InFoBERT: Zero-Shot Approach to Natural Language Understanding Using Contextualized Word Embedding

Pavel Burnyshev Anfrey Bout Valentin Malykh Irina Piontkovskaya
Huawei Noah's Ark lab, Moscow Russia
lastname.firstname@huawei.com

Abstract

Natural language understanding is an important task in modern dialogue systems. It becomes more important with the rapid extension of the dialogue systems' functionality. In this work, we present an approach to zero-shot transfer learning for the tasks of intent classification and slot-filling based on pre-trained language models. We use deep contextualized models feeding them with utterances and natural language descriptions of user intents to get text embeddings. These embeddings then used by a small neural network to produce predictions for intent and slot probabilities. This architecture achieves new state-of-the-art results in two zero-shot scenarios. One is a single language new skill adaptation and another one is a cross-lingual adaptation.

1 Introduction

Dialogue systems become more and more popular in everyday life. A dialogue system has two main language understanding tasks which are of key importance for any skill: user intent recognition and information extraction from user input. The former task traditionally is interpreted as a classification problem, since the possible intents list assumed to be given. The latter task is formulated as slot filling, i.e. extraction of a text span considered to be a slot value. With dialogue systems becoming more popular their functionality is also growing in terms of both skills available to use and languages their functionality is accessible in. The rapid growth entails lack of usage data for the new skills and new languages and causes the lists of intents and slots to be dynamic thus make classical approaches inapplicable. We consider these newly emerged scenarios and demonstrate that our approach is well suited for both. In the first scenario, there are several known domains and a few unknown domains. The second scenario contains skills with training data for one language (English) and only test data for two other languages (Spanish & Thai).

In this paper, we present a new architecture which is based on pre-trained language models, like BERT (Devlin et al., 2019), and uses a small "head" (i.e. a neural network using the embeddings from the language models as input) evaluated in two scenarios: zero-shot single language natural language understanding on Schema-Guided Dialog dataset (Rastogi et al., 2019) and zero-shot cross-lingual natural language understanding on Multi-lingual Task-oriented Dialog dataset (Schuster et al., 2019). For the Schema-Guided Dialog dataset, there is a baseline described alongside the dataset which is evaluated in the zero-shot transfer learning setup. It is using deep contextualized embedding model (BERT) to embed the utterances and natural language descriptions for the slots and intents. Another approach was presented in (Ruan et al., 2020), where the authors present a system based on the baseline with an addition of handcrafted context features. Our model is also close to this baseline with some important differences (described in section devoted to our approach) allowing it to outperform both previously presented approaches. For the Multi-language Task-oriented Dialog dataset there are also published baselines in zero-shot cross-lingual transfer setup described alongside the dataset. The authors train a whole machine translation system to make embeddings of utterances and use them for prediction in both tasks. Another approach proposed in (Liu et al., 2019b), where the authors use combination of recurrent neural network and latent variable model. Also an approach was presented in (Liu et al., 2020). The authors use BERT model to embed utterances and use special context gate for prediction in both tasks. We also use a pre-trained language-model to embed natural language descriptions of slots and intents and to achieve state of the art results in the

Proceedings of Recent Advances in Natural Language Processing, pages 208–215
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_025

zero-shot scenario.

Formal contribution of this paper is four-fold: we present a new model architecture for (1) zero-shot intent classification and (2) zero-shot slot filling which with usage of appropriate embedding model achieves a new state of the art results on (3) cross-skill and cross-domain transfer and on (4) cross-language transfer.

2 Related Work

There are several baseline systems for MTD which we describe here, while baseline systems for SGD are described alongside our model in the specific section below. One baseline system (*Multi. CoVe*) presented in (Schuster et al., 2019) is based on deep contextualized embedding model CoVe (McCann et al., 2017). This system is using bidirectional LSTM (Hochreiter and Schmidhuber, 1997) over CoVe embeddings with CRF for slot prediction and an attention mechanism for intent recognition. This system has a variant with usage of an autoencoder for produced embedding refinement (*Multi. CoVe w/ auto*). (Liu et al., 2020) are using the same design as Multi. CoVe with other embedding models, namely multilingual BERT and XLM (Lample and Conneau, 2019), this approach is called *Transformer-MLT*. In this work is also presented adaptation of RCSLS model (Joulin et al., 2018) used as an embedding one, which aligns word embeddings with help of retrival loss (*RCSLS-MLT*). The authors of (Liu et al., 2019b) are using slightly different approach. Their system is using cross-lingually aligned word embeddings, they are fed into bidirectional LSTM, which output is used by attention mechanism for intent recognition and by a latent variable model for joint slot and intent prediction. In the work (Schuster et al., 2019) also presented another baseline called *Translate Train*. This system is using direct alignment of word tokens to fed this alignment into CRF for prediction.

There are models which use contextualized representations for natural language understanding in different setups. An approach to slot filling called Zero-Shot Adaptive Transfer described in (Lee and Jha, 2019). A model takes an utterance and natural language description of a slot and produces BIO-encoding of the utterance, i.e. points to the span, containing the value. To contextualize the word representations authors use bidirectional recurrent neural network. Another approach is presented in (Xu et al., 2020), where the authors use multilingual BERT as embedding model and predict jointly slots and intents with additional refinement task of data source prediction.

Deep contextualized embeddings, especially BERT, have been used for intent classification and slot filling tasks in other setups, e.g. (Chen et al., 2019; Chao and Lane, 2019; Zhang et al., 2019; He et al., 2020), where authors consider classic setup with shared slots and intents for train and test. There were also other works focused on zero-shot modeling (Bapna et al., 2017; Xia et al., 2018; Shah et al., 2019), domain adaptation and transfer learning techniques (Rastogi et al., 2017) in recent years. Deep learning based approaches have achieved state of the art performance on dialogue state tracking tasks. Popular approaches on small-scale datasets estimate the dialogue state as a distribution over all possible slot-values (Henderson et al., 2014; Wen et al., 2017) or individually score all slot-value combinations (Mrkšić et al., 2017; Zhong et al., 2018). Such approaches are not practical for deployment in virtual assistants operating over real-world services having a very large and dynamic set of possible values. Addressing these concerns, approaches utilizing a dynamic vocabulary of slot values have been proposed (Rastogi et al., 2018; Goel et al., 2019; Wu et al., 2019).

3 Task Description and Datasets

We consider two independent scenarios in this work, although they share an important feature: zero-shot transfer learning. The first scenario is a cross-skill and cross-domain transfer which is evaluated on Schema-Guided Dialog (SGD) dataset. It is formulated as follows: in each domain, there are one or more skills. Each skill has its intents and slots described. Also a skill could share an intent and/or a slot with another skill, but in general case, they are not sharing anything. To complicate things, in the dataset there is a domain (Alarm) which is presented only in the development set, but not in the train set.

The second scenario is a cross-lingual transfer and is evaluated on Multi-language Task-oriented Dialog dataset (MTD) dataset. Each skill presented in the dataset is described (to a reasonable extent) identically in all three languages. So a system could be trained on one language and evaluated on the other two languages.

Schema-Guided Dialog

Schema-Guided Dialog dataset is described in (Rastogi et al., 2019). This dataset contains task-oriented dialogues in different domains. Each skill has a so-called schema, which contains one-sentence descriptions of slots and intents used in this skill. Each dialogue has slots and intents marked up. It is important to mention that domain could be represented by more that one skill, e.g. a person could rent a car using two different services. The skill in one domain could be split into train and dev sets. Another important feature of this dataset is that it contains multi-domain conversations. The general statistics for this dataset is presented in Tab. 1. Additional statistics on the cross-domain dialogues distribution is presented in Tab. 2.

	Single-domain	Multi-domain	Combined
Dialogues	5,403/836	10,739/1,646	16,142/2,482
Utterances	82,588/11,928	247,376/36,978	329,964/48,726
Slots	201/134	214/132	214/136
Domains	14/16	16/15	16/16
Skills	24/17	26/16	26/17
Intents	35/28	37/26	37/28

Table 1: Summary statistics of the Schema-Guided Dialog dataset. Train/Dev values are separated with slash.

Domain	#Intents	#Dialogs	Domain	#Intents	#Dialogs
Alarm	2 (1)	37	Movie	4 (2)	1758
Bank	4 (2)	1021	Music	4 (2)	1486
Bus	4 (2)	2609	RentalCar	4 (2)	1966
Calendar	3 (1)	1602	Restaurant	4 (2)	2755
Event	5 (2)	3927	RideShare	2 (2)	1973
Flight	8 (3)	3138	Service	8 (4)	2090
Home	2 (1)	1027	Travel	1 (1)	2154
Hotel	8 (4)	3930	Weather	1 (1)	1308
Media	4 (2)	1292			

Table 2: The number of intents (services in parentheses) and dialogues for each domain in the train and dev sets. Multi-domain dialogues contribute to counts of each domain. The domain Service includes salons, dentists, doctors etc.

Multi-turn Task-oriented Dialog

Multi-turn Task-oriented Dialog dataset is described in (Schuster et al., 2019). This dataset consists of dialogues in three different languages, specifically English, Spanish, and Thai. The dialogues share semantics for intents and slots across the languages, which makes it possible to formulate a zero-shot cross-lingual task, i.e. a model could be trained on one language and evaluated on another language without any additional training. The semantics, in this case, is represented by one-sentence description for slots and intents. This dataset contains only one skill per domain and no multi-domain dialogues. The general statistics for this dataset is presented in Tab. 3. For this dataset, there is published train/validation/test split, which we follow in our experiments.

	English	Spanish	Thai
Utter-s	30,521/4,181/8,621	3,617/1,983/3,043	2,156/1,235/1,692
Slots	11	11	11
Dom-s	3	3	3
Intents	12	12	12

Table 3: Summary statistics of the Multi-turn Task-oriented Dialog dataset. Note that the slot type *datetime* is shared across all three domains and therefore the total number of slot types is only 11. Train/Dev/Test values are separated with slash.

4 Instruction Following BERT

We think of the natural language descriptions of intents and slots as instructions for a model to follow in order to achieve a result, which is either intent classification or slot filling. We base our approach on BERT-like models, especially their contextualization ability, and we show that these models could solve both parts with a small help of an additional simple module. We add to the pre-trained language-models a recurrence module for intent classification task and a feed-forward module for slot filling one. We call our approach Instruction Following BERT (*InFoBERT*).

4.1 BERT Model

To better present our work we start with brief description of the Bidirectional Encoder Representations from Transformers (BERT) model described in (Devlin et al., 2019). The modules in it are interconnected, so the model has access to the whole sequence at once. BERT model has a specific training procedure, which consists of two tasks: next sentence prediction and masked language modeling, that requires the model to use specific tokens, like `[SEP]` which separates different sentences in the text sequence, `[CLS]` which requires a model to make a decision (binary in the original setup), or `[MASK]` which hides a particular token from model input, so model is required use a context to perform a prediction of a masked token. The usage of these special tokens in training procedure is illustrated in Fig. 1. We have experimented with BERT and several derivative models, namely RoBERTa (Liu et al., 2019a) and ToD-BERT (Wu et al., 2019) for SGD dataset and

multilingual BERT, XLM (Lample and Conneau, 2019), XLM-RoBERTa (Conneau et al., 2020), and Language-Agnostic BERT Sentence Embedding (Feng et al., 2020) for MTD one.

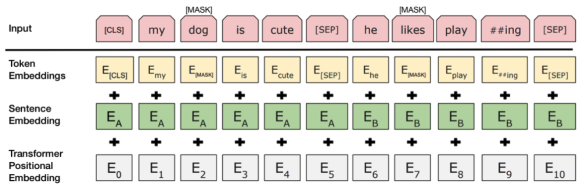

Figure 1: BERT training setup. The figure adopted from (Devlin et al., 2019).

4.2 Intent Classification

For each skill, there is a known list of possible intents, and each intent is accompanied with a natural language description. To represent rejection in the classification we add special intent "NONE", for which we use a separate description. This design is close to previous art for SGD. Although (Rastogi et al., 2019) system ignores a dialog history entirely, and (Ruan et al., 2020) system uses only hand-crafted history-based features. While our approach takes into account whole dialogue history using a recurrent neural network which is fed by a sequence of `[CLS]` token embeddings from all the utterances (including system ones). We denote `[CLS]` token embedding for i-th utterance as ${u_i}^{\texttt{CLS}}$, and ${\hat{u}_i}^{\texttt{CLS}}$ RNN output vector for this input. More formally: $\{{\hat{u}_i}^{\texttt{CLS}}\}_{i=1}^{n} = \text{RNN}\left(\{u_i^{\texttt{CLS}}, c_{i-1}\}_{i=1}^{n}\right)$, where c_i is a state of the recurrent module on i-th step, and n is a number of utterances in a dialogue.

We denote $I_j^{\texttt{CLS}}$ `[CLS]` token embedding from j-th intent description. To obtain the logits for possible intents, we compute dot-product between utterance vector representation and intent vector representation and normalize obtained scores using softmax (SM) function. More formally: $\text{P}_i^{intent} = \text{SM}(\{({\hat{u}_i}^{\texttt{CLS}})^\top \cdot I_j^{\texttt{CLS}} \mid j = \overline{0..m}\})$, where m is a number of the intents in a skill, and I_0 is reserved for "NONE" intent. The output logits are fed into conditional random field to incorporate further intent usage statistics.

To make the model more robust we use a Gaussian noise in form of d_{emb} normal distributions with zero mean and zero covariance adding it to all the embedding outputs of a language model. d_{emb} is an output language model embedding size.

4.3 Slot Filling

Generally, a slot has no list of possible values available, we call such slot the non-categorical one. In a case where the list of possible values is available, the slot is called categorical. MTD dataset contains only non-categorical slots, while SGD dataset contains both types. In this work we concentrate only on the non-categorical slots since they are more general and presented in both datasets we consider. The task is to extract a value from an utterance represented as a sequence of tokens, so the extracted value will be a span.

We denote s_k^{CLS} an embedding of [CLS] token for k-th slot description. Then to compute probability distributions for an utterance tokens $(u_{il} \mid l = \overline{1..L})$ to be a start $(P_i^{k,\texttt{start}})$ or an end $(P_i^{k,\texttt{end}})$ of a span we compute: $\text{P}_i^{k,\texttt{start}} = \text{SM}(\{u_{il}^\top \cdot W_{\texttt{start}} \cdot s_k^{\texttt{CLS}} \mid l = \overline{1..L}\})$ $P_i^{k,\texttt{end}} = \text{SM}(\{u_{il}^\top \cdot W_{\texttt{end}} \cdot s_k^{\texttt{CLS}} \mid l = \overline{1..L}\})$, where L is an utterance length, W_{start} and W_{end} are trainable matrices. It is important to mention, that these matrices are the only additional weights we use to perform the task.

To train the model for non-categorical slots extraction we use classic cross-entropy loss, summing it over all possible slots and utterances.

To make the model more robust we add slot-value dropout. This technique is related to one described in (Xu and Sarikaya, 2014), but in our work we are replacing the whole value span with `[MASK]` tokens thus requiring the model to rely on the value context.

Our setup is close to previous art for SGD. Our approach differs on the one hand from (Ruan et al., 2020) with usage of additional matrices for start and end token prediction, and on the other hand from (Rastogi et al., 2019) with usage of only one matrix for each prediction, while the baseline is using two matrices. Both these approaches do not use slot-value dropout.

5 Experiments

We conduct experiments on two datasets, described in the section devoted to the datasets. For both SGD and MTD datasets we measure quality for tasks of intent classification and non-categorical slots extraction in the zero-shot scenario. To measure the intent classification quality, we use *accuracy* metric, due to a model has a dynamic list of intents to choose from. To measure the slot extraction quality we use *F1 metric*, which allows us to measure both

Model	Single-dom.		Combined	
	Int.	Sl.	Int.	Sl.
InFoBERT-B	0.890	0.899	0.879	0.903
InFoBERT-R	0.954	0.959	0.940	0.925
InFoBERT-T	**0.982**	**0.965**	**0.955**	0.967
(Ruan et al., 2020)	N/A	N/A	0.948	**0.983**
(Rastogi et al., 2019)	0.748	0.883	0.773	0.891

Table 4: Schema-Guided Dialog dataset. **Int**ent classification accuracy and **Sl**ot tagging F1 measure.

the precision and recall of a slot detection.

In our experiments we use LSTM as a unit of the recurrent module in the intent classification task. We use one layer with hidden size equal to the output embedding size of a underlying language model throughout all the setups.

In our experiments on the SGD dataset we measure quality only for intents and slots not present in the original training set. To conduct the experiments on the SGD dataset we use BERT (Devlin et al., 2019) as an embedding model, specifically BERT-base variant for both tasks (*InFoBERT-B*); and RoBERTa (Liu et al., 2019a) model, specifically RoBERTa-large for intent classification and RoBERTa-base for slot filling (*InFoBERT-R*). We also use ToD-BERT (Wu et al., 2020) model denoting it *InFoBERT-T*. For the slot filling task the slot-value dropout probability of 0.35 was used. Since there were no explicitly presented metrics we measure in the original SGD paper, we reproduce their results using publicly available code[1]. The results are presented in Tab. 4. As one could see from the table InFoBERT-T outperforms both the baselines in the combined domain intent recognition task. Unfortunately, (Ruan et al., 2020) did not presented results for single domain task, but basing on our experiments we expect the results for their model would be close to presented ones.

In our experiments on the MTD dataset we use English data as training set and Spanish & Thai data as a validation and test sets. For the MTD dataset we use multilingual BERT (Devlin et al., 2019) as an embedding model (*InFoBERT-M*), XLM (Lample and Conneau, 2019) model (*InFoBERT-X*), XLM-RoBERTa (Conneau et al., 2020) model (*InFoBERT-XR*), and Language-Agnostic BERT Sentence Embedding (Feng et al., 2020) model (*InFoBERT-L*). The noise standard deviation was set to 0.1 for InFoBERT-M model and to 0.01 for InFoBERT-X one. For the slot filling task the slot-value dropout was used with two probabilities 0.2 and 0.35. The results are presented in Tab. 5. In the table there are results named "Translate Train". This is a special setup considered as a strong baseline. In this setup English data is translated to target language and used to train a model. These results and the results for Multi. CoVe [w/ auto] marked with * are adopted from (Schuster et al., 2019). The results of RCSLS-MLT and Transformer-MLT marked with † are adopted from (Liu et al., 2020).

As one could see form Tab. 5 InFoBERT-L variant of our model significantly outperforms all the baselines for Spanish language data, although for Thai language InFoBERT-XR is better than other baselines in intent prediction, with exception of Translate Train. For Thai slot tagging task InFoBERT-X variant is the best outperforming the Translate Train (and all the baselines).

Model	Spanish		Thai	
	Int.	Sl.	Int.	Sl.
InFoBERT-M	76.10	66.28	48.40	6.75
InFoBERT-X	90.32	77.48	65.65	**56.02**
InFoBERT-XR	88.51	74.84	**89.68**	17.50
InFoBERT-L	**96.65**	**86.74**	61.79	13.84
RCSLS-MLT†	87.05	59.12	81.44	30.42
Transformer-MLT†	87.88	74.88	73.46	28.47
(Liu et al., 2019b)	90.20	65.79	73.43	32.24
Multi. CoVe*	53.34	22.50	66.35	32.52
Multi. CoVe w/ auto*	53.89	19.25	70.70	35.62
Translate Train*	85.39	72.87	**95.85**	55.43

Table 5: Multi-language Task-oriented Dialog dataset. **Int**ent classification accuracy and **Sl**ot tagging F1 measure.

5.1 Thai Language Performance Analysis

The analysis of Tab. 5 shows that all the models with exception of InFoBERT-X variant of our model show significantly lower results on Thai slot filling task. The models we present and Transformer-MLT baseline could be split into two groups: the one using BERT tokenization model and the one using an external engine. We found out that all the models, including multilingual BERT, XLM-RoBERTa, Language-Agnostic BERT Sentence Embedding model, are using the same tokenization engine originally presented in BERT code[2]. Even XLM model by default is using this engine - this fact could explain low results for Transformer-MLT baseline. We found out that this

[1] https://github.com/google-research/google-research/tree/master/schema_guided_dst

[2] https://github.com/google-research/bert

default engine is broken against Thai text, thus it corrupts input text during tokenization. But XLM model could be set to use external tokenization[3] which handles the text correctly. This allows it to significantly improve the results on Thai language data.

It is interesting that the intent recognition results though being affected by this fault are still could achieve high performance. We suppose this fact is related to classification task structure where the whole utterance is regarded, allowing a model to overcome tokenization issues.

6 Conclusion

In this work we presented a model architecture which allows us to use different embedding models as a building block. This architecture is demonstrated to be effective in two zero-shot transfer tasks, namely cross-domain and cross-lingual. Our model using an appropriate embedding model (ToD-BERT for cross-domain task and several multi-lingual models for cross-lingual one) shows state of the art performance on the intent recognition and slot filling tasks.

It is interesting to mention that in our experiments we found that the best results for Spanish and Thai languages in the cross-lingual transfer task are not achievable at the same time. We suppose that this fact could be explained by the errors in the tokenization model used for most of the multi-lingual models. Thus usage of Language-Agnostic BERT Sentence Embedding model as an embedding one allows our system to outperform all the other systems on Spanish language data, but broken tokenization in this model does not allow to show any improvement for Thai language.

As future work authors consider the study of other BERT descendant models, which are plenty nowadays. Another direction of the work is research in low resource scenario, when some data is available for a model to tune onto. In closing, we hope that this work will facilitate the research in the direction of transfer learning in dialogue systems.

References

Ankur Bapna, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*.

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. *arXiv preprint arXiv:1907.00883*.

K. He, W. Xu, and Y. Yan. 2020. Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access*, 8:29407–29416.

M. Henderson, B. Thomson, and S. Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

[3] PyThaiNLP package. The results for tokenization benchmarking are available here: https://github.com/PyThaiNLP/pythainlp/blob/dev/tokenization-benchmark.md.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Sungjin Lee and Rahul Jha. 2019. Zero-shot adaptive transfer for conversational language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6642–6649.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019b. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303.

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. *Proceedings of 34th Association for the Advancement of Artificial Intelligence Conference*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1777–1788.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task learning for joint language understanding and dialogue state tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 376–384.

Abhinav Rastogi, Dilek Hakkani-Tür, and Larry Heck. 2017. Scalable multi-domain dialogue state tracking. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 561–568. IEEE.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*.

Yu-Ping Ruan, Zhen-Hua Ling, Jia-Chen Gu, and Quan Liu. 2020. Fine-tuning bert for schema-guided zero-shot dialogue state tracking. *Proceedings of DTSC8 Workshop @ AAAI 2020*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805.

Darsh Shah, Raghav Gupta, Amir Fayazi, and Dilek Hakkani-Tur. 2019. Robust zero-shot cross-domain slot filling with example values. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5484–5490, Florence, Italy. Association for Computational Linguistics.

TH Wen, D Vandyke, N Mrkšíc, M Gašíc, LM Rojas-Barahona, PH Su, S Ultes, and S Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference*, volume 1, pages 438–449.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogues. *arXiv preprint arXiv:2004.06871*.

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

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099. Association for Computational Linguistics.

Puyang Xu and Ruhi Sarikaya. 2014. Targeted feature dropout for robust slot filling in natural language understanding. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher, and Caiming Xiong. 2019. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv:1910.03544*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1458–1467, Melbourne, Australia. Association for Computational Linguistics.