Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5797–5804
Marseille, 20-25 June 2022

SHONGLAP: A Large Bengali Open-Domain Dialogue Corpus

Syed Mostofa Monsur[†], Sakib Chowdhury[†*], Md Shahrar Fatemi[†*], Shafayat Ahmed[‡]
[†]Celloscope Ltd.,[‡]Virginia Polytechnic Institute and State University
mostofa.monsur@cellosco.pe, 1606044@eee.buet.ac.bd, 1605007@ugrad.cse.buet.ac.bd
shafayatpiyal@vt.edu

Abstract

We introduce SHONGLAP, a large annotated open-domain dialogue corpus in Bengali language. Due to unavailability of high-quality dialogue datasets for low-resource languages like Bengali, existing neural open-domain dialogue systems suffer from data scarcity. We propose a framework to prepare large-scale open-domain dialogue datasets from publicly available multi-party discussion podcasts, talk-shows and label them based on weak-supervision techniques which is particularly suitable for low-resource settings. Using this framework, we prepared our corpus, the first reported Bengali open-domain dialogue corpus (7.7k+ fully annotated dialogues in total) which can serve as a strong baseline for future works. Experimental results show that our corpus improves performance of large language models (BanglaBERT) in case of downstream classification tasks during fine-tuning.

Keywords: open-domain dialogues, weak supervision, low-resource dialogue

1. Introduction

Due to recent advances in conversational systems research, a number of high quality dialogue datasets are being made publicly available (Rameshkumar and Bailey, 2020; Li et al., 2017b; McCowan et al., 2005; Janin et al., 2004; Canavan et al., 1997). These datasets have a wide range of applications including dialogue summarization, dialogue understanding, dialogue act classification, disagreement modeling and dialogue state tracking.

In general, dialogue systems can be classified into two major types: task-oriented and open-domain. Task-oriented dialogue systems are intended to assist people in achieving certain goals. Tasks are well-defined according to their use cases, and conversational systems are customized to these domains (Yan, 2018). Because of their attractive business value, these systems have been successfully deployed in a wide range of real-world applications such as online shopping (Yan et al., 2017), restaurant searching and reservation (Wen et al., 2017), customer care services at financial organizations and technical assistance (Huang et al., 2020a; Peng et al., 2018; Li et al., 2017a). Additionally, an increasing demand for task-oriented dialogue agents is observed in healthcare services such as automatic diagnosis,(Wei et al., 2018), mental state classification (Dosovitsky et al., 2020) and promoting health education (Liednikova et al., 2021; Brixey et al., 2017). Open-domain dialogue systems, on the other hand, are far more challenging to build due to their open-ended objective. Task-oriented dialogue systems have a predefined task-dependent workflow. In contrast with that, most of the currently available open-domain neural dialogue systems are not grounded in the real world which prevents these systems from seamlessly conversing about subjects that relate to the user's environmental context (Huang et al., 2020b). Researchers have recently begun to investigate methods to ground open-domain conversation systems in real-world events and knowledge (Ghazvininejad et al., 2018; Mostafazadeh et al., 2017; Qin et al., 2019).

Key challenges in building neural open-domain dialogue systems include semantic understanding, commonsense reasoning, contextual consistency, interactiveness, sentiment detection, domain and intent detection etc. Current state of the art dialogue systems are essentially powered by language models which require a massive amount of data for deep understanding of complex human language structures and generalizing them well (Devlin et al., 2019). For English and other languages, there are numerous structured, semi-structured, archived conversational data that can be collected in volume from publicly available news articles and radio transcriptions (NPR, CNN). A large number of these high-quality dialogue data are often made available by the owners of the archived resources (Zhu et al., 2021; Majumder et al., 2020). Unfortunately, in the case of low-resource languages like Bengali, high-quality open-domain transcribed dialogue data are not readily available in large volumes although there are a lot of publicly available Bengali podcasts, debates and discussion audio which can be found in the web. We propose a novel approach to prepare high-quality annotated dialogue data from multi-party discussion podcasts using weak-supervision techniques. To the best of our knowledge, this is the first fully labeled large-scale dataset on Bengali open-domain dialogues.

The key contributions of this work can be summarized as follows:

- We propose a new framework to collect dialogue data for low-resource languages in volume by leveraging weak supervision.
- We have prepared an annotated Bengali dialogue

[*]These authors contributed equally to this work.

corpus with around 7.7k+ transcripts collected from publicly available talk shows and podcasts.

- We evaluate our corpus on speaker bias detection task and show that *SHONGLAP*[*] enhances classification performance during fine-tuning of BanglaBERT.

2. Related Works

A large number of high-quality open-domain dialogue corpus is available for English and other languages. (Zhang et al., 2018; Li et al., 2017b; Zhou et al., 2018) are some examples of the crowd-sourced open-domain dialogue corpus in English. These dialogue datasets contain knowledge-grounded features and cover a wide range of domains. (Rashkin et al., 2019) presented an empathetic dialogue dataset used in emotional dialogue modeling. OpenDialKG (Moon et al., 2019) is another crowd-sourced open-domain dialogue dataset which covered a wide range of topics including movies, books, sports and music. (Tang et al., 2019) introduced Target-Guided-Conversation, a crowd-sourced open-domain dialogue dataset enriched with features for proactivity, behavioral and strategy learning. They proposed a structured approach that introduces coarse-grained keywords to control the intended content of system responses and then attained smooth conversation transition through turn-level supervised learning.
Application of weak-supervised strategies have recently caught attention in dialogue research. (Hudeček et al., 2021) used weak-supervision to annotate slots which resulted in significant improvement in the performance of an end-to-end dialogue response generation model. (Badene et al., 2019) uses weak supervision to learn discourse structures outperforming the combination of deep learning methods and handcrafted features.

3. Our Approach

3.1. Overview

We use publicly available podcasts and news recordings from various sources. The audio files mostly contain political discussions and debates on related topics. The files are converted into 16kHz single channel WAV audio files. First, we pass the audio files through a background noise removal layer. Then we perform speaker diarization on the clean audio files to determine which speaker spoke exactly when in the actual discussion audio. We then segment the diarized audio files based on speaker turns. We employ an end-to-end pretrained speech to text architecture to perform automatic transcription of the discussion audio files. After restoring the punctuations (as part of post-processing) of the transcriptions, we employ weak-supervision to annotate the dialogue corpus. Figure 1 depicts the workflow of the dialogue corpus preparation process. Details of the stages are described in subsections 2.2 and 2.3.

[*]SHONGLAP means "Conversation" in Bengali

3.2. Preparing Corpus from Raw Audio

3.2.1. Collecting and Cleaning Audio

We collect political discussion and debate audio from various publicly available podcasts and TV recordings (Matra, 2022). We convert the raw audio files to 16kHz mono channel WAV audio. The duration of the audio files is 54 minutes on average. We follow (Kashyap et al., 2021) to eliminate background noise signals from the collected audio. The background noise removal module is a 20 layered Deep Complex U-Net based architecture which achieves superior denoising performance over conventional training regimes utilizing clean training audio targets, in cases involving complex noise distributions and low Signal-to-Noise ratios (high noise environments).

3.2.2. Speaker Diarization

Speaker Diarization is the task of segmenting a multi-speaker audio stream based on the speakers i.e. answering the question of *"who spoke when?"*. To apply speaker diarization on our collected multi-party discussion audio, we follow (Dawalatabad et al., 2021) which is an enhanced version of the standard TDNN model based on Emphasized Channel Attention, Propagation, and Aggregation (ECAPA-TDNN) (Desplanques et al., 2020). The ECAPA-TDNN model makes use of a channel and context-dependent attention mechanism, as well as Multilayer Feature Aggregation (MFA), Squeeze-Excitation (SE), and residual blocks. It has also demonstrated superior performance in the domain of speaker verification (Desplanques et al., 2020). The ECAPA-TDNN model is being used in the context of speaker diarization by improving its robustness even further by training the ECAPA-TDNN model with an extensive on-the-fly augmentation scheme that relies on the combination of various techniques for speech contamination and Spectral Clustering (SC) method which shows highly competitive performance compared to the traditional Agglomerative Hierarchical Clustering (AHC) (Dawalatabad et al., 2021).

3.2.3. Splitting and Cleaning

We use the *.rttm* files generated by the diarization module to segment the raw audio files based on speakers. For long audio clips, we further split them (based on silence segments) into smaller clips using PyAudioAnalysis (Giannakopoulos, 2015). We take 0.4 seconds as minimum silence length and 0.0001 as silence threshold for clip generation.

3.2.4. Speech to Text

For low-resource languages like Bengali, it is difficult to collect reasonably large amounts of transcribed audio. Nonetheless, there are some publicly available datasets like (Kjartansson et al., 2018) which has around 229 hours of transcribed Bengali audio in total. Additionally, we collect another 150+ hours of raw Bengali audio from publicly available sources and transcribe them using Google Speech API. We

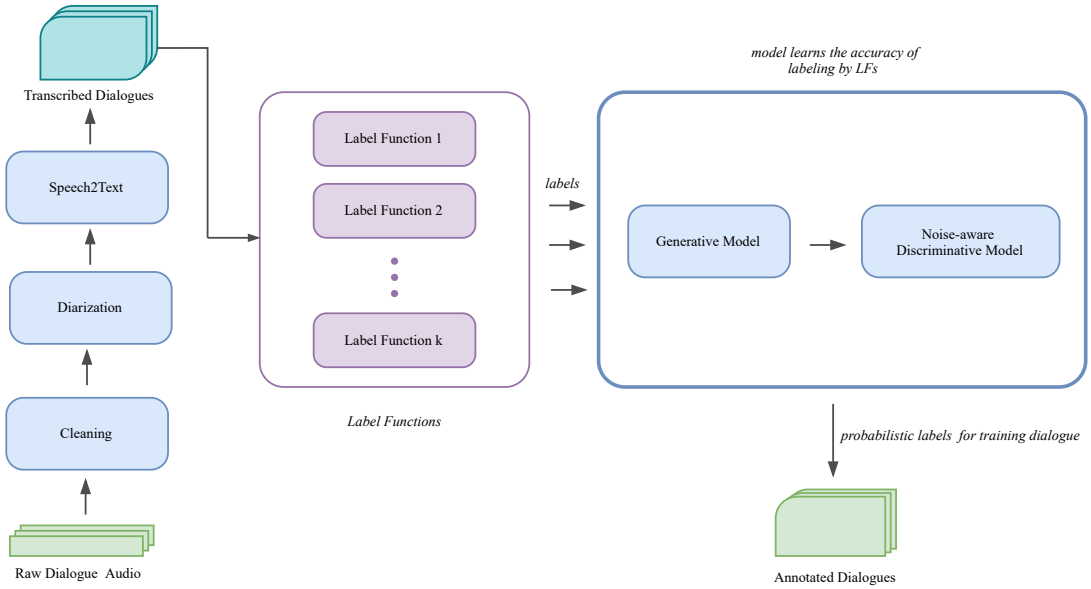

Figure 1: Workflow of preparing annotated open-domain dialogue data using weak supervision.

Topics	Keywords
Election	ইলেকশনের (Election), নির্বাচন (Election), গণতন্ত্রের (Democracy), কমিটি (Committee), অংশগ্রহণ (Participation), প্রতিরোধ (Prevention), সরকার (Government), নিরপেক্ষ (Neutral), সাংসদ (MP), প্রেসিডেন্ট (President)
International	আবুধাবি (Abu Dhabi), বাংলাদেশ (Bangladesh), ইন্ডাস্ট্রি (Industry), ইন্টারন্যাশনাল (International), জাতিসংঘে (UN), ভারতের (India's), কলকাতা (Kolkata), পাকিস্তান (Pakistan), পলিসিজ্‌গুলো (Policies), রাষ্ট্রীয় (State)
Bangladesh	জনগণ (Public), টিউবস (Tubes), বাংলা (Bangla), সংস্কৃতি (Culture), মন্ত্রী (Minister), বাংলাদেশকে (Bangladesh), ২০২৪ (2024), ২০২০ (2020), জাতীয় (National), উন্নয়নের (Development)
Economics	ইঞ্জিনিয়ারিং (Engineering), ইউনিয়ন (Union), ডক্টর (Doctor), অর্থমন্ত্রী (Minister of Finance), কক্সবাজার (Cox's Bazar), দক্ষিণের (South), অংখের (Amount), নারীর (Women's), গুগোল (Google), কলেজ্ (College)
Miscellaneous	আমার (My), কথা (Talk), থেকে (From), আপনি (You), করতে (Do), যদি (If), আপনার (Your), হচ্ছে (Happening), এটা (This), কোন (What), তারা (They), তো (So)

Figure 2: Most relevant keywords based on topics generated by Latent Dirichlet Allocation.

use wav2vec2.0 (Baevski et al., 2020) which is a framework for learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech. Their method is particularly useful for low-resource settings as they leverage pre-trained models which are trained on large scale audio data from multiple languages. We use XLSR-53 (Conneau et al., 2021) as the base model which is pre-trained on multiple language datasets including MLS: Multilingual LibriSpeech (Pratap et al., 2020), CommonVoice (Mozilla, 2022), Babel (UPenn, 2022), covering 53 languages in total (Ott et al., 2019). The base model is then fine-tuned on 400 hours of transcribed Bengali audio data. We use a lightweight language model KenLM (Heafield, 2011) to further improve the speech to text model output.

3.2.5. Punctuation Restoration

As part of post-processing of the generated dialogue transcriptions, we use a punctuation restoration module to improve the transcription quality. We employ (Alam et al., 2020) as our punctuation restoration module. They propose a layered architecture consisting of a pre-trained BERT (Devlin et al., 2019) variant, a bi-directional LSTM and finally a linear layer on top of it. The layered architecture is then fine-tuned on pre-processed punctuation labeled dataset extracted from the target language. They show that their architecture achieved comparable state-of-the-art results for English but for Bengali, it was the first reported work. Following their architecture, we fine-tuned the base model of XLM-RoBERTa (Conneau et al., 2020) using pre-processed punctuation labeled Bengali corpus which achieved a test accuracy of 97.2%.

3.3. Dialogue Dataset Preparation

3.3.1. Labeling Speaker Roles

Each dialogue in the corpus consists of multiple speakers (2.68 on average). Most of the dialogues include a speaker who acts as "Host". The host conducts a discussion session focusing on a particular topic where one or more guest speakers take part in the discussion

Statistics	Count
Total number of dialogues	7703
Total number of sentences	66413
Total number of unique tokens	138445
Avg. duration (in minutes)	5.62
Avg. speakers per dialogue	2.68
Avg. turns per dialogue	7.6
Avg. number of questions per dialogue	2.14

Table 1: Corpus statistics of the dataset.

Sentiment	Percentage
Positive	36.20%
Negative	24.36%
Neutral	39.44%

Table 2: Sentiment analysis results on the dialogue corpus.

in turns. In the speaker diarization module, we segment the audio portions w.r.t the speakers but we need to label the roles of the participants properly. For this task, we leverage weak-supervision techniques incorporating multiple label functions which are then applied to unlabeled (without the *speaker role* label) dialogue corpus. In our observation, we found that in most dialogues, the host generally tends to ask more questions than guests. Also, the host makes remarks standing on a neutral ground and the guest speakers generally answer long descriptive questions. During the design of label functions, we considered facts like number of questions asked, utterance sentiment, descriptiveness of the replies etc. With the help of a generative model, accuracy of the labeling functions are learnt on the fly and weights are assigned to the corresponding outputs. The generative model generates a collection of probabilistic training labels that are used to train a strong, flexible discriminative model which generalizes beyond the signal expressed in our labeling functions (Ratner et al., 2019). Thus we can seamlessly provide high-quality speaker role labels to the dialogue samples which would take a very long time to label with manual supervision. We employ *Snorkel* (SnorkelTeam, 2022), a popular weak-supervision toolkit, for annotating the dialogues with speaker roles.

4. Corpus Description

In this section, we will use statistics to explore several characteristics of our prepared corpus.

We organize our dialogue corpus following the works of (Majumder et al., 2020) and (Zhu et al., 2021). Our prepared dialogue corpus contains 7703 dialogues in total covering mostly debates and multi-party discussions in Bengali language. In each dialogue, the speakers take turns while talking. A sample dialogue from our corpus is shown in Figure 3. Each dialogue has an average duration of 5.6 minutes. The dialogues have on average 7.6 turns, 2.68 speakers and 17.97 unique tokens. The corpus statistics are summarized in Table 1.

We analyze the sentiment of the utterances in our prepared dialogue corpus. As part of sentiment detection, we classified the 7703 dialogue files into 3 classes - positive, negative, neutral. We follow a pre-trained BERT based model for detecting sentiment in our dialogue utterances from here (SBNLTK, 2022). Their pre-trained model (trained on 20k+ samples) reports 93.2% accuracy. We feed the utterances through the model and consider the predicted highest scoring class to be the label of that particular text. Table 2 shows the sentiment statistics of our dialogue corpus. In the case of dialogues, we have split them into sentences which makes a total of 66413 sentences. Among them, 26194 are detected to be 'neutral'. The number of 'positive' and 'negative' sentences were found 24041 and 16178 respectively.


```
{
 "utts":[
    "নির্বাস্তু হচ্ছে",
    "সেটা আমি ... হচ্ছে ভোটের পাশে",
    "ভোটের পাশে যে মানুষের নির্বাচন ... করে না",
    "২০ তারিখের নিউইয়র্কের উপনির্বাচনের ... পেছনে কত টাকা",
    "এখনো ... বড় টাকা",
    "এখন এখন এই যে ... এটা কথা আমি প্রশ্ন করি",
    "এই জায়গাটা",
    "এই দেশের ট্যাক্স ... আমার স্ত্রী বা আমার",
    "ইনভেস্ট করেন আবার ইনভেস্ট ... ইনভেস্টমেন্ট হবে"
  ],
  "speakers":[
    "HOST",
    "SPEAKER_1",
    "HOST",
    "SPEAKER_1",
    "SPEAKER_0",
    "SPEAKER_1",
    "SPEAKER_0",
    "SPEAKER_1",
    "SPEAKER_0"
  ]
}
```


Figure 3: Sample dialogue from the prepared dialogue corpus. The lists *utts* and *speakers* have equal length.

Also, we analyze our prepared dialogue corpus using topic modeling techniques to identify key topics of discussion and the relevant words. We use Latent Dirichlet Allocation (LDA) which is a widely used algorithm in topic modeling (Blei et al., 2003). The key idea behind LDA is that a document is a cluster of topics and each topic is a combination of words or tokens. The LDA model assigns each word in a document to the best-suited topic. The Dirichlet model identifies the patterns in a group of words that are repeating together. These words are similar to each other and occur frequently.

After tokenizing the corpus, a TF-IDF matrix is calculated on the entire corpus. In vector space, the text utterances of the dialogues are represented as a document-term matrix (DTM). To perform LDA, we

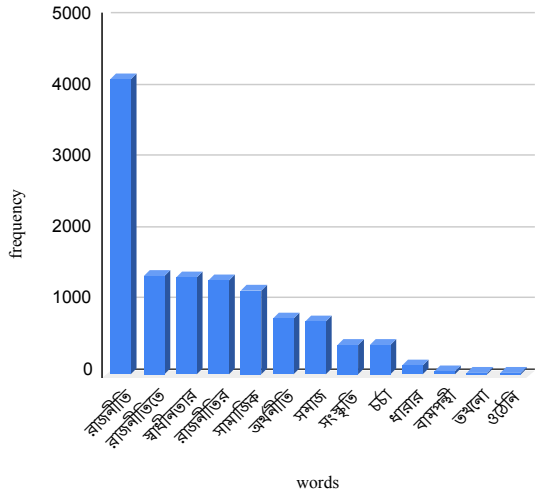

Figure 4: Top political keywords and their frequencies. English translations from left: *Politics, In politics, For Liberation, For politics, Social, Economics, Society, Culture, Practice, Rules, Leftist, Then, Haven't (done) yet.*

use popular machine learning toolbox Scikit-Learn (Pedregosa et al., 2011). It comes with *LatentDirichletAllocation* model which performs LDA on DTM matrix. Since LDA is a clustering method, we cluster all the sub-topics into a group of 5 major clusters (i.e. topics). The LDA model provides a list of words under each topic which is open to interpretation by an observer. The words having the highest frequency of occurrence under a particular cluster is called a 'Topic'. We then manually identify the topic name after analyzing the topic keywords contents. Figure 2 shows the topic-wise relevant top keywords that were found by applying LDA on the dialogue corpus. We observe *Elections, International Affairs, Bangladesh, Economics* were the most highlighted topics in our dialogue corpus which is consistent with the fact that most of the dialogues were collected from publicly available multi-party political discussions and debate podcasts in Bengali language.

Additionally, we use GloVe (Pennington et al., 2014) embeddings to visualize the most frequently mentioned political words in our dialogue corpus. The embeddings were pre-trained on a Bengali corpus collected from various publicly available sources with around 39M tokens and a vocabulary size of around 0.18M. The dimensions of trained embeddings were 300. Since a large number of dialogues were based on political debates and multi-party discussions, we take the top words which have the highest cosine proximity with political words in Bengali language. The top political keywords and their corresponding frequencies are shown in Figure 4.

5. Evaluation

In this section, we present the results of experimental evaluation on our dialogue corpus. Our prepared dialogue corpus contains speaker role sequences (host and guest) in the order of their appearance in the audio. As most of the dialogues are collected from debates, talk-shows and podcasts, a host and one or more participants are present in each of the dialogues. The host generally conducts the discussion or debate session from a neutral standpoint whereas the other participants present mutually contradictory views and opinions in the discussion.

Our task is to classify biased speakers from the utterances. The key assumption here is that the host generally remains unbiased and the other participants are somehow biased i.e. inclined towards a particular view. So we generate our dataset with utterance samples as data and the labels based on the speaker bias (biased vs unbiased). We assign negative labels to utterances by hosts and positive labels to those spoken by guest speakers.

We run experiments on our dataset using BanglaBERT (Bhattacharjee et al., 2022) as our base pre-trained model. BanglaBERT is a NLU model for Bangla which is based on Transformers (Vaswani et al., 2017). The base model is based on the ELECTRA (Clark et al., 2020) model and pre-trained on 18.6 GB Bangla text data. It outperformed previous state-of-the-art results on five downstream tasks by up to 3.5% (Bhattacharjee et al., 2022).

We fine-tuned BanglaBERT on our downstream task - biased speaker classification. We customized the fine-tuning recipe which is provided by the owners of BanglaBERT for this binary classification task. We generated a training set with 10000 samples and assigned each of them a positive (biased) or negative (unbiased) label based on speaker type. We also keep 500 samples as test set and another 500 samples as a validation set. We use the base BanglaBERT model which is based on ELECTRA. We set our learning rate to 0.00002, gradient accumulation steps to 2, weight decay value to 0.1 and maximum sequence length to 512. Due to memory constraints, we set both per device training batch size and per device evaluation batch sizes to 1. All the experiments were run on a Corei9-9900K CPU @ 3.60GHz machine with 32GB memory and Nvidia RTX 3070 GPU.

After 3 epochs, the fine-tuned model achieves a test accuracy of 0.735 which is shown in Table 3. We also conduct experiments analyzing the effect of training set size on F1 score of the test set. From figure 5, we observe that the F1 score value increases with the increasing size of training data set. Which means that adding samples from our prepared dialogue corpus improves performance of BanglaBERT during fine-tuning for the classification task. This proves the effectiveness of our dialogue corpus in case of downstream tasks.

Metric	Score
Accuracy	0.735
Precision	0.7328
Recall	0.735
F1 Score	0.733

Table 3: Accuracy, Precision, Recall and F1 scores on test set for biased speaker classification task (using fine-tuned BanglaBERT model).

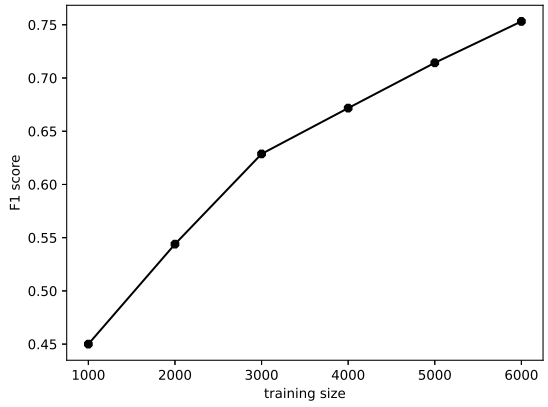

Figure 5: Effect of training data size on test F1 score

6. Conclusion

In this work, we presented a novel framework to prepare and annotate large-scale open-domain dialogue corpus (from publicly available discussions, podcasts, talk-show audio) which is suitable for low-resource settings. Using the framework, we prepared a large Bengali open-domain dialogue corpus: *SHONGLAP*. We also performed various analyses on our corpus and showed its competency in improving BanglaBERT's performance during fine-tuning for downstream tasks. In future, we will further extend our approach to collect and prepare an even larger dialogue corpus covering a wide range of topics. In addition, tasks like dialogue-summarization, agreement-disagreement modeling, dialogue state tracking, open-domain dialogue generation, mining similar dialogues in the context of Bengali will be explored in depth.

7. Acknowledgements

We thank DOER Services Ltd. for funding and supporting this research.

8. Bibliographical References

Alam, T., Khan, A., and Alam, F. (2020). Punctuation restoration using transformer models for high- and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 132–142, Online, November. Association for Computational Linguistics.

Badene, S., Thompson, K., Lorré, J.-P., and Asher, N. (2019). Weak supervision for learning discourse structure. 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 2296–2305, Hong Kong, China, November. Association for Computational Linguistics.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Bhattacharjee, A., Hasan, T., Mubasshir, K., Islam, M. S., Uddin, W. A., Iqbal, A., Rahman, M. S., and Shahriyar, R. (2022). Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar.

Brixey, J., Hoegen, R., Lan, W., Rusow, J., Singla, K., Yin, X., Artstein, R., and Leuski, A. (2017). SHIHbot: A Facebook chatbot for sexual health information on HIV/AIDS. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 370–373, Saarbrücken, Germany, August. Association for Computational Linguistics.

Canavan, Alexandra, Graff, D., and Zipperlen, G. (1997). CALLHOME American English Speech LDC97S42. Web Download. *Philadelphia: Linguistic Data Consortium*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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, July. Association for Computational Linguistics.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., and Na, H. (2021). Ecapa-tdnn embeddings for speaker diarization. *Interspeech 2021*, Aug.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention,

propagation and aggregation in tdnn based speaker verification. *Interspeech 2020*, Oct.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., Escoredo, M., and Bunge, E. L. (2020). Artificial intelligence chatbot for depression: Descriptive study of usage. *JMIR Form Res*, 4(11):e17065, Nov.

Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., and Galley, M. (2018). A knowledge-grounded neural conversation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12).

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Huang, M., Zhu, X., and Gao, J. (2020a). Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3), apr.

Huang, M., Zhu, X., and Gao, J. (2020b). Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3):21:1–21:32.

Hudeček, V., Dušek, O., and Yu, Z. (2021). Discovering dialogue slots with weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2430–2442, Online, August. Association for Computational Linguistics.

Janin, A., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2004). ICSI Meeting Speech LDC2004S02. Web Download. *Philadelphia: Linguistic Data Consortium*.

Kashyap, M. M., Tambwekar, A., Manohara, K., and Natarajan, S. (2021). Speech Denoising Without Clean Training Data: A Noise2Noise Approach. In *Proc. Interspeech 2021*, pages 2716–2720.

Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M., and Ha, L. (2018). Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 52–55, Gurugram, India, August.

Li, X., Chen, Y.-N., Li, L., Gao, J., and Celikyilmaz, A. (2017a). End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017b). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Liednikova, A., Jolivet, P., Durand-Salmon, A., and Gardent, C. (2021). Gathering information and engaging the user ComBot: A task-based, serendipitous dialog model for patient-doctor interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, Online, June. Association for Computational Linguistics.

Majumder, B. P., Li, S., Ni, J., and McAuley, J. (2020). Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online, November. Association for Computational Linguistics.

Matra, T. (2022). Tritiyo matra.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. In L.P.J.J. Noldus, et al., editors, *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140. Noldus Information Technology, August.

Moon, S., Shah, P., Kumar, A., and Subba, R. (2019). OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy, July. Association for Computational Linguistics.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G., and Vanderwende, L. (2017). Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Mozilla, C. V. (2022). Common voice.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S.,

Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, B., Li, X., Gao, J., Liu, J., Chen, Y.-N., and Wong, K.-F. (2018). Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6149–6153.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. *Interspeech 2020*, Oct.

Qin, L., Galley, M., Brockett, C., Liu, X., Gao, X., Dolan, B., Choi, Y., and Gao, J. (2019). Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy, jul. Association for Computational Linguistics.

Rameshkumar, R. and Bailey, P. (2020). Storytelling with dialogue: A critical role dungeons and dragons dataset. Association for Computational Linguistics.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July. Association for Computational Linguistics.

Ratner, A., Bach, S., Varma, P., and Ré, C. (2019). Weak supervision: the new programming paradigm for machine learning. *Hazy Research. Available via https://dawn. cs. stanford. edu//2017/07/16/weak-supervision/. Accessed*, pages 05–09.

SBNLTK. (2022). Sbnltk.

SnorkelTeam. (2022). Snorkel.

Tang, J., Zhao, T., Xiong, C., Liang, X., Xing, E., and Hu, Z. (2019). Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy, July. Association for Computational Linguistics.

UPenn, B. (2022). Babel.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei, Z., Liu, Q., Peng, B., Tou, H., Chen, T., Huang, X., Wong, K.-f., and Dai, X. (2018). Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia, July. Association for Computational Linguistics.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.

Yan, Z., Duan, N., Chen, P., Zhou, M., Zhou, J., and Li, Z. (2017). Building task-oriented dialogue systems for online shopping. In Satinder P. Singh et al., editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4618–4626. AAAI Press.

Yan, R. (2018). "chitty-chitty-chat bot": Deep learning for conversational ai. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5520–5526. International Joint Conferences on Artificial Intelligence Organization, 7.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July. Association for Computational Linguistics.

Zhou, K., Prabhumoye, S., and Black, A. W. (2018). A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zhu, C., Liu, Y., Mei, J., and Zeng, M. (2021). MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online, June. Association for Computational Linguistics.