Building Dialogue Understanding Models for Low-resource Language Indonesian from Scratch

DONGLIN DI, Advance.AI
XIANYANG SONG, Northeast Forestry University
WEINAN ZHANG*, Harbin Institute of Technology
YUE ZHANG, Westlake University
FANGLIN WANG, Advance.AI

Using off-the-shelf resources from resource-rich languages to transfer knowledge to low-resource languages has received a lot of attention. The requirements of enabling the model to achieve the reliable performance, including the scale of required annotated data and the effective framework, are not well guided. To address the first question, we empirically investigate the cost-effectiveness of several methods for training intent classification and slot-filling models from scratch in Indonesia (ID) using English data. Confronting the second challenge, we propose a Bi-Confidence-Frequency Cross-Lingual transfer framework (BiCF), which consists of "BiCF Mixing", "Latent Space Refinement" and "Joint Decoder", respectively, to overcome the lack of low-resource language dialogue data. BiCF Mixing based on the word-level alignment strategy generates code-mixed data by utilizing the importance-frequency and translating-confidence. Moreover, Latent Space Refinement trains a new dialogue understanding model using code-mixed data and word embedding models. Joint Decoder based on Bidirectional LSTM (BiLSTM) and Conditional Random Field (CRF) is used to obtain experimental results of intent classification and slot-filling. We also release a large-scale fine-labeled Indonesia dialogue dataset (ID-WOZ [1]) and ID-BERT for experiments. BiCF achieves 93.56% and 85.17% (F1 score) on intent classification and slot filling, respectively. Extensive experiments demonstrate that our framework performs reliably and cost-efficiently on different scales of manually annotated Indonesian data.

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Discourse, dialogue and pragmatics**.

Additional Key Words and Phrases: dialogue datasets, intent classification, slot-filling, Indonesian

1 INTRODUCTION

Dialogue is one of the vital ways for people to establish contact with each other, whether at work or in daily life. With the rapid development of the Internet, most people choose to chat online instead of having face-to-face communication. Therefore, dialogue texts in the network provide data resources with considerable quantity for studying neural dialogue understanding models, which rely heavily on the large scale of training data. [26] Whereas, there is a big gap in the number of people speaking different languages in the world. Most of the existing studies focus on rich-resource languages [49]. However, there are thousands of minority languages that have a limited range of usage and few available resources. It is impractical and cost-ineffective to collect and

[1]https://github.com/Davidddl/ID-WOZ

Authors' addresses: Donglin Di, donglin.ddl@gmail.com, Advance.AI; Xianyang Song, sxy56713@nefu.edu.cn, Northeast Forestry University; Weinan Zhang, wnzhang@ir.hit.edu.cn, Harbin Institute of Technology; Yue Zhang, yue.zhang@wias.org.cn, Westlake University; Fanglin Wang, fanglin.wang@advance.ai, Advance.AI.

2375-4699/2022/12-ART $15.00
https://doi.org/10.1145/3575803

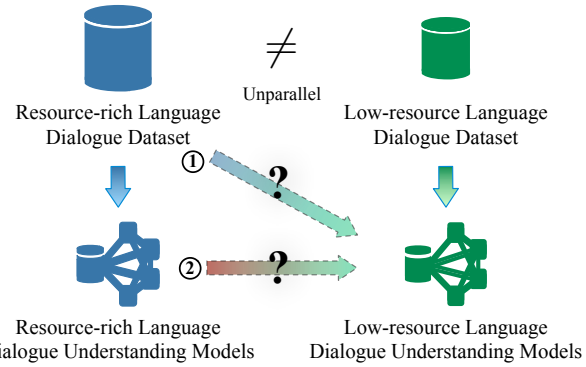

Fig. 1. It is time-consuming to construct large-scale high-quality datasets of low-resource language. Therefore, there are two mainstream methods for constructing low-resource language models: $Arrow_1$ indicates that a high-resource language dialogue dataset is applied to train the low-resource language dialogue understanding models through machine translation or cross-lingual pre-trained embeddings. $Arrow_2$ indicates that the contextual word embeddings are used to transfer existing resource-rich language dialogue understanding models to low-resource language dialogue understanding models.

annotate enough large-scale datasets for low-resource languages [16] to train the dialogue understanding models. Thus, as shown in Fig. 1, it remains huge challenges on how to efficiently adapt existing research resources and findings to low-resource languages (*e.g.*, Indonesian (ID)), so that the need for understanding the multilingual task-oriented dialogue [39, 40] can be addressed effectively.

The first challenge is how to effectively obtain low-resource languages by utilizing rich-resource languages (*e.g.*, English [3]). The intuitive method is to use a neural machine translation system [6, 45] to translate the English dataset into Indonesian, and then train the dialogue understanding models on the translated data. Another strategy is to use multilingual word embeddings [10, 34], which allow the dialogue model trained on the English dataset to be directly applied to Indonesian since the pre-trained multilingual model contains the vocabulary of both English and Indonesian. Each of the above methods has both strengths and limitations. While the former can save vast resources for collecting low-resource language data, it needs to deal with machine translation errors and invalid dislocated annotations from the source corpus, which can significantly influence the subsequent dialogue modeling (*i.e.*, slot-filling). The latter suffers from intrinsic differences between English and Indonesian, including variations in syntactic and semantic patterns.

The second challenge relies on how to transfer existing models to the target low-resource language. One possible approach is to align the contextual word embeddings in the semantic latent space for sentence-level encoding [39, 40] in order to avoid semantic misunderstanding and syntactic errors. However, this method is susceptible to imperfect alignments, and its implementation is complex, making it difficult to deploy or apply models.

To address these challenges, we propose a **Bi**-**C**onfidence-**F**requency Cross-lingual Transfer framework (BiCF). For the first challenge, we adopt the word-level alignment strategy [52], which has been demonstrated as effective as phrase-level alignment yet much simpler and more stable [43, 44]. Specifically, the first stage of BiCF is Bi-confidence-frequency Mixing, utilizing the English dataset to generate code-mixed data, which avoids sentence-level translation errors as well as label dislocation. The mixed data includes gold annotations for Indonesian from English datasets as well as importance-frequency and translating-confidence. And for the second challenge, in our framework, Latent Space Refinement and Joint Decoder are designed on the top of the resulting high-quality mixed data, utilizing and refining pre-trained off-the-shelf word embedding models, to

train the dialogue understanding models (*i.e.*, intent classification and slot-filling) for Indonesian. To conduct extensive experiments for Indonesian, we follow the method of MultiWOZ [3], which is a large-scale task-oriented English dialogue dataset, to collect and annotate a counterpart and richer-domain dataset in Indonesian (ID), named ID-WOZ. Extensive experiments demonstrate that our proposed framework can tackle the practical intent classification and slot filling well in Indonesian. The main contributions of our work are summarized as follows:

- We propose the Bi-confidence-frequency Cross-lingual Transfer framework to utilize English datasets for training Indonesian dialogue understanding models and achieve satisfactory performance on Indonesian dialogue datasets.
- We release a large-scale manually annotated multi-domain ID-WOZ dialogue dataset along with a pre-trained ID-BERT model as the resource contributions for low-resource language dialogue understanding tasks.
- We investigate the demand for annotated data of well-performing dialogue understanding models, which may guide future research on collecting datasets or training models for other low-resource languages.

2 RELATED WORK

Low-resource Language. Natural language processing in low-resource languages is difficult, which has piqued the interest of many researchers. Low-resource languages are those lack sufficient monolingual or parallel corpus as well as artificially crafted linguistic resources for constructing statistical NLP tasks. To improve the low-resource language understanding, Guo et al. [18] propose two algorithms to generate the cross-lingual distributed representations of words. This method can map two different languages into a joint vector space and use distributed feature representations to remedy the defect of the lexical feature gap. Duong et al. [12] build an accurate dependency parser by training a model with shared structure across fewer training languages. Ammar et al. [1] propose a multilingual model for parsing dependency in multiple languages that includes a multilingual word cluster, token-level language information, and language-specific features. Wang et al. [47] propose to integrate English syntactic knowledge into a parser trained on the Singlish treebank, and demonstrates that it is reasonable to leverage English to improve low-resource language models. The above works mainly focus on multilingual parsing of low-resource languages, which are useful in improving the low-resource language understanding.

Pre-trained Language Models. Pre-trained language model is the deep neural network trained on large-scale unlabeled data, which can be fine-tuned to suit different downstream tasks. Pre-trained language models have a strong ability to encode semantic information in different languages. Therefore, it is promising to apply pre-trained language models to enhance the understanding of cross-lingual dialogue systems. The success of Transformer [46] has led researchers to improve various pre-trained models based on it. GPT [35] is a semi-supervised method based on Transformer decoder using a combination of pre-training and fine-tuning. **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) [10] uses masked language models to enable pre-trained deep bidirectional representations from unlabeled texts by jointly combining both left and right contexts. Later, pre-trained language models, including XLNet [51], RoBERTa [28], T5 [36], and mBART [27] are proposed. XLNet and RoBERTa are based on BERT, while mBART and T5 are based on the encoder-decoder structure.

Cross-lingual Transfer. In general, standard NLP techniques are difficult to directly apply to low-resource languages. State-of-the-art models required a large number of training data which is unavailable for most languages. Cross-lingual transfer learning is a common paradigm for low-resource dialogue models and is usually divided into two categories: Transfer of annotations and Transfer of models. Based on most methods applied large parallel corpora to learn cross-lingual word embeddings, Artetxe et al. [2] propose a self-learning framework and a small size of word dictionary to learn a mapping between source and target word embeddings. Utilizing syntactic dependency features is one of the classical methods for treebank translation. Relying on aligned parallel

sentence pairs suffers from noise and imperfect alignments. Zhang et al. [52] focus on improving dependency parsing by translating confident words into a source treebank. Schuster et al. [40] utilize Multilingual CoVe embeddings obtained from Machine Translation systems [31] in Thai and Spanish for zero-shot dependency parsing. In line with these methods, encoding the semantic information directly within the same cross-lingual latent space could avoid semantic misunderstanding from machine translation or wrong alignments.

Cross-lingual Dialogue Systems. The lack of high-quality training data in low-resource language has hampered the development of task-oriented cross-lingual dialogue systems. Many researchers incorporated the concept of zero-shot learning into cross-lingual dialogue systems to cope with the scarcity of low-resource language dialogue data. Liu et al. [29] propose a zero-shot method that uses parallel word pairs to refine cross-lingual word representations. Then researchers use a latent variable model to deal with inherent differences across languages. Later, Liu et al. [30] further improve the zero-shot adaptation method and propose a novel model that uses the attention mechanism to extract source words. Xiang et al. [50] propose to develop an end-to-end cross-lingual dialogue system based on the idea of zero-shot under the guidance of machine translation and resource-rich language dialogue dataset. Although existing dialogue systems based on zero-shot learning have greatly improved in terms of generating fluent responses, there is still a gap when compared to human-to-human communication. Sun et al. argue that the dialogue system's responses are generally simple and blunt, preventing the conversation from transferring to a specific topic. Therefore, Sun et al. [41] propose a novel task, named cross-lingual knowledge grounded conversation, which employs knowledge distillation and a large-scale dialogue corpus to improve cross-lingual knowledge selection in the target language. Kim et al. [23] propose a Korean Wizard of Wikipedia dataset to extract knowledge across languages. Experiments confirm that using only English datasets can improve the performance of the non-English dialogue system.

Above all, existing studies tend to adopt new strategies or neural models (*e.g.*, zero-shot learning, knowledge-grounded method, and pre-trained word embeddings) to circumvent the difficulty of constructing low-resource language dialogue datasets. In our work, we applied the word-level alignment strategy, which is more convenient and stable than the traditional dependency parser, to generate code-mixed data with labels in Indonesian, importance-frequency, and translating-confidence. We use Multilingual-BERT, which was trained by traditional BERT on a large-scale corpora of 102 languages, as the word embedding model. Pre-trained Multilingual-BERT trained with our code-mixed data is capable of capturing latent semantic information in both English and Indonesian. We employ traditional BiLSTM and CRF as the decoder for intent classification and slot filling.

3 METHOD

In this section, We introduce the proposed pipeline framework "Bi-Confidence-Frequency Cross-Lingual transfer framework (BiCF)" in detail. It mainly consists of three components, namely "BiCF Mixing", "Latent Space Refinement", and "Joint Decoder". As shown in Fig. 2, the BiCF mixing step replaces a few English words with Indonesian. Then we train and refine the cross-lingual semantic embedding latent space based on the mixed data with gold annotations from English dataset. Finally, we adopt the combination of BiLSTM and CRF to decode the intent and slots jointly.

3.1 BiCF Mixing

The first stage of our framework is "**Bi**-**C**onfidence-**F**requency Mixing" (BiCF Mixing). As shown in Fig. 2, we use the English data in two steps. The first is to generate the frequency-word set ($\mathbf{W}_{freq}$) of English data. The second is to obtain the word alignment with the translating-confidence (λ_{conf}) of each word and generate confidence-word set ($\mathbf{W}_{conf}$). The goal of this stage is to select both frequent and high-confidence word pairs for English and Indonesian, and yield mixed data $\mathcal{T}_{mix}$.

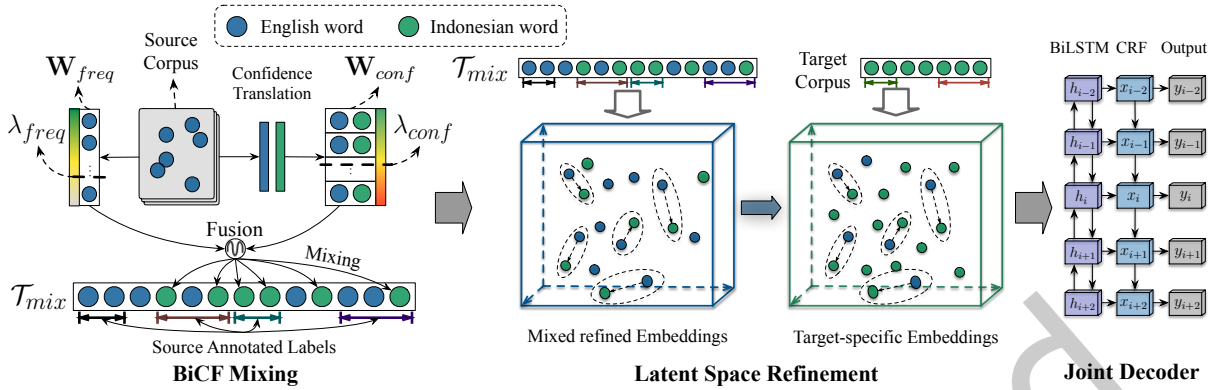

Fig. 2. Illustration of the proposed framework (BiCF), which consists of BiCF Mixing, Latent Space Refinement, and Joint Decoder. The frequency-word and confidence-word set in the first stage are derived from English dataset and confidence-translated parallel sentences, respectively. By fusion and mixing, the mixed data is generated. The cross-lingual space refinement module will generate a target-specific embedding model to represent Indonesian better. The final stage is to decode and output intent and slots jointly.

Given the set of English sentences $\mathcal{S} = \{s_1, s_2, ..., s_n\}$, we calculate TF-IDF [37, 38] for each word in the English dialogue corpus, as shown in Eq. 1:

$$\begin{cases} tf_{(i,j)} & = \frac{\mathcal{N}(w_i^s, s_j)}{\sum_k \mathcal{N}(w_k^s, s_j)} \\ idf_{(i)} & = \log \frac{|\mathcal{S}|}{1+\left|j : w_i^s \in d_j\right|} \\ tf\text{-}idf_{(i,j)} & = tf_{(i,j)} \times idf_{(i)} \end{cases} \tag{1}$$

where $\mathcal{N}(w_i^s, s_j)$ is the number of occurrences of the word w_i^s in the sentence s_j, and the denominator $(\sum_k \mathcal{N}(w_k^s, s_j))$ is the sum of occurrences of all terms w_k^s in the sentence $s_j \in \mathcal{S}$. $|S|$ represents the number of sentences and $\left|j : w_i^s \in d_j\right|$ denotes the number of sentences containing the word w_i^s. The frequency-word set $\mathbf{W}_{freq} = \langle (w_i^s, r_i), ..., (w_j^s, r_j) \rangle$ are obtained by sorting the output $(tf\text{-}idf_{(i,j)})$ from the TF-IDF algorithm, where r_i denotes the frequency score.

Then, we adopt small-scale high-quality parallel sentences (*i.e.*, 1K), translated by skilled bilingual translators, to generate the alignments of words by using *fast_align* [13]. Given a few English sentences and their corresponding confidently translated Indonesian sentences, the *fast_align* model uses a log-linear reparameterization of IBM Model 2 [8] to generate a set of confidence-word pairs $\mathbf{W}_{conf} = \langle (w_i^s, w_i^t), p_i \rangle, ..., \langle (w_j^s, w_j^t), p_j \rangle$ with Indonesian word and confidence score, denoted by w_i^t and p_i, respectively.

As shown in Algorithm 1, after selecting the words both over the frequency threshold λ_{freq} and the confidence threshold λ_{conf}, we then fuse words to generate the substitute words set $\mathbf{W}_{sub}$. *Thresh* function of line 1 and 2 in Algorithm 1 are designed as Eq. 2:

$$\widehat{\mathbf{W}} = Sort(\mathbf{W}(\cdot), \mathcal{P}(\cdot)) \odot \lambda_{(\cdot)} \tag{2}$$

where $\mathbf{W}(\cdot)$ denotes frequency-word set $(\mathbf{W}_{freq})$ or confidence-word set $(\mathbf{W}_{conf})$. $\mathcal{P}(\cdot)$ denotes frequency scores r_i or confidence score p_i. $\odot$ is selecting the top subset operation. And *Fusion* function in line 3 can be implemented as Eq. 3:

$$\widetilde{\mathbf{W}} = (\widehat{\mathbf{W}}_{freq} \odot \theta) \cap (\widehat{\mathbf{W}}_{conf} \odot (1 - \theta)) \tag{3}$$

Algorithm 1 BiCF Mixing

Input: $\mathcal{S}, \mathbf{W}_{freq}, \lambda_{freq}, \mathbf{W}_{conf}, \lambda_{conf}, \theta$
Output: $\mathcal{T}_{mix}$
1: $\widehat{\mathbf{W}}_{freq} \leftarrow \textsc{Thresh}(\mathbf{W}_{freq}, \lambda_{freq})$
2: $\widehat{\mathbf{W}}_{conf} \leftarrow \textsc{Thresh}(\mathbf{W}_{conf}, \lambda_{conf})$
3: $\widetilde{\mathbf{W}}_{sub} \leftarrow \textsc{Fusion}(\widehat{\mathbf{W}}_{freq}, \widehat{\mathbf{W}}_{conf}, \theta)$
4: $\mathcal{T}_{mix} \leftarrow \Phi$
5: **for** $s \in \mathcal{S}$ **do**
6: $\quad \widehat{s} \leftarrow s$
7: $\quad$ **for** $w^s \in s$ **do**
8: $\quad\quad$ **if** $w^s \in \widetilde{\mathbf{W}}_{sub}$ **then**
9: $\quad\quad\quad w^t \leftarrow \textsc{Get}(\widetilde{\mathbf{W}}_{sub}, w^s)$
10: $\quad\quad\quad \widehat{s} \leftarrow \textsc{Mixing}(\widehat{s}, w^s, w^t)$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: $\quad \mathcal{T}_{mix} \leftarrow \mathcal{T}_{mix} \cup \widehat{s}$
14: **end for**
15: **return** $\mathcal{T}_{mix}$

where θ is the hyper-parameter to adjust the ratio of two branch of word sets. Lines 4 to 13 in Algorithm 1 illustrate the mixing procedure. The algorithm first traverses each English word, which is in the substitute word set $\widetilde{\mathbf{W}}_{sub}$. Then, the English word is replaced with the corresponding Indonesian word w^t from the confidence word pair (w_i^s, w_i^t). We incrementally substitute the English word w^s of a temporarily copied sentence $\widehat{s}$ with the corresponding Indonesian word w^t. In this way, the mixed corpus $\mathcal{T}_{mix}$ is generated, consisting of both English words and Indonesian words.

3.2 Latent Space Refinement

We train and refine the initially pre-trained multilingual model (*i.e.*, Multilingual-BERT) on the mixed corpus $\mathcal{T}_{mix}$ with annotations from the source English dataset. This operation could update the embeddings of English words as well as the Indonesian words. Therefore, this stage allows our model to make use of English corpora and obtain a refined latent space to improve semantic representations. The multilingual latent space can be updated with the discriminative training process as Eq. 4:

$$\begin{cases} \Theta_{i+1}^l = \Theta_i^l - \eta^l \cdot \nabla_{\Theta_i^l} J(\Theta_i) \\ \eta^{i-1} = \xi \cdot \eta^i \end{cases} \tag{4}$$

where η^l denotes the learning rate of the l-th layer. Θ_i^l represents the parameters of the model at l-th layer in i step. $\nabla_{\Theta_i^l} J(\Theta_i)$ is the gradient of parameters Θ_i^l at l-th layer with regard to the model's objective function, *i.e.*, supervised by intent classification and slot-filling annotations in our model.

When the performance is stable on the training set (around 25 epochs in our experiments), we save the model that performs best on the validation set as the mixed refined embedding model, denoted in blue embedding space in the middle of Fig. 2. Then we feed fine-labeled Indonesian data into the mixed refined embedding model and transfer one more time to obtain a refined target-specific embedding model. In this way, by utilizing the English dataset, we generate a better representation latent space for Indonesian, *i.e.*, encoding each sentence into $\mathbb{R}^{1\times 768}$ representation feature vector.

3.3 Joint Decoder

The decoder of our framework performs two tasks, *i.e.*, intent classification and slot-filling sequence labeler, respectively. We apply Bi-directional Long Short-Term Memory (BiLSTM) and a conditional random fields (CRF) layer, as shown in Eq. 5, to predict the classifications for the input words [4, 5, 11, 47].

$$h_t = [f_l(\overrightarrow{h_{t-1}}, x_t); f_r(\overleftarrow{h_{t+1}}, x_t)] \Rightarrow BiLSTM \tag{5}$$

where f_l and f_r denote the hidden state of backward propagation and the hidden state of forward feeding in BiLSTM, respectively. $\overrightarrow{h_{t-1}}$ and $\overleftarrow{h_{t+1}}$ denote the hidden layer's output of the previous timestamp of the sentence forward and backward input, respectively. x_t denotes the input word embedding at the current moment. The final output h_t of each word embedding is generated by concatenating both $f_l(\overrightarrow{h_{t-1}}, x_t)$ and $f_r(\overleftarrow{h_{t+1}}, x_t)$.

And then CRF layer is appended to decode slot classes further and generate results of the framework. Specifically, given a sentence $S = \{w_1, w_2, ..., w_n\}$ and a label sequence $Y = \{y_1, y_2, ..., y_m\}$ predicted by BiLSTM for each word w_i. n and m indicate the number of words and labels, respectively. N indicates the number of all possible label paths. The principle of the CRF layer is to find the best path that is most similar to the real label path among all possible paths.

$$P_{total} = P_1 + P_2 + P_3 \cdots + P_N = e^{S_1} + e^{S_2} + e^{S_3} \cdots + e^{S_N} \tag{6}$$

$$S_k = \sum_{i=1}^{n} (X_{w_i, y_i} + T_{y_{i-1} \to y_i})(k = 1, 2, 3, \cdots, N) \tag{7}$$

$$Loss = -\log \frac{P_{RealPath}}{P_{total}} \tag{8}$$

where S_k denotes the score for each possible path. X_{w_i, y_i} denotes the confidence score of the word w_i on the label y_i. $T_{y_{i-1} \to y_i}$ represents the score from the transfer of the label y_{i-1} to y_i. Each pair of $T_{y_{i-1} \to y_i}$ is stored as parameters and is continuously updated with training in CRF layer. $P_{RealPath}$ represents the label path with gold annotations for each sentence. The proportion of the $P_{RealPath}$ increases with the training of the CRF, while the value of *Loss* decreases.

4 EXPERIMENTS

4.1 Dataset and Metrics

There has been a lack of available datasets for training natural dialogue understanding systems in regional low-resource languages, such as Indonesian [7, 24, 42]. In this section, we mainly introduce the self-established ID-WOZ dataset in detail. The statistics of ID-WOZ are reported in tables 1 to 3. We conduct experiments on two branches of datasets, MultiWOZ and our collected ID-WOZ, respectively. We take widely used F1-Score for evaluation to comprehensively compare the performance of our BiCF with baselines.

4.1.1 ID-WOZ construction. ID-WOZ is constructed to obtain highly natural conversations between a customer and an agent or a query information center focusing on daily life. We consider various possible dialogue scenarios ranging from basic requests like *hotel*, *restaurant*, to a few emergency situations such as *hospital* or *police*. Our dataset consists of nine domains, namely *plane, taxi, wear, restaurant, movie, hotel, attraction, hospital, and police*, most of which are extended domains that include the sub-task *Booking* (with the exception of *police*). In terms of collection and annotation, we adopt the Wizard-of-OZ [22] dialogue-collecting approach, which has been shown to be effective for obtaining a high-quality corpus at relatively low costs and with a small-time effort. Following the success of MultiWOZ [3], we conduct a large-scale corpus of natural human-human conversations

Fig. 3. An example of editing template interface. We design a corresponding generic template for the topic to which the dialogue belongs. First, we add the topic and result buttons to indicate whether the conversation on this topic has ended successfully. Second, we define multiple Indonesia alternative buttons for intent and slot and English buttons for action and domain. The annotation platform display the corresponding templates based on the topic selected by annotators.

on a similar scale. Based on the given templates for various domains, users and wizards generate conversations using heuristic-based rules to prevent the overflow of information. We design and develop a collection-annotation pipeline platform with a user-friendly structure for building the dataset.

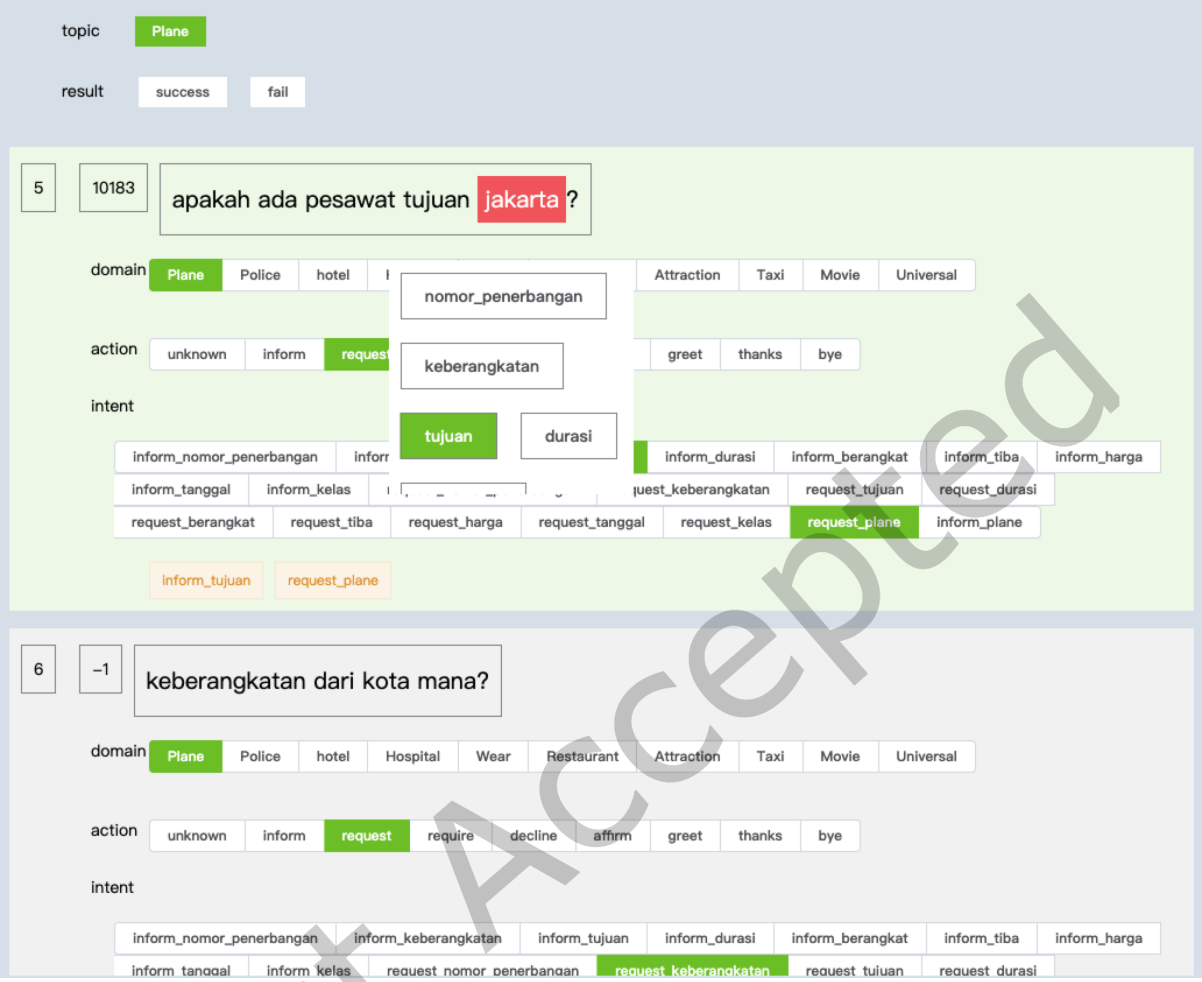

Fig. 4. An example of the annotation procedure. The annotator first click the topic button and obtain the corresponding template. For domain/action/intent classification, the annotator could click these multiple labels, defined before for each attributes. As for slot-filling, our platform provides a fashion approach: click and underscore the content (the red part) and select its slot type (under the red part), pop-out automatically when any words are picked.

In order to accelerate and optimize the process of collection and annotating, we design and develop a pipeline platform. Our platform consists of three stages, "collection - annotation - statistics & analysis", which are executed synchronously after the initialization process. We divided a number of well-trained annotators (*i.e.*, 80 local people, 70 of whom spoken ID as their native language, 10 of whom were bilingual citizens, plus 2 main organizers) into two groups to produce dialogue and annotation. A quarter of annotators (*i.e.*, 20) are trained following the templates we provide to play the wizard role. After collecting 1k dialogues initially (about one week), while the collecting conversation is still ongoing, the second group of annotators (*i.e.*, 62) joins in to work towards the detailed full-labeled corpus, including domains, actions, intents, and slots. As shown in Fig. 3, we design corresponding annotation templates for each topic and define multiple options for intent, slot value, action, and

Table 1. Comparison of ID-WOZ with other related datasets in several statistics metrics.

Dataset	ID Chat	Dyadic Chat	MultiWOZ	ID-WOZ
Domains	None	None	7	9
Language	ID	ID	En	ID
Total # dials	300	79	8, 438	9, 189
Total # tokens	150, 000	3164	1, 520, 970	1, 551, 591
Total # utters	1, 000	158	142, 974	251, 184
Avg. # turns	3	3	13.68	13.67
Avg. # slots	-	-	25	8.8

domain. Specific labeling process is depicted in Fig. 4. First, the annotator decides which topic this round of dialogue belongs to. The platform will display the template associated with the topic. Then, annotators use tools to click buttons for determining domain, action, and intent. When annotating a slot value, the annotator uses the mouse to underscore corresponding words and select a slot value from the pre-defined options.

The quality is assured in three processes, namely "scripts-checking", "cross-checking", and "supervisor-checking". Specifically, the scripts can filter hypothesis cases which have potential flaws such as vacant labels or being malformed. For cross-checking process, annotators are assigned not only to fresh unlabeled annotation tasks, but also to a few sampling labeled cases (*i.e.*, 20%) from their peers. Cases that pass the cross-checking procedure will be sampled and handed over to supervisors (*i.e.*, the two organizers), who are more familiar with the details of the entire annotation task to control the overall consistency and accuracy of the annotation. We adopt the inter-annotator agreement (IAA) [15] to measure how well our recruited annotators can make the same annotation decision for a certain category further, as follows:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{9}$$

where p_o and p_e denote the relative observed agreement among raters and hypothetical probability of chance agreement, respectively. The average score of our dataset is 0.834.

4.1.2 Statistics and Analysis. Table 1 compares our dataset with existing datasets in Indonesian as well as the English dialogue dataset MultiWOZ [3]. ID Chat [24] is the first publicly available Indonesian chat corpus, and draw a few related research on the Indonesian Language dialogue [7]. Dyadic Chat [42] is another public chat corpus on Indonesian, which focuses on the dyadic term. Dyadic is a term that describes the relationship between two people, such as a romantic relationship between two people. Compared with these small-scale datasets, ID-WOZ is the first to contain large-scale (about ten thousand dialogues across multiple domains) corpus focusing on general task-oriented chat.

MultiWOZ [3] is a large-scale multi-domain task-oriented English dialogue dataset, including seven distinct domains (*taxi, restaurant, hotel, attraction, hospital, police, and train*) and fine-labeled actions and slots in the spoken language understanding stage. Considering the regional cultural background, our collected dataset contains a few more general domains (*i.e.*, *wear, movie, plane*) and more corresponding slots types, such as *clothes type, movie genre, movie synopsis*.

Even though there are several off-the-shelf pre-trained BERT models for rich-resource languages such as English and Chinese, pre-trained language-specific models for low-resource languages like Indonesian, are still not available to our knowledge. We release a pre-trained model for Indonesian named ID-BERT as another resource contribution. Despite the fact that most related work [34, 40] relies on the pre-trained Multilingual-BERT model and fine-tunes it for low-resource languages, our main goal is to build a relatively reliable dialogue system and

Table 2. Statistics for total number in four domains.

Dataset	Domains	# Sentences	# Slots	# Intent
MultiWOZ	**Restaurant**	62, 703	28, 351	41, 177
	Hotel	64, 284	25, 985	42, 434
	Taxi	48, 080	7, 160	28, 976
	Attraction	55, 186	21, 004	34, 053
ID-WOZ	**Restaurant**	28, 095	5, 809	22, 312
	Hotel	30, 865	8, 720	24, 694
	Taxi	28, 178	6, 038	22, 168
	Attraction	36, 523	9, 198	29, 513

Table 3. Comparison of our dataset to similar well-known datasets.

Dataset	Twitter	Ubuntu	Sina Weibo	WOZ 2.0	Frames	M2M	MultiWOZ	ID-WOZ
Domains	Unrestricted	Ubuntu	Unrestricted	Unrestricted	Unrestricted	Unrestricted	7	9
Language	English	English	Chinese	English	English	English	English	Indonesian (+En)
Total # dialogues	1.3M	930K	4.5M	600	1, 369	1, 500	8, 438	9,189 (+1k)
Total # tokens	-	-	-	50, 264	251, 867	121, 977	1, 520, 970	1, 551, 591
Avg. # Turns	2.10	7.71	2.3	7.45	14.60	9.86	13.68	13.67
Avg. # slots	-	-	-	4	61	14	25	8.8

experiment with how to bridge the gap between language-specific BERT and the multilingual-BERT. Therefore, we put in effort to train ID-BERT for comparing appearance on spoken language understanding or furthermore tasks. We pre-train a BERT for Indonesian from scratch using approximately 3.3 billion tokens from the document-level corpus of Indonesian websites, which covers news reports, research papers, daily articles, and other text genres. The size of our ID-BERT vocabulary is 0.9M, which is much larger than Multilingual-BERT (0.12M). We believe that this size of the vocabulary is sufficient to cover most of the scenarios of daily multi-domain task-oriented dialogue in Indonesian. The training takes one week by using Google Cloud TPU v3_8, and our ID-BERT (Cased, L=12, H=768, A=12) is eventually obtained.

We take MultiWOZ [3] as the English dataset and our collected ID-WOZ as the target language Indonesian dataset. As the *hospital* and *police* domains in MultiWOZ contain very few dialogues (5% of total dialogues) and only appear in the training dataset, we choose to ignore them in our experiments, following [48]. The *train* domain is invalid in Indonesian data because it reflects the cultural difference between English and Indonesia. Therefore, we only adopt four domains as the main experiment *restaurant, hotel, taxi, attraction* shared by MultiWOZ and ID-WOZ. Statistics of them are shown in Table 2. In order to suit the testing set, we have to merge the annotations of English data with the Indonesian dataset, thereby abandoning a few types of labels, such as *reference, choice* in MultiWOZ. After processing, the statistics for the four domains in two datasets are reported in Table 2. Table 3 shows a comparative study of differences between our dataset and similar well-known datasets. All of the experiments are evaluated on the same test set from ID-WOZ (1K dialogues, 250 dialogues for each domain), which suits the local cultural background. We use the F1 score as the evaluation metric, which is calculated by the Precision and Recall.

4.2 Experimental Settings

There are several branches of methods to utilize English datasets and pre-trained models, *i.e.*, **M**achine **T**ranslation based (**MT**); **M**ulti**l**ingual pre-trained embedding model with **En**glish corpus (**MLEn**); and our proposed **BiCF**.

We also compared our method and the following popular state-of-the-art cross-lingual dialogue systems (**TLM**, **DST**, **Seq2Seq-DU**, **DSS-DST**, and **DiCoS-DST**).

1) MT. We adopt the machine translation preprocessing method and extract word embeddings (*i.e.*, $\mathbb{R}^{1\times768}$) by random initiation, pre-trained multilingual-BERT (ML-BERT), and our pre-trained ID-BERT. We also take Indonesian-fastText (ID-fastText) [21], Transformer [46] and Indonesian-Word2vec (ID-Word2vec) into comparison. ($\mathbb{R}^{1\times300}$)

2) MLEn. We adopt three pre-trained multilingual word embedding models in this baseline, namely multilingual fastText (ML-fastText) [21], multilingual Word2vec (ML-Word2vec) [9], and multilingual-BERT (ML-BERT) [10]. By extracting the embeddings of MultiWOZ and ID-WOZ, we encode each sentence into $\mathbb{R}^{1\times300}$, $\mathbb{R}^{1\times300}$ and $\mathbb{R}^{1\times768}$ dimensions, respectively.

3) TLM. Previous works have used cross-lingual models or machine translation to generate low-resource language data for task-oriented dialogue systems. Moghe et al. [32] use movie subtitle datasets as parallel related data to fine-tune the pre-trained multilingual models. This enhanced transfer method can achieve better performance of dialogue state tracking with a few target language data or zero-shot setup.

4) DSTC9. Lin et al. [25] use a multilingual pre-trained seq2seq model and high-resource training dataset to study the transfer-ability of the dialogue state tracking model of DSTC9 [17]. Researchers also conducted experiments with a variety of training strategies, including joint-training or pre-training, and different datasets (cross-lingual or cross-ontology) to further confirm the effectiveness of the cross-lingual dialogue models. We adopt multilingual-BERT (ML-BERT) and Indonesia-BERT(ID-BERT) as the pre-trained word embedding models in this baseline.

5) Seq2Seq-DU. Feng et al. [14] propose a new dialogue state tracking module (Seq2Seq-DU) comprised of two BERT-based encoders, one attender, and a decoder. Two BERT-based encoders are capable of generating utterance embeddings and schema embeddings in the dialogue. The attender and decoder can utilize word embeddings of utterances and schemas based on BERT to jointly model intent classification and slot-filling in both DST and NLU tasks.

6) DSS-DST. Traditional slot-filling methods treat the slot value in each round of dialogue equally, which may lead to unpredictable errors. Gue et al. [19] focus on slot-filling and propose a two-stage model (DSS-DST) to address the above problem. DSS-DST consists of the Dual Slot Selector and the Slot Value Generator. Each slot value is determined by the Dual Slot Selector whether to update or inherit the dialog of the last round. The Slot Value Generator updates the corresponding slot value based on the decision.

7) DiCoS-DST. The consistent dialogue history information is utilized by existing works on dialogue state when updating slot value, which can lead dialogue systems to produce inaccurate results. Therefore, Guo et al. [20] further propose DiCoS-DST to generate different historical dialogue for different slots as information for updating slot values. DiCoS-DST first retrieves the correlation between each round of dialogue and slot values by learning the whole dialogue history. Second, the dialogue state is generated only using the dialogue history with high relevance score.

8) BiCF. We generate about 1.5K confident word pairs from MultiWOZ and 1K translated parallel sentences. For our method BiCF, the training process converges after 20 epochs. It reaches 91.13, 87.84/ 90.17, 82.09/ 93.37, 82.98/ 89.55, 85.54 for the F1 score of intent classification and slot-filling on the MultiWOZ validation set of *restaurant, hotel, taxi, attraction* domains, respectively. And then the Indonesian training data of ID-WOZ is fed to refine the Indonesian embedding model.

4.3 Development Experiments

We feed 16K Indonesian sentences of ID-WOZ to each model and validate their performance on the same test set of ID-WOZ. In our implementation, five-fold cross-validation is employed to investigate the optimal parameter

Table 4. Experimental comparison on ID-WOZ dataset. ("†" denotes the significance testing, $p\text{-}value < 0.05$.)

Methods + Emb. \ Domains		Restaurant		Hotel		Taxi		Attraction	
		Intent	Slots	Intent	Slots	Intent	Slots	Intent	Slots
MT	**Random Init**	85.48	74.36	82.73	73.49	89.15	80.22	89.64	86.26
MT	**ID-fastText**	86.03	75.27	83.17	74.03	89.82	80.28	90.02	86.88
MT	**ID-Word2vec**	88.22	76.70	86.33	74.11	89.91	81.81	91.55	86.90
MT	**Transformer**	90.13	79.91	91.89	74.27	90.25	82.11	92.85	87.16
MT	**ML-BERT**	91.63	79.22	92.52	73.83	91.20	82.34	93.77	87.31
MT	**ID-BERT**	*92.37*	*81.88*	*93.78*	*75.79*	*91.76*	*83.59*	*94.07*	*89.63*
MLEn	**ML-fastText**	86.00	76.11	83.10	74.91	89.22	80.88	90.31	86.93
MLEn	**ML-Word2vec**	88.22	77.70	86.33	74.11	89.91	81.81	91.55	86.90
MLEn	**ML-BERT**	*90.42*	*79.79*	*92.01*	*74.28*	*90.47*	*82.91*	*93.18*	*87.77*
TLM	**ML-BERT**	87.65	74.78	81.17	75.36	89.08	81.40	92.32	85.89
TLM	**ID-BERT**	*90.71*	*76.16*	*90.58*	*75.89*	*91.02*	*84.29*	*92.80*	*86.13*
Seq2Seq-DU	**ML-BERT**	*91.10*	*80.21*	*90.52*	*74.40*	*92.01*	*85.78*	*91.67*	*85.20*
Seq2Seq-DU	**ID-BERT**	*92.11*	*81.78*	*91.49*	*75.03*	*92.53*	*87.43*	*92.12*	*87.81*
DSTC9	**ML-BERT**	*89.98*	*80.39*	*90.15*	*74.87*	*89.56*	*84.94*	*91.76*	*87.38*
DSTC9	**ID-BERT**	*90.68*	*81.14*	*92.26*	*75.07*	*91.66*	*87.79*	*92.31*	*89.83*
DSS-DST	**ML-BERT**	*90.44*	*81.60*	*91.82*	*75.34*	*90.01*	*85.55*	*91.29*	*87.77*
DSS-DST	**ID-BERT**	*91.37*	*82.71*	*92.67*	*75.07*	*91.45*	*88.36*	*92.80*	*88.29*
DiCoS-DST	**ML-BERT**	*91.73*	*82.64*	*92.38*	*74.59*	*91.76*	*86.14*	*93.41*	*90.23*
DiCoS-DST	**ID-BERT**	*92.27*	*82.31*	*93.42*	*75.94*	*92.54*	*88.35*	*93.78*	*91.18*
BiCF	**ML-fastText**	86.21	76.16	83.31	75.01	90.24	82.58	90.84	87.23
BiCF	**ID-fastText**	87.08	76.34	84.21	75.79	90.83	82.92	91.52	87.67
BiCF	**ML-Word2vec**	88.80	77.91	87.12	74.24	90.01	82.87	91.58	87.03
BiCF	**ID-Word2vec**	88.92	78.84	88.52	74.35	90.31	83.15	91.82	87.49
BiCF	**ML-BERT**	*92.92*†	*82.84*†	*94.30*†	*76.95*†	*92.23*†	*90.45*†	*94.80*†	*90.44*†
BiCF	**ID-BERT**	**93.02**†	**82.91**†	**94.73**†	**77.15**†	**92.73**†	**91.03**†	**94.88**†	**90.74**†

setting within training datasets ($learning\_rate = e^{-3}, batch\_size = 64, dropout\_rate = 0.1, optimizer = SGD$). To verify the stability of the proposed method, we run the experiments five times for each set of parameter settings and compare their mean performance, as reported in Table 4.

We conduct a series of experiments by feeding batches of annotated Indonesian data (*i.e.*, 1K sentences, 2K sentences, 4K sentences, ..., full-scale). We pick the results of *restaurant*, *hotel*, *taxi*, and *attraction* domains in Fig. 7 and Fig. 8, as they are widely usable domains and have the most scale of dialogue data and annotations both in MultiWOZ and ID-WOZ. The entire annotated dataset, experiment results, and codes are detailedly reported in Table 5 and Code 1. We also conduct a comparison experiment for Multilingual-BERT (ML-BERT) and ID-BERT on all domains of full-scale ID-WOZ, as reported in Table 6.

4.4 Results Analysis

The results of the method in Section 4.2 are shown in Table 5, with English data of MultiWOZ and 16k Indonesian data of ID-WOZ. The method of machine translation based methods (MT + ML-BERT/ ID-BERT) surpass multilingual model with English data (MLEn + ML-BERT) on the intent classification task, outperforming by about 1.21%, 1.95%; 0.51%, 1.77%; 0.73%, 1.29% and 0.59%, 0.89% on F1 score for *restaurant, hotel, taxi, attraction*, respectively. The main reason is that the machine translation methods enjoy much more Indonesian sentences with corresponding intention labels. However, on the slot-filling task, the machine translation methods are weaker. As shown in Fig. 6 and Fig. 5, the machine translation methods suffer from invalid or mismatching labels after translation and

Table 5. Performance comparison of different methods on the selected MultiWOZ and ID-WOZ with different amounts of feeding ID-WOZ data.

Methods	ID-WOZ	Restaurant		Hotel		Taxi		Attraction	
		Intent	Slots	Intent	Slots	Intent	Slots	Intent	Slots
MT (ID-BERT)	ID-WOZ-1000	87.33	56.67	90.83	60.14	86.66	40.28	90.01	62.29
	ID-WOZ-2000	88.97	59.74	91.67	66.63	86.98	59.88	91.02	76.05
	ID-WOZ-4000	90.01	70.67	93.23	69.35	89.50	74.09	93.05	83.73
	ID-WOZ-8000	91.67	80.57	93.65	73.75	90.63	82.09	93.95	87.96
	ID-WOZ-16000	92.37	81.88	93.78	75.79	91.76	83.59	94.07	89.63
	ID-WOZ-All	92.25	81.87	93.42	75.65	91.67	82.17	94.25	90.40
MLEn (ML-BERT)	ID-WOZ-1000	84.11	55.59	89.73	60.41	82.93	22.66	89.32	64.44
	ID-WOZ-2000	86.57	56.86	91.51	65.26	86.37	40.63	91.57	70.56
	ID-WOZ-4000	89.57	68.99	91.90	72.20	87.93	46.42	92.58	84.22
	ID-WOZ-8000	90.93	73.37	93.42	75.15	88.08	58.63	94.03	86.85
	ID-WOZ-16000	90.92	74.24	93.28	75.89	88.12	64.12	94.11	87.67
	ID-WOZ-All	90.89	75.36	93.23	75.97	88.34	64.86	94.25	88.71
TLM (ML-BERT)	ID-WOZ-1000	84.27	70.34	76.29	69.58	84.30	78.44	85.73	81.02
	ID-WOZ-2000	85.12	71.62	77.38	70.41	86.56	79.23	87.73	82.11
	ID-WOZ-4000	86.02	72.30	78.35	71.50	87.48	80.02	89.49	83.57
	ID-WOZ-8000	86.51	73.75	80.43	73.90	88.62	80.42	91.67	84.14
	ID-WOZ-16000	87.65	74.78	81.17	75.36	89.08	81.40	92.32	85.89
	ID-WOZ-All	88.18	74.89	81.74	75.73	89.68	82.24	92.19	85.64
Seq2Seq-DU (ML-BERT)	ID-WOZ-1000	85.83	76.02	87.34	68.79	86.25	81.46	86.13	80.24
	ID-WOZ-2000	87.31	77.35	88.23	70.94	87.47	82.26	87.80	82.44
	ID-WOZ-4000	88.34	78.51	89.75	72.67	88.72	83.53	89.48	83.85
	ID-WOZ-8000	90.71	79.63	90.10	73.62	90.11	84.65	90.40	84.43
	ID-WOZ-16000	91.10	80.21	90.52	74.40	92.01	85.78	91.67	85.20
	ID-WOZ-All	92.04	81.16	91.07	74.24	92.76	86.34	91.55	85.45
DSTC9 (ML-BERT)	ID-WOZ-1000	84.54	75.58	85.80	70.67	84.28	79.40	86.56	82.15
	ID-WOZ-2000	86.36	77.04	87.17	71.49	85.12	81.41	87.79	83.27
	ID-WOZ-4000	88.18	78.72	88.56	72.48	86.31	82.70	89.43	84.83
	ID-WOZ-8000	89.07	79.83	89.24	73.67	88.39	84.10	91.13	86.34
	ID-WOZ-16000	89.98	80.39	90.15	74.87	89.56	84.94	91.76	87.38
	ID-WOZ-All	90.16	81.04	90.79	75.52	90.37	85.41	91.29	87.15
DSS-DST (ML-BERT)	ID-WOZ-1000	86.32	76.17	86.61	69.13	84.41	78.23	86.26	84.98
	ID-WOZ-2000	87.28	77.53	87.26	70.85	85.35	80.75	87.02	85.01
	ID-WOZ-4000	88.28	79.48	88.41	71.09	86.62	82.13	88.48	86.72
	ID-WOZ-8000	89.02	80.31	90.95	73.78	89.73	84.19	90.61	85.21
	ID-WOZ-16000	90.44	81.60	91.82	75.34	90.01	85.55	91.29	87.70
	ID-WOZ-All	90.35	81.71	91.67	75.41	90.87	85.24	91.38	87.84
DiCoS-DST (ML-BERT)	ID-WOZ-1000	84.73	77.34	87.79	68.26	84.40	80.30	86.78	84.65
	ID-WOZ-2000	87.37	79.52	89.49	70.75	86.25	81.92	88.61	86.58
	ID-WOZ-4000	89.45	80.57	90.77	71.08	88.34	83.56	90.76	87.48
	ID-WOZ-8000	90.17	81.47	91.03	73.50	90.79	85.21	92.35	89.05
	ID-WOZ-16000	91.73	82.64	92.38	76.59	91.76	86.14	93.41	90.23
	ID-WOZ-All	91.80	82.41	92.24	76.65	91.59	86.20	93.37	90.85
BiCF (ML-BERT)	ID-WOZ-1000	84.23	59.92	87.66	59.81	84.78	72.31	87.87	69.41
	ID-WOZ-2000	86.69	66.67	90.35	61.93	86.69	75.25	90.04	80.05
	ID-WOZ-4000	89.07	76.10	91.77	68.85	88.82	81.87	92.72	85.52
	ID-WOZ-8000	92.23	78.34	93.13	73.71	91.55	86.48	93.46	88.41
	ID-WOZ-16000	**92.92**	**82.84**	**94.30**	**76.95**	92.23	**90.45**	**94.80**	90.44
	ID-WOZ-All	92.60	82.67	94.24	76.91	**92.25**	89.43	94.77	**90.45**
ID-BERT	ID-WOZ-All	92.22	82.14	93.91	76.88	91.97	88.13	93.96	90.20

some annotations are invalid in translation tasks. Overall, our proposed framework (BiCF + ML-BERT / ID-BERT) performs better than others in both tasks, as we are capable of utilizing the English intention labels and correct slot-filling annotations effectively. And from Table 6, we can see that ID-BERT outperforms ML-BERT across all domains, demonstrating Indonesian-specific word-embedding model (ID-BERT) is capable of representing more information and semantic knowledge than the general multilingual model (ML-BERT) in all domains.

4.5 Effectiveness of Using ID-WOZ

The statistics line chart is shown in Fig. 7, where the four upmost sub-graphs denote intent classification, the four downmost sub-graphs denote slot-filling and the red line is the performance of ID-BERT baseline. The detailed results and all of the line charts of rest domains are in Fig 8.

Table 6. Experimental comparison of ML-BERT and ID-BERT on full-scale ID-WOZ.

Domains	ML-BERT		ID-BERT	
	Intent	Slots	Intent	Slots
Restaurant	91.07	77.68	92.22	82.14
Hotel	92.78	74.91	93.91	76.88
Taxi	90.84	82.91	91.97	88.13
Attraction	93.25	88.04	93.96	90.20
Plane	91.36	92.77	93.42	93.11
Police	90.02	88.89	92.78	90.07
Movie	90.57	86.14	91.76	87.98
Hospital	92.64	84.15	93.85	86.09
Wear	90.77	87.02	91.80	88.34

Christ	's	College	is	located	in	the	centre	at	saint	Andrew	's	street	.
B-name	I-name	I-name	O	O	O	O	B-area	O	B-addr	I-addr	I-addr	I-addr	O
Chunks: {'name': 'Christ's College', 'area': 'centre', 'addr': 'saint Andrew 's street'}													
Christ	's	College	terletak	di	pusat	di	jalan	suci	dan	suci	.		
O	O	O	O	O	B-area	O	B-addr	I-addr	O	B-addr	O		
Translated Chunks: {'name': 'Perguruan Tinggi Kristus', 'area': 'pusat', 'addr': 'jalan suci orang suci'}													
Christ	's	College	terletak	di	pusat	jalan	saint	Andrew	's	.			
B-name	I-name	I-name	O	O	B-area	B-addr	I-addr	I-addr	I-addr	O			
Human Annotated Chunks: {'name': 'Christ's College', 'area': 'pusat', 'addr': 'jalan saint Andrew 's'}													

Fig. 5. Illustrate the situation that annotations getting invalid in the machine translation.

tiket pesawatnya kalo beli besok berapaan ya?	request_price
how much is the airplane ticket if i buy it tomorrow?	request_price
What are the plane tickets for tomorrow?	request_type
kalo mau pesen tiket, lewat mana mas pesennya?	request_ticket
if i want to order the ticket, how do i order it?	request_ticket
if you want to order a ticket, where do you order it?	request_location

Fig. 6. Illustration of the mistakes from machine translation. The green sentence is true mean, and the red is the result of machine translation. These two examples show that a tiny mistake that happened during translation may cause complete misunderstanding.

1) MT relies heavily on the quality of translation. We conduct the BLEU [33] test for the entire MultiWOZ, and the performance of translation is 28.46 (BLEU-5) on 30k sentences. However, during the translation of the dialogue, one incorrect word would cause misunderstanding. Several examples are shown in Fig. 6. Different sentences in English may be translated from the same source sentence in Indonesian. In the first case, the true meaning is requesting "*how much*" but the model may misunderstand the customer's intent in requesting the type of plane ticket. And in the second, the customer is wondering "how" to order a ticket, but the translator gives the result that the customer's request is "*request location*". Based on Fig. 7 and Fig. 8, when the scale of ID-WOZ is negligible, the machine translation has a large advantage on intent classification but performs badly on slot-filling. The reason is that the MT method has the ability to adjust or reset the grammar and syntactic

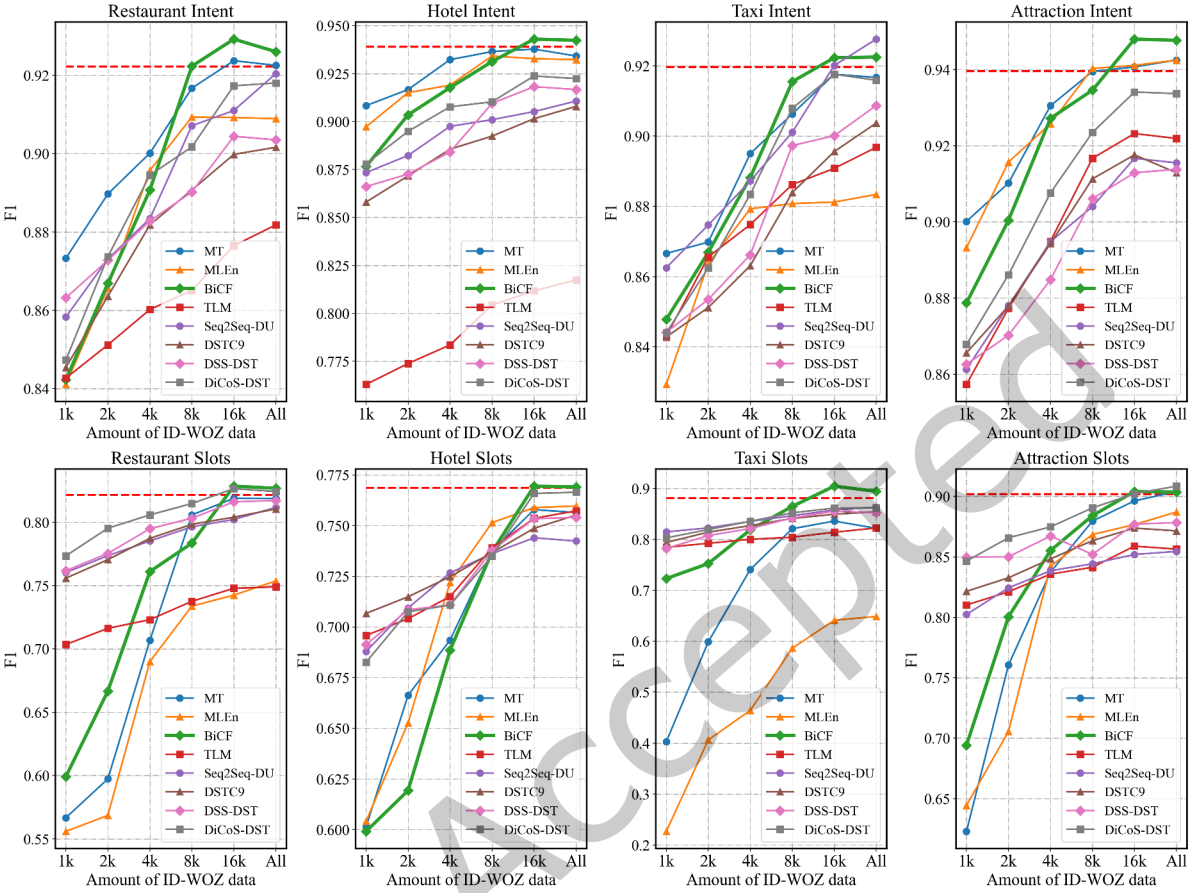

Fig. 7. The comparison of different methods on four domains.

structure of the target language, whose characteristic leads to the bad consequences that make the English slot labels dislocated, invalid and wrong.

2) MLEn only learns semantic information from English data at the beginning, which causes low accuracy on intent classification than others. When feeding this model with ID-WOZ, it has weakness coming from the English data because the large-scale English data shrinks the feeding ID-WOZ data. This method has strength in slot-filling when the comparison is under a small scale of ID-WOZ. Because labels of slot-filling in the English data are accurate and complete. But the performance does not improve when more ID-WOZ data is further used, which shows ML-BERT has a limitation on reaching higher performance. Overall, this method is not recommended for building stable low-resource language dialogue understanding models even with gold annotated data.

3) TLM enhances the transfer learning process by fine-tuning pre-trained models. Using 16,000 Indonesian data and a pre-trained model (ML-BERT), TLM achieves F1 Score of 87.65/74.78%, 81.17/75.36%, 89.08/81.40%, and 92.32/85.89% on four domains (Restaurant, Hotel, Taxi, Attraction), respectively. As reported in Table 4, using the same amount of training data, but replacing the pre-trained model with ID-BERT, the performance of TLM is improved by 3.06/1.38%, 9.41/0.53%, 1.94/2.89%, and 0.48/0.24% in four domains, respectively. When using

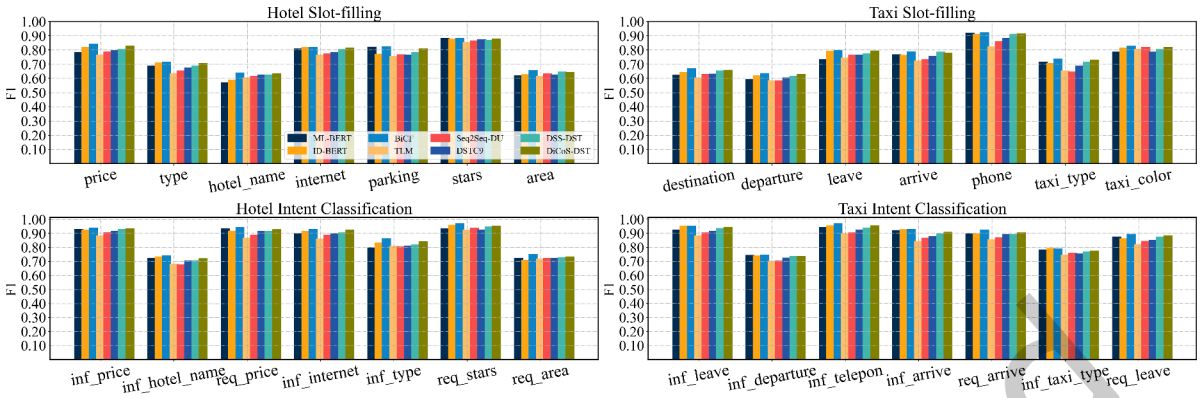

Fig. 8. The comparison of different methods on four domains.

few Indonesian data, TLM performs poorly compared to other baseline models. As the amount of data used for training gradually increases, the performance of TLM improves to a certain extent.

4) Seq2Seq-DU is superior to TLM in Restaurant, Hotel, and Taxi, but slightly worse than TLM in Attraction. The amount of Indonesian data gradually increased from 1,000 to 16,000, and F1 Score of Seq2Seq-DU improved by 6.21/5.14%, 3.73/5.45%, 6.51/4.88%, and 5.42/5.21%. With the pre-trained model (ID-BERT), the performance of Seq2Seq-DU is further improved to 92.11/81.78%, 91.49/75.03%, 92.53/87.43%, and 92.12/87.81%.

5) DSTC9 uses a pre-trained Seq2Seq and resource-rich training dataset to learn the transfer-ability of traditional dialogue state tracking models. Compared with TLM, the performance of DSTC9 is generally similar, but the performance on the slot-filling is better. With the support of ID-BERT, DSTC9 reaches improvements of 0.7/0.75%, 2.11/0.2%, 2.1/2.85%, and 0.55/2.45%, respectively. This further demonstrates that ID-BERT is capable of capturing the semantic information of Indonesian.

6) DSS-DST achieves satisfactory performance on four domains with the F1-Score of 90.44/81.71%, 91.82/75.41%, 90.87/85.55%, and 91.38/87.84%. By learning from the dialogue history, DSS-DST utilizes a combination of extraction methods and classification methods for slot filling. As the amount of Indonesian language data used for training increases, the performance of DSS-DST improves and stabilizes when the number of datasets exceeds 16,000.

7) DiCoS-DST is an extension of DSS-DST that uses the dialogue history of different rounds to update the slot value with a strong correlation. Compared with DSS-DST, DiCoS-DST respectively reaches improvements of 1.36/0.93%, 0.56/1.25%, 1.75/0.59%, and 2.12/2.53%. As with DSS-DST, the performance of DiCoS-DST can be continuously improved as the amount of training data increases. However, few Indonesian dialogue data cause low F1-Score on four domains. The F1-score of DiCoS-DST varied by 7.07/5.07%, 4.45/8.39%, 7.19/5.9%, and 6.59/6.2% in the four domains when using all training data and only 1,000 training data, respectively.

8) BiCF does not outperform machine translation when the scale of fed Indonesian data is negligible on the intent classification. When the scale of ID-WOZ data gets larger, the strength of BiCF becomes more obvious. It starts to outperform significantly better than the other methods while the Indonesian data grows. It is capable of avoiding misunderstanding caused by translation and mitigating the shrink effect of the English corpus, which makes it achieve the best performance and even better than the baseline ID-BERT, when the ID-WOZ data reaches around 16k for *restaurant, hotel, taxi, attraction* domains on the intent classification, *i.e.*, 92.92%, 94.30%, 92.23%, 94.80% on F1 score, respectively. This method outperforms other methods on slot-filling when the ID-WOZ data fed is negligible. Not only it makes use of correct slot-filling annotations from the English dataset, but it can also reduce the bad effects of large-scale English corpus. The accuracy reaches 82.84%, 76.95%, 90.45%, 90.44% on

F1 score for *restaurant, hotel, taxi, attraction* on the slot-filling, respectively. Fig. 8 reports the results of three methods trained by 16K of ID-WOZ. It shows that the cross-lingual method performs better than others when the slots need more words to describe.

5 CONCLUSION AND FUTURE WORK

This paper presents a Bi-Confidence-Frequency Cross-lingual Transfer framework, which consists of BiCF Mixing, Cross-lingual Space Refinement, and Joint Decoder, to address the challenge of the lack of sufficient training data for existing dialogue models based on low-resource languages. BiCF is capable of utilizing English datasets to generate code-mixed data. Then, cross-lingual space refinement and joint decoder are used to train the dialogue understanding models based on the high-quality mixed data from results of BiCF mixing. We also collect and annotate a richer-domain dataset in Indonesian (ID-WOZ) for experiments. Results demonstrate that our framework enjoys the enriched and accurate English dataset, performs effectively, and achieves reliable performance on intent detection and slot filling. In the further, we consider building a large Indonesian dialogue dataset and an upgraded ID-BERT to model a large-scale dataset for better experimental results.

ACKNOWLEDGMENTS

This paper is supported by the Science and Technology Innovation 2030 Major Project of China (No. 2021ZD0113302) and National Natural Science Foundation of China (No. 62076081, No. 61772153 and No. 61936010).

REFERENCES

[1] Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics* 4 (2016), 431–444.
[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 451–462.
[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278* (2018).
[4] Hongshen Chen, Yue Zhang, and Qun Liu. 2016. Neural network for heterogeneous annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 731–741.
[5] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications* 72 (2017), 221–230.
[6] Yong Cheng. 2019. Semi-supervised learning for neural machine translation. In *Joint Training for Neural Machine Translation*. Springer, 25–40.
[7] Andry Chowanda and Alan Darmasaputra Chowanda. 2017. Recurrent neural network to deep learn conversation in indonesian. *Procedia computer science* 116 (2017), 579–586.
[8] Michael Collins. 2011. Statistical machine translation: IBM models 1 and 2. *Columbia Columbia Univ* (2011).
[9] Gerard de Melo. 2017. Multilingual vector representations of words, sentences, and documents. In *Proceedings of the IJCNLP 2017, Tutorial Abstracts*. 3–5.
[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).
[11] Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734* (2016).
[12] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 339–348.
[13] Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. (2013).
[14] Yue Feng, Yang Wang, and Hang Li. 2020. A Sequence-to-Sequence Approach to Dialogue State Tracking. *arXiv preprint arXiv:2011.09553* (2020).
[15] Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin* 72, 5 (1969), 323.
[16] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[17] 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* (2020).
[18] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*. 1234–1244.
[19] Jinyu Guo, Kai Shuang, Jijie Li, and Zihan Wang. 2021. Dual Slot Selector via Local Reliability Verification for Dialogue State Tracking. *arXiv preprint arXiv:2107.12578* (2021).
[20] Jinyu Guo, Kai Shuang, Jijie Li, Zihan Wang, and Yixuan Liu. 2022. Beyond the Granularity: Multi-Perspective Dialogue Collaborative Selection for Dialogue State Tracking. *arXiv preprint arXiv:2205.10059* (2022).
[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).
[22] John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2, 1 (1984), 26–41.
[23] San Kim, Jin Yea Jang, Minyoung Jung, and Saim Shin. 2021. A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 352–365.
[24] Fajri Koto. 2016. A publicly available indonesian corpora for automatic abstractive and extractive chat summarization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 801–805.
[25] Yen-Ting Lin and Yun-Nung Chen. 2021. An empirical study of cross-lingual transferability in generative dialogue state tracker. *arXiv preprint arXiv:2101.11360* (2021).
[26] Hui Liu, Qingyu Yin, and William Yang Wang. 2018. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. *arXiv preprint arXiv:1811.00196* (2018).
[27] 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 (2020), 726–742.
[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[29] Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. *arXiv preprint arXiv:1911.04081* (2019).
[30] Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8433–8440.
[31] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*. 6294–6305.
[32] Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. *arXiv preprint arXiv:2109.13620* (2021).
[33] 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 on association for computational linguistics*. Association for Computational Linguistics, 311–318.
[34] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is Multilingual BERT? *arXiv preprint arXiv:1906.01502* (2019).
[35] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.
[37] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.
[38] Gerard Salton, Edward A Fox, and Harry Wu. 1982. *Extended Boolean information retrieval.* Technical Report. Cornell University.
[39] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. *arXiv preprint arXiv:1810.13327* (2018).
[40] Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. *arXiv preprint arXiv:1902.09492* (2019).
[41] Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Conversations Powered by Cross-Lingual Knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1442–1451.
[42] Cuk Tho, Arden S Setiawan, and Andry Chowanda. 2018. Forming of Dyadic Conversation Dataset for Bahasa Indonesia. *Procedia Computer Science* 135 (2018), 315–322.

[43] Jörg Tiedemann. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. Linköping University Electronic Press, 191–199.
[44] Jörg Tiedemann and Zeljko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* 55 (2016), 209–248.
[45] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, et al. 2018. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416* (2018).
[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[47] Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial singaporean english. *arXiv preprint arXiv:1705.06463* (2017).
[48] 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. *arXiv preprint arXiv:1905.08743* (2019).
[49] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627* (2016).
[50] Lu Xiang, Yang Zhao, Junnan Zhu, Yu Zhou, and Chengqing Zong. 2021. Zero-Shot Deployment for Cross-Lingual Dialogue System. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 193–205.
[51] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237* (2019).
[52] Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-Lingual Dependency Parsing Using Code-Mixed TreeBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 996–1005.