved earlier. This process ends with an evaluation of 100 randomly selected newly created Indonesian dialogues by human evaluators to detect any errors to ensure the quality of the resulting dialogues. If there are any errors, we ask them to manually edit that utterance. The sample of corrected Indonesian utterances is presented in Appendix Table 1.

2.3 Dataset Statistics

We collect a total of 999 dialogues with 6,338 utterances from two datasets, named as IndoCamRest and IndoSMD, which are derived from CamRest676 and SMD datasets, respectively. We com-

Statistics	IndoCamRest	IndoSMD		
	Restaurant Search	POI Navigation	Calendar Scheduling	Weather Information
# dialogues	676	100	114	109
# slot types	6 (area, food, price range, postcode, phone, address)	5 (POI name, traffic info, POI category, address, distance)	6 (event, time, date, party, room, agenda)	4 (location, weekly time, temperature, weather attr.)
# distinct slot values	88	131	79	78

Table 2: Per-domain statistics of IndoCamRest and IndoSMD.

pare several statistic attributes between our datasets (IndoCamRest and IndoSMD) and COD that are shown in Table 1. Furthermore, the per-domain statistics of IndoCamRest and IndoSMD are presented in Table 2. IndoCamRest and IndoSMD share some of the same statistical characteristics as the previous datasets, including the number of domains, intents, and slot types. It is worth noting that both of our datasets have a substantially larger number of dialogues per domain compared to COD. In addition, IndoCamRest and IndoSMD comprise both the lexicalized and delexicalized forms in the dialogue. The inclusion of the delexicalized form is intended to facilitate the generation of dialogues by lexicalizing the delexicalized sentences.

3 Experimental Setup

3.1 Experiment Settings

We set up a benchmark for both Indonesian and English ToD to evaluate the performance of the current ToD systems. We explore three different training and evaluation settings:

- **Monolingual – id/en**. We train the end-to-end ToD systems using monolingual corpus in each language independently.

- **Cross-lingual**. We train the systems using the English end-to-end ToD dataset before testing it using an Indonesian test set to analyze the effectiveness of the cross-lingual approach in understanding the context of the dialogue.

- **Bilingual – id+en**. We train the systems by combining English and Indonesian datasets, utilizing parallel corpora from CamRest and SMD. This allows us to examine how the system's performance is affected by exposure to bilingual languages.

For monolingual and bilingual settings, we test each ToD framework in English and Indonesian test sets separately to analyze the impact of monolingual and bilingual training in each language. We also split the data into 3:1:1 and 8:1:1 for the CamRest and SMD, respectively.

3.2 ToD Baselines

As baselines, We evaluate four ToD frameworks: Sequicity (Lei et al., 2018), LABES (Zhang et al., 2020a), MinTL (Lin et al., 2020), and GALAXY (He et al., 2022). Since the four ToD system frameworks are designed for English, we have made adjustments to the frameworks to suit the requirements of our experiments. We largely adopt the configuration settings used in the original paper, including batch size, decoding method, early stop count, and learning rate.

To adapt the Sequicity and LABES frameworks for our experiments, we utilize fastText (Bojanowski et al., 2017) instead of GloVe (Pennington et al., 2014). This was necessary because fastText can accommodate English, Indonesian, and bilingual Indonesian-English word vectors (Conneau et al., 2017). Specifically, we incorporate common crawl[4] English and Indonesian fastText word vectors for monolingual English and Indonesian experiment settings respectively. Meanwhile, we use bilingual English-Indonesian word vectors (Conneau et al., 2017) for other settings.

Unlike the other frameworks, MinTL utilizes T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) as its backbone, which sets it apart regarding approach and potential performance. Note that for this experiment, we only use T5 (Raffel et al., 2020) and mT5 (Xue et al., 2021) as the backbone models, and use this framework without utilizing

---
[4]https://fasttext.cc/docs/en/crawl-vectors.html

Baselines	IndoCamRest				IndoSMD			
	BLEU	Match	Success	Combined	BLEU	Match	Success	Combined
monolingual								
Sequicity (Lei et al., 2018)	16.74	92.31	81.29	103.54	**8.75**	**44.83**	**61.04**	**61.69**
LABES (Zhang et al., 2020a)	17.59	70.00	78.49	91.83	1.01	0.00	34.82	18.42
MinTL (Lin et al., 2020)	14.61	80.33	80.31	94.93	0.45	24.14	23.70	24.37
GALAXY (He et al., 2022)	**17.89**	**96.24**	**84.62**	**108.32**	-	-	-	-
cross-lingual								
Sequicity (Lei et al., 2018)	0.08	8.46	31.58	20.10	0.00	15.52	29.46	22.50
LABES (Zhang et al., 2020a)	0.02	0.00	60.96	30.50	0.00	**15.79**	55.84	**35.82**
MinTL (Lin et al., 2020)	**1.61**	**16.67**	**68.81**	**44.35**	0.00	13.33	**56.75**	35.04
GALAXY (He et al., 2022)	0.05	0.00	47.24	23.67	0.00	0.00	37.11	18.56
bilingual								
Sequicity (Lei et al., 2018)	15.84	89.23	**84.01**	102.46	10.95	**58.33**	75.66	**77.94**
LABES (Zhang et al., 2020a)	**18.77**	**96.36**	81.68	**107.79**	8.78	47.37	71.11	68.02
MinTL (Lin et al., 2020)	18.65	87.30	79.63	102.12	**11.59**	48.33	**77.86**	74.68
GALAXY (He et al., 2022)	17.09	94.74	83.60	106.26	0.08	45.00	9.17	27.17

Table 3: Experiment settings result on Indonesian test set. **bold** denotes the best score per metric.

Levenshtein Belief Spans (Lev) resulting in an inference process similar to Sequicity. In all of the experiment settings, mT5-small is implemented in the framework and used as the backbone of the framework. We also try to use T5 for monolingual and bilingual settings to analyze the impact of using two different kinds of pre-trained language models (LMs) that will be discussed later in the Subsection 4.2. We implement reader and evaluator modules for both CamRest and SMD datasets in MinTL, as their paper only included the reader and evaluator for the MultiWOZ (Budzianowski et al., 2018) dataset.

3.3 Hyper-parameters

We conduct the learning rate search [1e-4, 6e-4, 1e-5] to optimize the performance of MinTL on the new datasets, and the best learning rate is 6e-4. For other ToD frameworks, we adopt their proposed learning rate in our experiments. We use a learning rate of 3e-3 for Sequicity, 1e-4 for GALAXY, and, for LABES, we set the learning rate to 3e-4 and 1e-4 in the CamRest and SMD experiments, respectively. We follow the inference sampling strategies recommended by the original papers that introduce the frameworks. We use a beam size of 5 for generating inference responses in GALAXY, while in the other frameworks, we employ greedy search. We run each experiment once per experiment setting for uniformity with an A100 40GB GPU.

3.4 Evaluation Metrics

We use four evaluation metrics for end-to-end ToD tasks to evaluate the quality of responses. There are 1) **BLEU** score (Papineni et al., 2002) to assess how fluent the generated response is, 2) **Match** rate (Lin et al., 2020), checks whether the ToD system can produce the entity constraints specified by the user, thereby measuring task completion, 3) **Success** F1 score (Lei et al., 2018), which determines whether the system has successfully provided all of the information that the user has requested, such as the address or phone number, and 4) **Combined** score (Lin et al., 2020; Mehri et al., 2019) using the equation **Combined** = (Match + Success) × 0.5 + BLEU. We assess the performance of all frameworks using the CamRest and SMD datasets, along with the designated experiment settings.

4 Results & Analysis

4.1 Results

Tables 3 and 4 show the experiment results for Indonesian and English test sets, respectively. We conclude that GALAXY has a decent score in both monolingual settings for CamRest, and Sequicity with fastText word vector has the best model in monolingual settings for SMD. Furthermore, MinTL outperforms other frameworks in zero-shot cross-lingual settings for both datasets. The non-contextual word-embedding frameworks, such as Sequicity and LABES achieve the highest combined scores for bilingual settings. Furthermore, it was observed that GALAXY does not perform

Baselines	CamRest				SMD			
	BLEU	**Match**	**Success**	**Combined**	**BLEU**	**Match**	**Success**	**Combined**
monolingual								
Sequicity (Lei et al., 2018)	**21.17**	90.30	80.99	106.81	**21.04**	82.12	78.21	**101.20**
LABES (Zhang et al., 2020a)	19.98	93.64	72.93	103.26	1.94	80.81	60.07	72.39
MinTL (Lin et al., 2020)	20.34	83.21	83.64	103.77	18.53	**83.90**	**78.60**	99.78
GALAXY (He et al., 2022)	20.53	**96.24**	**84.62**	**110.96**	18.25	80.00	77.32	96.91
bilingual								
Sequicity (Lei et al., 2018)	22.53	91.04	82.06	109.08	21.04	75.56	**83.48**	100.56
LABES (Zhang et al., 2020a)	22.58	**97.27**	79.33	**110.88**	**22.35**	**83.83**	74.02	**101.28**
MinTL (Lin et al., 2020)	**22.88**	86.26	82.26	107.14	18.39	80.23	79.10	98.05
GALAXY (He et al., 2022)	20.31	95.49	**84.67**	110.39	16.58	77.37	76.84	93.68

Table 4: Experiment settings result on English test set. **bold** denotes the best score per metric.

as well as other frameworks when dealing with Indonesian, as we denote its results by using '-' in Table 3.

Additionally, we present the combined metric results of the ToD frameworks on the IndoSMD test set, categorized by domain, in Figure 2. We discover that POI navigation is relatively easier than other domains to learn, while calendar scheduling is the hardest domain for ToD frameworks on SMD dataset. Fine-grained results can be found in Appendix Table 2.

4.2 Analysis

Monolingual English vs. Indonesian Performance. In general, the monolingual model trained and tested in English performs better than in Indonesian across different frameworks and datasets. We hypothesize that the pre-trained LMs used in the experiments are better trained to handle English data. GALAXY outperforms the monolingual setting on both Indonesian and English CamRest datasets. However, we find that the GALAXY framework fails to adapt to the IndoSMD dataset and the model is not converged. We further observe that the experiments using GALAXY in Indonesian are not as effective as English, as evidenced by the superior performance on the English CamRest dataset, despite having the same amount of dialogue and utterance as IndoCamRest. Furthermore, it is worth noting on the IndoToD benchmark that we are not using Indonesian T5 models. This is because there is no paper working on releasing the Indonesian T5 models that are comparable to the English T5 models we use in the experiment.

Transferability from English to Indonesian. Based on the cross-lingual results, it can be concluded that existing ToD frameworks are unable to handle the task when trained in a different language. While not all training data needs to be in the same language as the test set, it is still essential for the framework to have some knowledge of the target language in order to understand the task at hand. Based on the inference results, we observe that all of the ToD frameworks generate responses in English despite Indonesian user inputs. This can be correlated to the fact that the training data used in the cross-lingual setting is in English. Among all the ToD frameworks that are evaluated in this setting, MinTL performs the best in the cross-lingual setting (due to mT5 as its pre-trained model, trained on multilingual datasets), while GALAXY and Sequicity do not perform well.

Monolingual vs. Bilingual. The results demonstrate that, in most cases, the bilingual setting yields higher scores than the monolingual (i.e., Indonesian and English monolingual setting), especially in the Indonesian test set. The bilingual setting has the advantage of using a larger amount of training data and performing tasks in both languages (Lin et al., 2021b), increasing both the overall metric scores and the individual metric components. We also compare and analyze the framework's performance closely based on the test set we used. The results show that most of the bilingual settings scores are higher than monolingual settings in both the Indonesian and English test sets showing the effective usage of language data counterpart for domain adaptation.

However, based on the inference result in Table 5, we observe that the responses generated between bilingual and monolingual settings are not substantially different in terms of the semantics, confirmed by the fact that the difference in
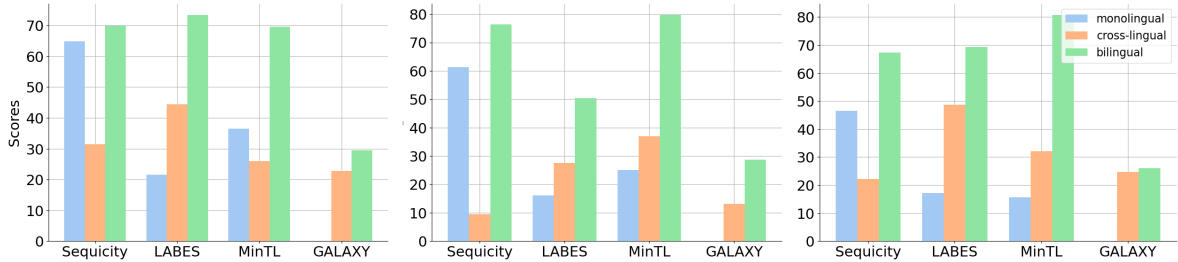
Figure 2: ToD framework results on IndoSMD per domain. **Left**: `POI navigation` domain, **Center**: `Calendar scheduling` domain, and **Right**: `Weather information` domain.
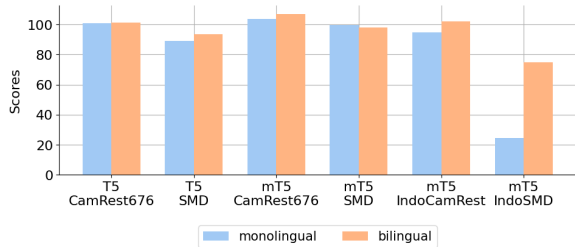

Figure 3: Monolingual and bilingual MinTL combined score using T5 and mT5 models on CamRest676, SMD, IndoCamRest, and IndoSMD as test sets.

BLEU scores between these two settings is not significant. However, we see some improvements in belief spans generation in the bilingual setting. ToD frameworks can understand the context from the user better when they are trained with a larger amount of training data, although the training data are a mixture of English and Indonesian. This highlights the significance of leveraging cross-lingual data, in this case, using English data on top of Indonesian data, in the training process of the ToD system to enhance the model performance and achieve better metric scores in the target language.

Furthermore, we delve deeper into analyzing the potential of bilingual training using the MinTL framework. Interestingly, this framework is flexible since we can modify the pre-trained transformer LM unlike other frameworks. We explore not only English T5, but also multilingual version of T5. We compare the performance of T5-small and mT5-small as the backbone of the MinTL framework, shown in Figure 3 and Appendix Table 3 for per metric details. Based on our findings, we observe that bilingual training positively impacts the BLEU score, match rate, and combined metrics for the MinTL framework, either using T5 or mT5. The results indicate that the bilingual setting can yield benefits on the efficient training process of T5 and mT5, especially in Indonesian, that is considered as

underrepresented language. Moreover, we notice that mT5-small is more effective than T5-small in end-to-end ToD tasks. Overall, the bilingual training is beneficial for training end-to-end ToD, especially to improve the performance on non-English languages.

Error Analysis. We conduct an error analysis to investigate the limitations of existing frameworks. In general, most issues occur in the form of a "template response". The end-to-end ToD systems sometimes generate responses that appear to be template-like despite varying user inputs which might be because of the limited trainable parameters and dataset used in the training process. Despite its goal to complete the user's task, the ToD must also be able to maintain the user's satisfaction whenever using the ToD as a product. Users may become disinterested if they receive repeated responses, potentially leading to reduced usage. Another issue we notice is the occurrence of repeated tokens in the generated output by the framework during the inference phase. Specifically, we find that the Sequicity and MinTL frameworks tend to produce responses with repeated tokens more than the other ToD frameworks. This could be due to the decoding method applied by these frameworks in our research which relied on a greedy search approach. The other issue that is worth mentioning in inference examples is the hallucination in the generated outputs. Some of the outputs do not have the correct combination of slots that are found in the training set.

5 Related Work

ToD Datasets. Multiple studies have contributed to creating datasets for ToD using various approaches. bAbI (Bordes et al., 2016) is an English synthetic dataset combining five subtasks into ToD system datasets. SMD (Eric et al., 2017)

Input	Belief spans	Response
"carikan saya restoran mahal yang menyediakan makanan internasional." (*find me an expensive restaurant that serves international food*)	**gold:** "restaurant-food": "internasional", "restaurant-pricerange": "mahal" **monolingual:** "restaurant-food": "internasional", "restaurant-pricerange": "menengah" **bilingual:** "restaurant-food": "internasional", "restaurant-pricerange": "mahal"	**gold:** "maafkan saya. saya tidak bisa menemukan restoran apapun yang menyediakan makanan internasional di kisaran harga yang mahal" (*I am sorry. I couldn't find any restaurant that provides international food at such an expensive price range*) **monolingual:** "value_name adalah restoran mahal yang menyediakan makanan internasional. apakah kamu ingin tahu lokasinya?" (*value_name is an expensive restaurant that serves international food. Would you like to know its location?*) **bilingual:** "value_name menyediakan makanan internasional. apakah kamu ingin informasi lebih lanjut?" (*value_name serves international foods. Would you like other further information?*)

Table 5: Example of the LABES response on IndoCamRest using monolingual and bilingual settings. The English translation is in *italic*.

uses the WoZ scheme to build an English car assistant dialogue system that has knowledge of weather and places. Other English datasets using the WoZ scheme are CamRest (Wen et al., 2017), a dataset for a restaurant reservation, MultiWoZ (Budzianowski et al., 2018), a multi-domain WoZ dataset containing five domains collected using the human-to-human scheme, and OpenDialKG (Moon et al., 2019), a synthetic dataset containing four domains. On non-English datasets. such as Chinese, there are CrossWoZ (Zhu et al., 2020), a large-scale Chinese WoZ dataset with five domains, and RiSAWoZ (Quan et al., 2020), a large-scale multidomain Chinese WoZ dataset with rich semantic annotations up to 12 domains. MetaLWOz (Lee et al., 2019) is a goal-oriented dialogue corpus containing 51 domains and collected using WoZ. Other datasets include bilingual languages, such as BiToD (Lin et al., 2021b), a dataset for tourism assistants focusing on English and Chinese languages, and COD (Majewska et al., 2023), a dataset containing Russian, Arabic, Indonesian, and Swahili. Based on the aforementioned datasets, we conclude that the ToD dataset with contextual knowledge focusing on Indonesian is not yet available.

Indonesian Dialogue Systems Datasets. Research for the ToD system in the Indonesian language is still underway. IndoNLG (Cahyawijaya et al., 2021), a benchmark for natural language generation in low-resource languages, focuses on summarization, question answering, chit-chat, and machine translation for Indonesian, Javanese, and Sundanese languages. An Indonesian language subset of the XPersona (Lin et al., 2021a) dataset, which consists of approximately 17 thousand dialogues, is used for the chit-chat task. It is worth noting that although IndoNLG is making significant progress, it has not yet explored the ToD system. Furthermore, NusaCrowd (Cahyawijaya et al., 2022), a large pool of hundreds of Indonesian NLP data, only listed one Indonesian ToD dataset, i.e., COD (Majewska et al., 2023), which is a multilingual ToD dataset that is solely used for evaluation with only 194 dialogues spanning across 11 domains. Therefore, the Indonesian ToD dataset is still very limited, and urgent action is required to increase the coverage of Indonesian ToD datasets.

Frameworks for End-to-End ToD. GALAXY (He et al., 2022) is a pre-trained dialogue model that uses semi-supervised learning to learn dialogue policy from limited labeled dialogues and large-scale unlabeled dialogue corpora. LABES (Zhang et al., 2020a), conversely, represents belief states as discrete latent variables and models them jointly with system responses given user inputs. MinTL (Lin et al., 2020) efficiently uses pretrained LMs in developing task-oriented dialogue systems, eliminating the need for ad hoc modules in studies that use this framework. DAMD (Zhang et al., 2020b) is a network that accommodates state-action pair structure in generation and can leverage the proposed multi-action data augmentation

framework to address the multidomain response generation problem. Sequicity (Lei et al., 2018) uses the Seq2Seq model for dialogue state tracking and generating a response to the user. Some other recent works (Bang et al., 2023; Hudeček and Dušek, 2023) explore the potential of using large LMs for performing zero-shot ToD. In this paper, we utilize GALAXY, LABES, MinTL, and Sequicity to evaluate the performance of various end-to-end ToD frameworks.

6 Conclusion

We introduce IndoToD, an end-to-end multidomain Indonesian ToD benchmark. We extend the existing two English ToD datasets using an efficient framework by delexicalizing the conversation into templates and reconstruction using KB. We evaluate our benchmark in monolingual, cross-lingual, and bilingual settings. We also show the benefits of having English data to improve Indonesian performance. Furthermore, we conduct some analysis from multiple viewpoints that shows the existing ToD system's inability to understand dialogue context from unseen language and the current ToD system's behavior in handling underrepresented language such as Indonesian using our datasets. We also investigate some errors that usually occur in ToD system responses to explore its limitation.

7 Limitations

There are several limitations that hinder our progress in creating the first Indonesian end-to-end ToD benchmark. Firstly, we encountered a limitation in translating the English dataset. Due to the constraints of our resources, specifically having only three annotators, we could only translate the CamRest dataset and a mere 10% of the SMD dataset. Expanding the dataset size would significantly contribute to advancing the Indonesian ToD system. Second, we only use publicly available end-to-end ToD system framework repositories as our baselines. Based on MultiWoZ benchmark[5], there are several end-to-end ToD frameworks with slightly superior metric scores compared to our baselines (Budzianowski et al., 2018). But because of their unavailability, we could not evaluate them and only focused on evaluating the best available ToD frameworks possible.

[5] https://github.com/budzianowski/multiwoz

8 Ethics Statement

Our work spotlights a need to develop ToD systems for underrepresented languages such as Indonesian through our new ToD datasets: IndoCamRest and IndoSMD. During our study, we commit to the ethical principles of NLP research and are well aware of its impact on the language community. We strongly believe that there is no ethical issue within this work, including the data collection process, annotation, existing dataset usage, and experiments. Within our work, annotators are well-rewarded and give us informed consent as they understand and agree to provide their annotations to build new ToD datasets. We uphold our annotator's privacy and follow the data protection and privacy regulation for releasing the datasets. The dataset itself is free from abusive language and personal information as our ultimate goal in this work is to contribute and make an impact on society with provide useful NLP task resources and more linguistic diversity in the NLP field.

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, pages 135–146.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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*, pages 5016–5026.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Pascale Fung, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2022. Nusacrowd: Open source initiative for indonesian nlp resources.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10749–10757.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are llms all you need for task-oriented dialogue?

Tushar Kanakagiri and Karthik Radhakrishnan. 2021. Task-oriented dialog systems for dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 85–93.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2:26–41.

Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. 2019. Multi-domain task-completion dialog challenge. In *Dialog System Technology Challenges 8*.

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, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021a. Xpersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021b. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787*.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 165–177.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Evgeniia Razumovskaia, Goran Glavavs, Olga Majewska, E. Ponti, Anna Korhonen, and Ivan Vulic. 2021. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *J. Artif. Intell. Res.*, 74:1351–1402.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020a. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020b. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

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   Annotator Recruitment

We recruit two Indonesian native speakers that can speak English to translate the delexicalized English utterance template to Indonesian. The age of the annotators ranged from 19 to 25, and they were college students by profession. They work around 18 hours in total, and they are compensated approximately 22,583 IDR per hour for their work, 18,62% bigger than the worker minimum wage in West Java (19,038 IDR per hour)[6].

B   Sample Evaluation Correction by Annotators

Appendix Table 1 shows several examples of evaluation by annotators. The evaluation process is intended to maintain high-quality translation from the annotator. We report five sampled English utterances, incorrect Indonesian utterances, and corrected Indonesian utterances by annotators.

C   Per Domain Experiment Settings Result on IndoSMD Test Set

We report per domain results for all of the baselines on IndoSMD dataset on Appendix Table 2. We can conclude that Sequicity has the best score on monolingual Indonesian setting, meanwhile MinTL outperforms other frameworks on cross-lingual and bilingual settings by having the most biggest metric score compared to others when handling Indonesian test set.

D   MinTL Result Comparisons

We present MinTL result comparisons by using two pre-trained LM: T5 and mT5 on Appendix Table 3.

---

[6]https://www.bps.go.id/indicator/19/1172/1/upah-rata---rata-per-jam-pekerja-menurut-provinsi.html

The table compares monolingual and bilingual approach on English test set. The result shows that bilingual setting have a beneficial impact on MinTL to learn the task well compares to monolingual setting.

E   Annotator Instruction and Consent

We report instructions for annotators to translate several English CamRest and SMD utterance templates, as well as annotator informed consent templates on Appendix Figure 1, Appendix Figure 2, and Appendix Figure 3. The instructions are in Indonesian as all of the annotators are Indonesian native speakers.

English Utterance	Incorrect Indonesian utterance	Corrected Indonesian utterance
"remind me that i have a doctors appointment at 5 pm"	"ingatkan saya bahwa saya akan janji temu dengan dokter pada pukul 5 sore"	"ingatkan saya bahwa saya **memiliki** janji temu dengan dokter pada pukul 5 sore"
"that you, that's all I need to know."	"itu kamu, hanya itu saja yang perlu saya ketahui"	"**terimakasih**, hanya itu saja yang perlu saya ketahui" (there is a typo on the English utterance. It suppose to be 'thank you', not 'that you')
"are there ant jamaican restaurants in any part of town?"	"apakah ada restoran semut jamaika di bagian manapun dari kota ini?"	"apakah ada restoran jamaika di bagian manapun dari kota ini?" (there is a typo on the English utterance. It suppose to be 'any', not 'ant')
"la tasca is located at 14-16 Bridge Street and the postcode is C.B 2, 1 U.F. Their phone number is 1223 464630. May I help you with anything else?"	"la tasca berlokasi di 14 -16 bridge street dan kode pos nya adalah c.b 2, 1 u.f. nomornya adalah 1223 464630. bisakah saya membantu kamu dengan hal yang lain?"	"la tasca berlokasi di 14 -16 bridge street dan kode pos nya adalah c.b 2, 1 u.f. nomornya adalah **01223 464630**. bisakah saya membantu kamu dengan hal yang lain?" (we check in to the KB and we find out the correct entity for its phone number is '01223 464630')
"it is located at Quayside Off Bridge street. The phone number is is 01223 301030."	"alamatnya adalah quayside off bridge street. nomor teleponnya adalah 01223 301030."	"alamatnya adalah quayside off bridge street. nomor teleponnya adalah **01223 307030**." (we check in to the KB and find out that the correct entity is 01223 307030)

Table 1: Sample evaluation correction by annotators.

Baselines	Domain											
	POI Navigation				Calendar Scheduling				Weather Information			
	BL	Ma	Su	Co	BL	Ma	Su	Co	BL	Ma	Su	Co
monolingual												
Sequicity	**7.90**	25.00	**89.05**	64.93	**6.85**	**61.90**	46.91	61.26	10.57	33.33	38.71	46.59
LABES	1.30	0.00	40.50	21.55	0.00	0.00	32.30	16.15	0.00	14.30	20.00	17.15
MinTL	0.02	**50.00**	22.92	36.48	0.34	19.05	30.30	25.02	0.004	4.76	26.51	15.64
GALAXY	-	-	-	-	-	-	-	-	-	-	-	-
cross-lingual												
Sequicity	0.00	12.50	50.53	31.52	0.00	14.29	4.82	9.56	0.00	19.05	25.45	22.25
LABES	0.00	13.30	**75.50**	**44.40**	0.00	**16.70**	38.30	27.50	0.00	**42.90**	54.50	**48.70**
MinTL	0.00	**16.67**	35.29	25.98	**0.004**	14.29	**59.54**	36.92	0.00	4.76	**59.41**	32.09
GALAXY	**0.01**	11.76	33.66	22.72	0.00	0.00	26.17	13.09	0.00	0.00	49.54	24.77
bilingual												
Sequicity	**8.92**	44.44	77.47	69.88	11.73	**61.90**	67.37	76.37	11.72	**57.14**	54.05	67.32
LABES	4.50	**60.00**	**77.90**	**73.45**	6.90	33.30	53.70	50.40	9.70	42.90	76.30	69.30
MinTL	4.75	55.56	74.24	69.65	**15.94**	47.62	**80.00**	**79.75**	**13.85**	52.38	**81.54**	**80.81**
GALAXY	0.06	47.06	11.76	29.47	0.06	50.00	7.23	28.68	0.12	42.86	8.96	26.03

Table 2: Per domain experiment settings result on Indonesian test set using SMD dataset. **bold** denotes the best score per metric. BL, Ma, Su, and Co denote the **BL**EU score, **Ma**tch rate, **Su**ccess F1, and **Co**mbined score respectively.

Experiment Setting	CamRest				SMD			
	BLEU	**Match**	**Success**	**Combined**	**BLEU**	**Match**	**Success**	**Combined**
T5-small								
Monolingual	17.57	83.21	**83.25**	100.80	15.31	71.54	76.22	89.19
Bilingual	**18.50**	**85.50**	80.35	**101.42**	**16.23**	**75.29**	**79.52**	**93.63**
mT5-small								
Monolingual	20.34	83.21	**83.64**	103.77	**18.53**	**83.90**	78.60	**99.78**
Bilingual	**22.88**	**86.26**	82.26	**107.14**	18.39	80.23	**79.10**	98.05

Table 3: MinTL (Lin et al., 2020) result comparisons using pre-trained LM: T5 and mT5 as backbone model on English test set. **bold** denotes the best score per metric.

Figure 1: Instruction for annotators to translating English CamRest templates.

TRANSLATING SMD DATASET: README
Task : mengubah kalimat bahasa inggris yang terdapat pada sheet `usr_utterance` dan `sys_utterance` menjadi kalimat bahasa indonesia
How?
- Translate di samping kalimat bahasa inggrisnya di tiap sheet `usr_utterance` dan `sys_utterance` :D
- Kalau bingung sama kalimatnya, bisa lihat keseluruhan conversationnya di sheet 'conversation' --> track pakai cid
- Gausah pakai kapital
- style : sopan tapi ngga terlalu kaku
Note:
- Terdapat beberapa kata / token yang tidak perlu di-translate, yaitu kata yang mengandung '@'
- Jika bertemu dengan kata yang mengandung '@', kata lain selain kata tersebut harus di-translate, sedangkan untuk kata '@' dapat dibayangkan menjadi beberapa kata di bawah ini (agar peletakan kata tersebut menjadi baik)
@event : jadwal berobat / makan malam
@time : jam 1 siang / 8 malam
@date : senin / tanggal 1
@party : tim eksekutif / HR / ibu / kakak
@room : ruangan konferensi 102
@agenda : membahas liburan perusahaan / penerimaan karyawan baru / membahas dresscode (aturan berbusana)
@location : oakland / san fransisco
@weekly_time : hari ini / besok / lusa
@temperature : 20 derajat / 30 derajat
@weather_attribute : berawan / badai / berembun
@traffic_info : tabrakan mobil / kemacetan / lancar
@poi_type : rumah teman / rumah / kafe / rumah sakit
@poi : Pizza hut / McDonalds / Taman bermain
@poi_address : jalan kacang panjang
@distance : 1 km / 1 mil

Figure 2: Instruction for annotators to translating English SMD templates.

Data Collection Informed Consent

Name   :
Age   :
E-mail   :

I agree to participate as an annotator in the research project conducted by:

Name   :
E-mail   :
Research title :

As an annotator, I understand that I will be asked to translate 748 English sentences into Indonesian and validate the translated samples from November 1st to November 9th. I will also be required to follow specific guidelines and instructions the research team provides. The fee that I will receive is Rp

I agree to allow the research team to use and publish the annotations I provide to create an open-source dataset. I also understand that the translation results will be published as part of the research, and my personal information may be included in the publication (i.e. ethnicity, age, profession).

I have read and understood the information provided in this consent form, and I have had the opportunity to ask any questions about the research project and my participation in it.

Annotator,

Figure 3: Annotator informed consent template.