Low-Resource Chat Translation: A Benchmark for Hindi–English Language Pair

Baban Gain[1] gainbaban@gmail.com
Ramakrishna Appicharla[1] ramakrishnaappicharla@gmail.com
Soumya Chennabasavraj[2] soumya.cb@flipkart.com
Nikesh Garera[2] nikesh.garera@flipkart.com
Asif Ekbal[1] asif@iitp.ac.in
Muthusamy Chelliah[2] muthusamy.c@flipkart.com
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]Flipkart, India

Abstract

Chatbots or conversational systems are used in various sectors such as banking, healthcare, e-commerce, customer support, etc. These chatbots are mainly available for resource-rich languages like English, often limiting their widespread usage to multilingual users. Therefore, making these services or agents available in non-English languages has become essential for their broader applicability. Machine Translation (MT) could be an effective way to develop multilingual chatbots. Further, to help users be confident about a product, feedback and recommendation from the end-user community are essential. However, these question-answers (QnA) can be in a different language than the users. The use of MT systems can reduce these issues to a large extent. In this paper, we provide a benchmark setup for Chat and QnA translation for English-Hindi, a relatively low-resource language pair. We first create the English-Hindi parallel corpus comprising of synthetic and gold standard parallel sentences. Thereafter, we develop several sentence-level and context-level neural machine translation (NMT) models, and measure their effectiveness on the newly created datasets. We achieve a BLEU score of 58.7 and 62.6 on the English-Hindi and Hindi-English subset of the gold-standard version of the WMT20 Chat dataset. Further, we achieve BLEU scores of 52.9 and 76.9 on the gold-standard Multi-modal Dialogue Dataset (MMD) English-Hindi and Hindi-English datasets. For QnA, we achieve a BLEU score of 49.9. Further, we achieve BLEU scores of 50.3 and 50.4 on question and answers subsets, respectively. We also perform thorough qualitative analysis of the outputs by the real users.

1 Introduction

Chatbots or conversational systems serve a crucial role in various sectors such as banking, healthcare, e-commerce, customer support, etc. While chatbots are convenient and fast, most are available only for resource-rich languages like English, limiting their widespread usage to users from languages other than the chatbot's language. It is essential to make these services available in non-English languages for their broader

applicability. Machine Translation (MT) can be an effective technology for developing multilingual chatbots to meet the need of multilingual societies. In recent years, there has been significant progress in Neural Machine Translation (NMT) with applications in a variety of domains, such as document (Yu et al., 2020), biomedical Yeganova et al. (2021), news (Hassan et al., 2018; Ng et al., 2019) etc. However, NMT requires a huge amount of data (Koehn and Knowles, 2017), and its manual creation requires significantly a lot of human effort. Chat translation poses more challenges than sentence-level translation due to the following: (i). Translation of the current utterance may depend on the contextual sentences. To translate the current sentence properly, we may need to refer to the previous sentences in the chat to grasp the meaning and generate appropriate translation; (ii). Chat often contains informal sentences, urban slang, stretched words (e.g., niceeee), code-mixed phrases, etc.

Further, communicating with chatbots or customer service may not be enough to make an informed decision regarding a product. Queries regarding a product can effectively be answered by the community of customers who use a similar product, leading to the requirement of a QnA system. Hence, it is also vital to perform QnA translation. This paper provides a benchmark setup for Chat and QnA translation for Hindi–English language pair, which has very little to no resources for such tasks. Firstly, we create a synthetic and gold-standard parallel corpus for English-Hindi chat and QnA translation. We use the existing MT systems described in section 5 to generate synthetic translation. With this model, we get a BLEU score of 47.8, indicating the translation is of good quality. Further, we create gold-standard parallel corpus by professional translators.

We employ the transfer-learning technique by initializing the transformer architecture using the model trained on similar domain data (Zoph et al., 2016). Further, we report the results on using same-speaker context to generate context-aware translation, which was useful for better translations for English–German (Gain et al., 2021). For QnA, we combine the QnA pair to generate consistent translation. We used to question and answer tags during training to help the model learn QnA-specific properties. Further, we compare the results with models trained on only question corpus or only answer corpus. We report BLEU (Papineni et al., 2002; Post, 2018) and TER (Snover et al., 2005) scores. To our knowledge, there is no publicly available dialogue dataset for the English–Hindi Chat or QnA translation. We introduce two datasets containing a total of 68.7K synthetic sentences, as well as 3,037 sentences of gold-standard data for chat translation. Further, we provide about 2.1 million QnA pairs (4.2 million sentences), their corresponding translation (synthetic), and 1,000 gold-standard sentences for each validation and test set, respectively. We make all the data and codes publicly available[1]. To summarize, our work has the following attributes: (i). introduce three English-Hindi parallel corpora for Chat Translation in Service and E-Commerce domains; (ii). use several context-aware and context-agnostic methods and observe their effectiveness on Chat Translation in (extremely) low-resource language settings, especially for the task at hand; (iii). propose a method to generate consistent translation performance in question-answer settings, where question-answers can be treated as a short chat with only two utterances; (iv). MT outputs have been quality checked by the actual users recruited by the well-known e-commerce company.

[1] https://github.com/babangain/en_hi_chat_qna_translation

2 Related Work

There are several approaches to applying Chatbots in E-commerce. Zhang et al. (2018) proposed a system that asks aspect-based questions to the user in a sequence, and recommendations are provided when the system is confident. Sun and Zhang (2018) proposed a personalized recommendation model which considers past ratings of products rated by the user and queries in the current conversation session to generate product recommendations. Chen et al. (2019) integrated the recommender system and dialog system, where the dialog system enhanced the recommendation system by introducing knowledge-grounded information about users' preferences. Further, the recommender system improved the dialog generation system by providing recommendation-aware vocabulary bias. Lai et al. (2018) showed that a simple transfer learning technique from an existing large-scale community question answering helped to generate 10% more accurate answers. Qu et al. (2019) proposed positional history answer embedding method to encode conversation history with position information. Further, they proposed a method that attends to conversational utterances with different weights based on their helpfulness in answering the current question. Deng et al. (2020) generated opinion-aware answers by jointly learning answer generation and opinion mining with a unified model. However, it is to be noted that all of the systems are monolingual.

Despite its demand, the field of dialogue translation remains mostly unexplored due to the lack of publicly available chat corpus. Farajian et al. (2020) introduced an English-German parallel conversational corpus. Berard et al. (2020) adopted several methods including replacement of rare characters with a special '<copy>' token, inline casing, tagged back-translation (BT) (Caswell et al., 2019), Byte-Pair-Encoding (BPE) (Sennrich et al., 2016), dropout (Provilkov et al., 2020), tagged synthetic noise, and ensemble of models using domain-specific adaptive layers, etc. While the largest single contributor to the improvement of translation quality was fine-tuning, the ensemble method with a domain-specific adaptor layer generated the best translation on WMT20 Chat data. Moghe et al. (2020) used the pre-trained models (Ng et al., 2019) and fine-tuned them on the pseudo-in-domain and in-domain data. Wang et al. (2020) adapted Cross-lingual Language Model Pre-training (Conneau and Lample, 2019) objectives into document-level NMT by using three previous contexts along with the current sentence. Bao et al. (2020) used the transformer architecture, modified with an additional encoder to process one previous context. Additional encoder failed to generate better overall translation in terms of BLEU score.

Gain et al. (2021) proposed a rule-based context selection technique where previous sentences by the same user are used to enhance the translation quality. Liang et al. (2021a) introduced an English-Chinese dialogue dataset named BMELD, which is an automatically translated and manually post-edited version of the MELD dataset (Poria et al., 2019). Further, they introduced a conditional variational auto-encoder (CVAE) model that captures role preference, dialogue coherence, and translation consistency. Liang et al. (2021b) proposed a multi-tasking system performing monolingual response generation, cross-lingual response generation, subsequent utterance discrimination, and speaker identification along with NMT. Here, the context-aware multi-tasking methods could generate better translation than context-agnostic models. Liang et al. (2022b) extended the same by introducing an additional objective, cross-lingual subsequent utterance discrimination, and evaluated the models with BMELD (English–Chinese) and BConTrast (English–German) datasets. Wang et al. (2021) proposed a multi-task learning-based NMT system to identify missing pronouns and typos and utilize context

to translate dialogue utterances for English-Chinese language pairs. Liang et al. (2022a) observed visual features helps to generate better quality translation on multi-modal dialogue.

Our task is different from others in the following aspects: a). we focus on dialogue or chat translation in low-resource language settings and the e-commerce domain; b). the existing works primarily focus on dialogues or formal conversations. In contrast, we focus on noisy and informal conversations or chat (specifically for QnA); c). we perform experiments on the sentence-level system (transformer) with domain adaptation and transfer learning. We also report the evaluation results on a context-based system exploiting source-side context.

3 Corpus Creation

We create the English-Hindi parallel corpus from the existing English dialogue corpora. Specifically, we choose MultiModal Dialogue (MMD) corpus (Saha et al., 2018) and English part of German-English corpus from WMT20 chat translation task (Farajian et al., 2020), which is based on Taskmaster-1 corpus (Byrne et al., 2019). Further, we translate the Flipkart QnA corpus, consisting of 2.1M QnA pairs. This dataset contains the queries about a product asked by users of the website, which have been answered by either of (a). Other Customers or (b). Seller of the product. The QnA covers queries of a large range of products, including Electronics, Clothing, Appliances, etc., on the E-Commerce website, while the aspect is about the quality of the product, its features, compatibility, durability, and others.

We create two types of parallel corpora, synthetic corpus and gold standard corpus. The synthetic corpus is created through forward-translation or backward-translation (depending upon the translation direction of the task) of English corpora into Hindi with the existing NMT models. The gold standard corpus is created by manually translating English into Hindi. Table 1 shows the statistics of the prepared synthetic and gold standard corpora.

Subset	#Dialogues	Dialogue Avg. length	#Sentences	English Avg. length	Hindi Avg. length	Type
WMT20 Chat Corpus						
Train	550	25.17	13,845	8.07	9.08	Synthetic
Validation	78	24.61	1,902	8.08	9.11	Synthetic
Test	18	27.89	502	7.29	8.20	Manual
MMD Corpus						
Train	1,366	35.6	50,000	8.72	9.42	Synthetic
Validation	25	37.04	926	12.50	13.01	Manual
Test	46	34.97	1,609	12.78	13.61	Manual
QnA Corpus						
Train	-	-	4,191,608	5.31	6.15	Synthetic
Validation	-	-	1,000	25	28.9	Manual
Test	-	-	1,000	25.8	28.2	Manual

Table 1: Statistics of the created English-Hindi WMT20 Chat, MMD, and QnA parallel corpora. **#Dialogues**: Total number of dialogues, **Dialogue Avg. length**: Average dialogue length, **#Sentences**: Total number of sentences, **English Avg. length**: Average English sentence length, **Hindi Avg. length**: Average Hindi sentence length, **Type**: Type of translation (synthetic/manual). The QnA corpus consists of the question and answers pairs.

3.1 Synthetic Corpus Creation

We extract all English sentences from the training and validation sets of the WMT20 chat translation task. Similarly, we extract the first 50k sentences from the MMD corpus out of 5 million sentences in the dataset (Saha et al., 2018). The synthetic corpus is prepared by translating the extracted data through google translate[2] (Wu et al., 2016). Note that the translations generated are with a sentence-level NMT model. For Flipkart QnA data, we translate all the available data with a sentence-level MT system, which is trained on Samanantar Corpus-v3 (Ramesh et al., 2021) containing general-domain sentences. Table 1 shows the statistics of the prepared synthetic corpora. To assess the quality of synthetic corpus, we randomly select 100 English-Hindi sentence pairs from the prepared WMT20 chat, MMD, and QnA corpora and evaluate the quality of generated Hindi sentences based on adequacy and fluency. Table 2 shows the average adequacy and fluency scores of the generated Hindi sentences for all three domains. Based on the scores, we observe that the prepared synthetic corpora are of good quality, specifically the WMT20 chat and MMD corpora, where sentences are usually more formal and less noisy than QnA data. However, it is to be noted that despite its excellent quality, it lacks discourse awareness. Therefore, using these translation models directly (instead of training using discourse-aware models) may not be preferred.

Corpus	Adequacy	Fluency
WMT20 Chat	4.87	4.75
MMD	4.93	4.85
QnA	4.59	4.53

Table 2: Adequacy and Fluency of the prepared synthetic corpora. The scores are averaged from randomly picked 100 sentences from each corpus.

3.2 Gold-standard Corpus Creation

We create a gold-standard corpus by manually translating English data into Hindi. We extract the first 502 sentences (18 dialogues) from the WMT20 chat translation task testset and the first 1,436 sentences (43 dialogues) from the MMD dataset. We employed two annotators who are fluent in Hindi and English. The annotators are instructed to prefer commonly used words over conventional words for translation To be consistent with the nature of the dataset. They are further instructed to translate the sentences based on the previous sentences in a particular dialogue. For QnA, we observe that a large proportion of the answers are yes/no type, which may not be indicative of the true performance of the MT systems. Therefore, we extract 1,000 QnA pairs with large sentence lengths for answers to curate test and validation sets. We divide them into validation and test sets containing 500 QnA pairs (1,000 sentences) for each set. Then the annotators are instructed to translate the questions manually based on the corresponding answers and vice-versa. Table 1 shows the statistics of prepared gold-standard corpora.

4 Description of Corpora

This section describes the WMT20 chat, MMD, and QnA corpus.

[2]Translation through google translate is done between September-November 2021

WMT20 Chat Corpus: The WMT20 chat translation dataset (Farajian et al., 2020) is initially released for German–English language pair. The dataset is derived from Taskmaster-1 (Byrne et al., 2019) corpus, which is a monolingual conversational dataset.

The dataset consists of conversations from six domains. We create a Hindi version of this corpus following the process as described in Section 3. If the customer speaks in Hindi, it is to be translated into English, and if the assistant speaks in English, then it is to be translated into Hindi. The training, validation, and test sets contain 13,845, 1,902, and 502 sentences, respectively. The dialogues contain an average of 25 sentences. Table 1 shows the detailed statistics of the prepared Hindi–English WMT20 Chat corpus. The datasets can be divided into two subsets[3]. The agent and customer subsets contain 271 and 231 utterances on the English–Hindi test set.

MMD Corpus: MultiModal Dialogue Corpus (MMD) (Saha et al., 2018) is a monolingual dialogue dataset containing an English sentence or an image as an utterance, where the conversation is between shoppers and sales agents. The dataset is created by a semi-automatic method using automata and feedback from the experts. Since we require only conversations in text, we remove all the occurrences of the images. The training, validation, and test sets contain 50,000, 926, and 1,609 sentences. Table 1 shows the detailed statistics of prepared Hindi–English MMD corpus.

QnA Corpus We introduce a parallel (synthetic target side) QnA corpus containing 2.1M QA pairs (4.2M sentences). The dataset contains questions asked by the users and their answers, which are provided by either other users or seller of the product. The dataset contains queries about a wide range of products, including Electronics, Lifestyle, Appliances, etc. We specifically choose longer sentences for validation and test sets to avoid high performance due to shorter "Yes/No" type questions. Table 1 shows the statistics of prepared Hindi–English QnA corpus.

5 Methodology and Experimental Setup

We use Samanantar (Ramesh et al., 2021) corpus as a general domain corpus that contains 10M sentence pairs for English–Hindi. We train the transformer (Vaswani et al., 2017) model on the Samanantar corpus and consider it a baseline model.

5.1 Methodology

Domain Adaptation: We fine-tune the baseline model with the prepared WMT20 chat, MMD, and QnA corpora. The models are trained at the sentence level, and no context information is utilized for training the models or generating translation.

Transfer learning: Since we have two conversational datasets (WMT20 chat and MMD), we try to use knowledge from the trained model on one corpus to generate a better translation for the model trained on the other corpus via two-stage fine-tuning. For the first stage, we fine-tune the baseline model on the MMD corpus, and in the second stage, we fine-tune the model on the WMT20 chat corpus. Finally, we evaluate on WMT20 chat testset. Similarly, in the case of MMD experiments, for the first stage, we fine-tune WMT20 chat data and then MMD data in the second stage.

Domain adaptation with user specific context: We follow context-aware domain adaptation strategy proposed by Gain et al. (2021). In this approach, the context is selected by considering the previous utterances until the occurrence of an utterance from different speakers, subject to a maximum of three previous sentences. After selecting the

[3]The structure of the dataset is described at: https://github.com/Unbabel/BConTrasT

context, we add a special '<context>' token, representing the beginning of the context. We concatenate all the contexts, followed by a special '<end>' token, representing the end of the context. Then we concatenate the current source sentence and context and use them as input to the encoder.

For the QnA corpus, since there is no context associated with the question and answer pairs, we consider *answer* as a context for *question*, and vice-versa. We concatenate the question and answer pairs by the '<sep>' token and feed them to the encoder as input. For QnA, the decoder is trained to generate the translation for both questions and answers separated by '<sep>' at the output side. Encoding and translating the question and answer simultaneously makes the model effectively encode both questions and answers.

Pre-processing and Experimental Setup: Chat utterances are often written in lowercase and do not follow usual Capitalization Rules. Therefore, for the experiments in the English-Hindi direction, we convert every source sentence to lowercase. For Hindi-English, we do not convert to lowercase as the target side is English, and the true case of the sentences should be generated. Then, we jointly learn byte-pair-encoding (Sennrich et al., 2016) by combining source and target sides with fastBPE. We use fairseq (Ott et al., 2019) to train all our models. We use six layered encoder-decoder stacks with eight attention heads. The embedding and feed-forward layer sizes are set to 512 and 2048, respectively, with a dropout of 0.2. We set the maximum tokens per training batch to 4,000 with update frequencies of 64 and 4 during pre-training and fine-tuning, respectively. Thus maximum effective token per update during pre-training is (4000 * 64 * 2 GPUs) = 768,000. During fine-tuning, we use one GPU. We set an update frequency of 4 during fine-tuning on WMT20 and MMD datasets. Thus, effective number of tokens per update is (4000 * 4 * 1 GPUs) = 16,000. For all settings mentioned above, the initial learning rate is set to 0.0005, whereas 0.1 is set for label smoothing. We train the model for 30 epochs for pre-training and a maximum of up to 5000 updates during fine-tuning. Due to a large number of data in QnA, we fine-tune QnA models for a maximum of 10,000 updates and set the update frequency to 16. We deploy early stopping with patience set to five. We select the model checkpoint with the lowest perplexity on the validation set. We train our systems on two GeForce RTX 2080 Ti GPU with half-precision (FP16) for faster training. During the inference, the beam size to set to 5.

6 Results and Analysis

6.1 Results

We report BLEU[4] and TER[5] scores of all the trained models, calculated with sacreBLEU (Post, 2018). Tables 3 and 4 shows the BLEU and TER scores of all models on prepared corpora. For the WMT20 Chat dataset, the Baseline model achieves 39.1 and 43.8 BLEU points for Agent and Customer subsets, respectively. After domain adaptation, we achieve 19.1 and 17.5 BLEU score improvements. Transfer from MMD boosted BLEU by 1.3 points for the Customer Subset test set, mostly consisting of informal utterances. However, the improvement was 0.5 BLEU for the Agent Subset, containing mostly formal utterances. Further, we report our results on Overall (En-Hi), combining both user and customer subsets, then flip the source and target for the Customer subset (where the source is English and the Target is Hindi). Similarly, for

[4]sacreBLEU Signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0

[5]TER signature: nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.1.0

Overall (Hi-En), we flip the source and target side of the Agent subset. Although the context-based method caused degradation in the BLEU score, it can achieve the best TER score on the Customer subset.

Model	Customer		Agent		Overall (En-to-Hi)		Overall (Hi-to-En)	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
WMT20 Chat Results								
Baseline	43.8	47.2	39.1	48.1	37.5	48.8	37.0	49.2
Domain Adaptation	61.3	29.8	58.2	30.9	57.4	31.7	61.4	25.3
Transfer Learning	**62.6**	29.6	**58.7**	**30.6**	**57.9**	**31.3**	**61.7**	**25.1**
Domain Adaptation with Context	62.0	**29.0**	57.9	31.1	-	-	-	-
MMD Results								
Baseline	30.8	47.3	38.3	47.0	38.3	46.0	29.9	51.0
Domain Adaptation	76.6	17.4	52.4	**33.0**	56.4	29.6	62.3	26.4
Transfer Learning	76.8	17.7	**52.9**	33.4	**57.1**	**29.5**	**62.4**	**26.0**
Domain Adaptation with Context	**76.9**	**17.1**	52.3	34.1	-	-	-	-

Table 3: Results on English–Hindi WMT20 Chat and MMD corpora. Transfer Learning: Fine-tuning the model trained on the MMD corpus for the WMT20 Chat model and the model trained on the WMT20 Chat corpus for the MMD model. Translation direction of Agent and Customer subsets are En-Hi and Hi-En, respectively. Note that Agent and Customer correspond to System and User Subsets of MMD Data. En: English, Hi: Hindi.

Model	Questions		Answers		Overall	
	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	48.2	36.5	47.8	39.5	47.8	39.0
Domain Adaptation	49.2	36.3	49.1	37.3	49.1	37.1
Domain Adaptation with Context	50.1	35.8	49.9	36.9	**49.9**	**36.7**
Tagged Fine-tuning	**50.3**	35.9	48.7	37.2	48.9	37.0
Domain Specific NMT	50.2	**35.1**	**50.4**	**36.1**	-	-

Table 4: Results on English–Hindi QnA corpus. En: English, Hi: Hindi. Domain Adaptation with Context: Used question as the context for the answer and vice-versa; Tagged-Finetuning: Provided Question or Answer tags with sentences; Domain-Specific NMT: Models Trained on either question or answer.

For the MMD dataset, the Baseline model achieves 30.8 and 38.3 BLEU scores for user and system subsets. The model trained with the Domain Adaptation method achieves 45.8 and 14.1 BLEU improvement over the Baseline on Customer and Agent subsets, respectively. Transfer learning from the WMT20 Chat dataset yields a better signal and improves performance on all subsets regarding the BLEU score. However, it is noted that transfer learning models use more data than other models as they are fine-tuned on top of already fine-tuned models on WMT20 chat data.

Context-based model was able to improve BLEU by 0.3 for *user* subset whereas BLEU score decreased for *System* subset. The system subsets usually consist of a long description of products and sufficient information to translate the sentence. The context here acts as noise and negatively affects the translation quality.

For the QnA dataset, we achieved a 47.8 BLEU score and 39.0 TER on the baseline MT system, which is trained on general domain data. After fine-tuning on in-domain

data, we achieve a BLEU score of 49.1, improving the 1.3 BLEU score over the Baseline. For fine-tuning with context, we achieve a BLEU score of 49.9, which is better by 0.8 than the previous model.

Effects of tagged fine-tune: Tagged fine-tune is a popular method to train the models on diverse datasets. (Caswell et al., 2019) used tags for back-translated (BT) sentences to train a model with a combination of bitext and BT data. We supply <question> tag for every question and <answer> tag for every answer. This is to help the model learn to question and answer specific properties. While this method improved the translation quality of Questions, the same was not reflected in Answers. This can be attributed to the fact that questions are usually ambiguous as there are missing question marks, grammatical errors framing questions as statements, etc. Tag for questions helped to disambiguate such errors, but the tags were not of much use for answers.

Domain specific NMT: We divide our data into two subsets: (i). Questions and (ii). Answers. We train two separate sentence-level MT systems on the two subsets to help the model learn question/answer property without other domain data. We achieve a 50.4 BLEU score with this method on the Answer subset, which is the best among all methods. We suggest this is due to the absence of a question subset during training. The question subsets negatively impact the learning of answer translation as they contain a more significant portion of erroneous references on the target side due to mistranslation. On the question subset, we obtain a BLEU score of 50.2, which is 0.1 less than that of the best model. It is to be noted that each of the systems is trained on only mutually exclusive 50% of the dataset[6]. Therefore, it generates a bit poor translation than the model trained on complete data.

Context	6, okay great! let me catch you the rates really quickly.
Source	The rate I found for an UberXL will be $45.66.
Translation without Context	मुझे uberxl के लिए $45.66 की दर मिलेगी। *(For an uberxl, the rate I will find is $45.66)*
Translation with Context	मुझे एक uberxl के लिए $45.66 की दर मिली। *(For an uberxl, the rate I found is $45.66)*

Table 5: Example of generated output sentences with and without context.

6.2 Analysis

In the example about ride-booking from Table 5, translations from non-context based systems are in future tense. This can be attributed to the presence of the word *will* in the source utterance. However, translation with context was able to translate it correctly to present tense.

6.3 Quality testing

The proposed model is evaluated in the well-known E-commerce industry, Flipkart[7] with the help of real-time human evaluators. The evaluators rated them with respect to the scale of 1-3, where 1-*Bad*, 2-*Can be Better* and 3- *Good*. During the evaluation, while assigning the labels to the output samples, 'tense preservation,' 'syntax of output

[6]The in-domain dataset contains 50% questions and 50% answers. However, these models are trained on either questions or answers, not both. These (questions or answers) are effectively 50% of the available in-domain dataset

[7]`https://www.flipkart.com/`

Model	WMT20 Chat			MMD			QnA		
	Good	Can be better	Bad	Good	Can be better	Bad	Good	Can be better	Bad
Baseline	28%	60%	12%	40%	56%	4%	29%	36%	34%
Domain-Adaptation	76%	20%	4%	68%	28%	4%	31%	35%	34%
Context-Based	76%	20%	4%	64%	32%	4%	30%	38%	32%

Table 6: Real-time quality evaluation of trained models on the prepared corpora between Baseline, Transfer Learning, and Domain Adaptation with Context models.

sentence,' 'choice of in-domain output tokens' are some important factors that are kept in mind. Table 6 shows the statistics of the real-time evaluation. A sample of 25 sentences from Baseline and Domain-Adaptation, Domain Adaptation with Context models for analysis for both MMD and WMT20 chat corpora. For the MMD corpus, 4% of the translations are rated as *Bad* from each model. While 56% and 40% are rated as *Can be better* and *Good*, respectively, for Baseline. Domain-Adaptation model achieves the best rating with 68% as *Good* and 28% as *Can be better*. Similarly for WMT20 Chat corpus, the baseline model outputs are rated as, 28% as *Good* quality, 60% & 12% as *Can be better* and *Bad* respectively. Domain Adaptation and Domain Adaptation with Context models achieves 76%, 20% and 4% as *Good*, *Can be better* and *Bad*, respectively.

For the QnA corpus, a sample of 250 sentences is taken for evaluation. 34% of the translations are rated as *Bad* from Baseline and Domain Adaptation and 32% for Domain Adaptation with Context model. While 35% and 31% are rated as *Can be better* and *Good* for Domain Adaptation model respectively. For Domain Adaptation with Context model, 38% were rated as *Can be better* and 30% were rated as *Good*. Based on the evaluation results, the Domain Adaptation method significantly improves the "Good" category from the "Can be better" category.

7 Conclusion

Dialogue translation is different from sentence translation due to several additional challenges. The particular field could not be adequately explored in Indian languages due to a lack of datasets. Multilingual chatbots are in high demand as many of the world's population are not fluent in English and other major languages. We introduce two English–Hindi parallel datasets for Dialogue Translation and one large-scale English–Hindi parallel corpus for QnA translation. The datasets contain various domains, including fashion, making reservations, ordering foods, E-commerce products, etc., and contain sentences from different roles and languages. Demands for online chatbot services involving the domains mentioned above are tremendous. The multilingual property of the datasets will be beneficial in building multilingual chatbots and QnA translators. We report a baseline result on some sentence-level and a context-level model exploiting context. In the future, we would like to introduce better ways to utilize context and use other chat-specific characteristics, including informality, code-mixing, etc. Further, we would like to extend the work into other Indian languages, including Tamil, Telugu, Malayalam, etc.

Acknowledgment

Authors gratefully acknowledge the unrestricted research grant received from the Flipkart Internet Private Limited to carry out the research. Authors thank Anubhav Tripathee for gold standard parallel corpus creation and translation quality evaluation.

References

Bao, C., Shiue, Y.-T., Song, C., Li, J., and Carpuat, M. (2020). The University of Maryland's Submissions to the WMT20 Chat Translation Task: Searching for More Data to Adapt Discourse-Aware Neural Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 456–461.

Berard, A., Calapodescu, I., Nikoulina, V., and Philip, J. (2020). Naver Labs Europe's Participation in the Robustness, Chat, and Biomedical Tasks at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 460–470.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., and Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy.

Chen, Q., Lin, J., Zhang, Y., Ding, M., Cen, Y., Yang, H., and Tang, J. (2019). Towards knowledge-based recommender dialog system. 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 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069.

Deng, Y., Zhanng, W., and Lam, W. (2020). Opinion-aware answer generation for review-driven question answering in e-commerce. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Farajian, M. A., Lopes, A. V., Martins, A. F. T., Maruf, S., and Haffari, G. (2020). Findings of the WMT 2020 Shared Task on Chat Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 64–74, Online.

Gain, B., Haque, R., and Ekbal, A. (2021). Not all contexts are important: The impact of effective context in conversational neural machine translation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Lai, T. M., Bui, T., Lipka, N., and Li, S. (2018). Supervised transfer learning for product information question answering. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1109–1114.

Liang, Y., Meng, F., Chen, Y., Xu, J., and Zhou, J. (2021a). Modeling bilingual conversational characteristics for neural chat translation. 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 5711–5724, Online. Association for Computational Linguistics.

Liang, Y., Meng, F., Xu, J., Chen, Y., and Zhou, J. (2022a). MSCTD: A multimodal sentiment chat translation dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2601–2613, Dublin, Ireland. Association for Computational Linguistics.

Liang, Y., Meng, F., Xu, J., Chen, Y., and Zhou, J. (2022b). Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388, Dublin, Ireland. Association for Computational Linguistics.

Liang, Y., Zhou, C., Meng, F., Xu, J., Chen, Y., Su, J., and Zhou, J. (2021b). Towards making the most of dialogue characteristics for neural chat translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 67–79, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Moghe, N., Hardmeier, C., and Bawden, R. (2020). The University of Edinburgh-Uppsala University's Submission to the WMT 2020 Chat Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 471–476.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation*, pages 314–319, Florence, Italy.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, MN.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels.

Provilkov, I., Emelianenko, D., and Voita, E. (2020). BPE-Dropout: Simple and Effective Subword Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892.

Qu, C., Yang, L., Qiu, M., Zhang, Y., Chen, C., Croft, W. B., and Iyyer, M. (2019). Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1391–1400, New York, NY, USA. Association for Computing Machinery.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. (2021). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Saha, A., Khapra, M. M., and Sankaranarayanan, K. (2018). Towards building large scale multimodal domain-aware conversation systems. In *AAAI*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Snover, M. G., Dorr, B., Schwartz, R. M., Micciulla, L., and Weischedel, R. M. (2005). A study of translation error rate with targeted human annotation.

Sun, Y. and Zhang, Y. (2018). Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 235–244, New York, NY, USA. Association for Computing Machinery.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA.

Wang, L., Tu, Z., Wang, X., Ding, L., Ding, L., and Shi, S. (2020). Tencent AI Lab Machine Translation Systems for WMT20 Chat Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 483–491.

Wang, T., Zhao, C., Wang, M., Li, L., and Xiong, D. (2021). Autocorrect in the process of translation — multi-task learning improves dialogue machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 105–112, Online. Association for Computational Linguistics.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yeganova, L., Wiemann, D., Neves, M., Vezzani, F., Siu, A., Jauregi Unanue, I., Oronoz, M., Mah, N., Névéol, A., Martinez, D., Bawden, R., Di Nunzio, G. M., Roller, R., Thomas, P., Grozea, C., Perez-de Viñaspre, O., Vicente Navarro, M., and Jimeno Yepes, A. (2021). Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.

Yu, L., Sartran, L., Huang, P.-S., Stokowiec, W., Donato, D., Srinivasan, S., Andreev, A., Ling, W., Mokra, S., Dal Lago, A., Doron, Y., Young, S., Blunsom, P., and Dyer, C. (2020). The DeepMind Chinese–English document translation system at WMT2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 326–337, Online. Association for Computational Linguistics.

Zhang, Y., Chen, X., Ai, Q., Yang, L., and Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 177–186, New York, NY, USA. Association for Computing Machinery.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.