Investigating Effectiveness of Multi-Encoder for Conversational Neural Machine Translation

Baban Gain$^1$, Ramakrishna Appicharla$^1$, Soumya Chennabasavraj$^2$, Nikesh Garera$^2$, Asif Ekbal$^1$ and Muthusamy Chelliah$^2$

$^1$Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
{gainbaban, ramakrishnaappicharla, asif.ekbal}@gmail.com

$^2$Flipkart, India
{ soumya.cb, nikesh.garera, muthusamy.c}@flipkart.com

Abstract
Multilingual chatbots are the need of the hour for modern business. There is increasing demand for such systems all over the world. A multilingual chatbot can help to connect distant parts of the world together, without sharing a common language. We participated in WMT22 Chat Translation Shared Task. In this paper, we report descriptions of methodologies used for participation. We submit outputs from multi-encoder based transformer model, where one encoder is for context and another for source utterance. We consider one previous utterance as context. We obtain COMET scores of 0.768 and 0.907 on English-to-German and German-to-English directions, respectively. We submitted outputs without using context at all, which generated worse results in English-to-German direction. While for German-to-English, the model achieved a lower COMET score but slightly higher chrF and BLEU scores. Further, to understand the effectiveness of the context encoder, we submitted a run after removing the context encoder during testing and we obtain similar results.

1 Introduction
Translation of Dialogues is a crucial part of building multilingual chatbots. With easier access to the internet than ever, we have the opportunity to connect with different people with different languages. However, language remains a barrier to smooth communication. Using automated machine translation systems can alleviate such issues. However, most of the general MT systems are not very suitable for conversations. This is due to additional challenges chat translation possesses that general domains do not have. This includes the presence of noisy utterances. Compared to other domains, chat is more prone to contain noisy sentences. This comes from multiple sources, as follows. a) Keyboard typos: Spelling mistakes that occurred due to quick typing. In this case, often, some characters are replaced by nearby characters on the keyboard. Further, the insertion of extra characters or the absence of some characters is also common. b) Intentional shortening of Words: Users often use short forms of words by removing certain characters (primarily vowels) while keeping the pronunciation similar to the correct word (For example, ‘hw’ instead of ‘how’). c) Grammatical Errors: Conversations usually occur in an informal setting, and grammar is mostly ignored as long as the meaning is understood correctly. However, this makes it difficult to translate. Further, there are other challenges, like context dependency. That is, the utterances can be ambiguous, and the correct meaning of an utterance can not be understood without referring to its dialogue history.

In this paper, we use a multi-encoder transformer to translate chat utterances. We use six encoder layers for source text and one encoder layer for context. For better comparison, we have submitted translations from two other models. To test the effectiveness of context, we did not provide context during the testing phase as described in section 3.3.2. Further, we train another model without using any context at all as described in 3.3.3. We achieved very competitive results for the Agent subset (English-to-German), where we obtained 0.551 BLEU, 0.730 chrF, and 0.768 COMET scores, where the best result among primary submissions of the participants are 0.555, 0.735, and 0.810 BLEU, chrF and COMET score respectively. For German-to-English, our method produced 0.907, 0.729, and 0.587 COMET, chrF, and BLEU scores, respectively.

2 Related Work
The area of chat translation mostly remained unexplored until recent years. This is in part due to the unavailability of suitable dialogue datasets. Farajian et al. (2020) introduced an German–English parallel conversational corpus. Berard et al. (2020) proposed a method that replaced rare characters...
Figure 1: Diagram of our model; The weight is determined by a FFN from concatenated representations of the attentions.

with a special ‘<copy>’ token, which helps the model to learn when to copy the tokens from source to target. Further, they used methods like inline casing, tagged back-translation (BT) (Caswell et al., 2019), Byte-Pair-Encoding (BPE) (Sennrich et al., 2016), and ensemble of models using domain-specific adaptive layers, etc. Ensemble model with a domain-specific adaptor layer generated the best translation on WMT20 Chat data. Moghe et al. (2020) used fine-tuned pre-trained models (Ng et al., 2019) on the pseudo-in-domain and in-domain data. Wang et al. (2020) used using three previous contexts along with the current sentence for adaptation of Cross-lingual Language Model Pre-training (Conneau and Lample, 2019) objectives into document-level NMT. Bao et al. (2020) used an additional encoder to process one previous context. However, adding an additional encoder did not result in consistent improvement in translation. Gain et al. (2021c) proposed a rule-based context selection technique where previous sentences by the same user are used to enhance the translation quality. This mainly helped to translate anaphoric pronouns correctly. Liang et al. (2021a) introduced a conditional variational autoencoder (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 objective. 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. Further, they propose a multi-tasking algorithm that helped to generate better translation than traditional multi-tasking. Wang et al. (2021) proposed a multi-task learning-based model that identifies missing pronouns, typos and utilizes context to translate chat utterances. Liang et al. (2022a) observed visual features helps to generate better quality translation on multi-modal dialogue. Apart from chat translation, context is commonly used in other translation tasks as well. This include document translation (Kim et al., 2019; Zhang et al., 2018; Läubli et al., 2018) where other sentences from the document is used as context, multimodal translation (Yao and Wan, 2020; Gain et al., 2021a,b) where image features are used as context, etc. Gain et al. (2022) proposed a method where context is concatenated with source on both source and target side, requiring the model to translate context also, thus avoiding ignorance of context in Question-Answer translation.
3 Methodology

3.1 Pre-Training

Pre-training models with general domain data and transferring the knowledge to intended domain is standard practice in MT. We use Facebook AI’s pre-trained models (Ng et al., 2019) from WMT19. The pre-training methodology consists of data processing techniques like normalize punctuation and tokenizing all data with the Moses tokenizer (Koehn et al., 2007) and byte-pair-encoding (Sennrich et al., 2016). Further, sentences with wrong language on either source or target side filtered out with language identification (Lui and Baldwin, 2012) filtering.

3.2 Model

We use a dual encoder-based transformer model. The components of the models are as follows:

- **Source Encoder**: Source Encoder consists of 6 standard transformer encoder layers. For all our models, the encoder weights are initialized from the pre-trained models. The input language of source encoder is the input language of the translation direction. That is, for English-to-German model, the language for Source Encoder is English.

- **Context Encoder**: Context Encoder of consists of 1 encoder layer. This is in part to keep model parameters lower. Further, context is supposed to assist the translation process. Thus has limited contribution compared to source. The language of the context encoder can be English or German, depending upon speaker of the previous utterance, irrespective of translation direction. We take one previous utterance from source side of previous speaker. That is, English if the speaker of previous utterance is agent or German if speaker of the previous utterance is Customer. For first utterance in a conversation, the context is empty.

- **Decoder**: Decoder consists of 6 layers of standard transformer decoder layers. We initialize the decoder from the pre-trained model. Further, in addition to encoder-decoder attention, we perform context-decoder attention. Then, we concatenate them before passing it to a feed-forward Neural Network (FFN) which determines weighted average factor g. Inspired from (Libovický et al., 2018), we take final attention output as g * context-decoder attention + (1-g) * encoder-decoder attention. The rest parts of the decoder is similar to standard transformer decoder.

3.2.1 Stage-1 Fine-tuning

For all our submissions, we perform two-stage fine-tuning. Due to the unavailability of the training set in the task, we fine-tune the model on WMT20 Chat Task (Farajian et al., 2020) data. However, since our objective is to get the highest results for WMT22 version of chat data, we use that as a validation set.

3.2.2 Stage-2 Fine-tuning

We finetune the models obtained from Stage-1 fine-tuning with WMT22 Chat Task Dev Subset. We fine-tune the models for 15 epochs. Since we are using validation set for training, we did not use any validation at this stage. We use last checkpoint from this stage as the final model and use it for testing.

3.3 Submitted Models

We submit our results for English-to-German and German-to-English directions. For each direction, we submit three results. We do not freeze any parameters during fine-tuning process for all of our submissions.

3.3.1 Primary

In our primary submission, we use the model as described in Section 3.2. We use one previous utterance as context during training, validation, and testing. This model consists of about 359M parameters.

3.3.2 Contrastive-1

Li et al. (2020) suggested that improvement in translation quality is observed after introduction of context encoder. However, it can be attributed to the contextual information acting as noise, rather than rich information relevant to the source or target. They showed that, even if context is not used during testing, the models produce similar results due to the fact that the context used during training helped the model for robust training. While this observation was for document translation, we use this method for chat translation. Thus, in this
## Table 1: Results of submissions at WMT22 Chat task for En–De; C1: contrastive-1 submission; C2: contrastive-2 submission

<table>
<thead>
<tr>
<th>Models</th>
<th>En-De (agent)</th>
<th>De-En (customer)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COMET</td>
<td>chrF</td>
</tr>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline without context</td>
<td>0.403</td>
<td>0.550</td>
</tr>
<tr>
<td>Baseline with context (N=2)</td>
<td>0.376</td>
<td>0.537</td>
</tr>
<tr>
<td><strong>Primary submissions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BJTU-WeChat</td>
<td>0.810</td>
<td>0.735</td>
</tr>
<tr>
<td>Unbabel-IST</td>
<td>0.774</td>
<td>0.733</td>
</tr>
<tr>
<td>Our Submission</td>
<td>0.768</td>
<td>0.730</td>
</tr>
<tr>
<td>HW-TSC</td>
<td>0.704</td>
<td>0.725</td>
</tr>
<tr>
<td><strong>Contrastive submissions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BJTU-WeChat, C1</td>
<td>0.804</td>
<td>0.731</td>
</tr>
<tr>
<td>BJTU-WeChat, C2</td>
<td>0.805</td>
<td>0.738</td>
</tr>
<tr>
<td>Unbabel-IST, C1</td>
<td>0.780</td>
<td>0.737</td>
</tr>
<tr>
<td>Unbabel-IST, C2</td>
<td>0.778</td>
<td>0.734</td>
</tr>
<tr>
<td>Our Submission (C1)</td>
<td>0.769</td>
<td>0.730</td>
</tr>
<tr>
<td>Our Submission (C2)</td>
<td>0.765</td>
<td>0.729</td>
</tr>
<tr>
<td>HW-TSC, C1</td>
<td>0.649</td>
<td>0.670</td>
</tr>
<tr>
<td>HW-TSC, C2</td>
<td>0.726</td>
<td>0.732</td>
</tr>
</tbody>
</table>

### 3.3.3 Contrastive-2

We submit the results from a model without using any context for better comparison. Note that this model is trained with all other methodologies similar to Primary and Contrastive-1, which includes two-stage pre-training with the same data.

### 3.4 Post-Processing

We remove `<unk>` from the output. Further, we observe tags and modify them to the original tag, if mistranslated. For Example, we change "#PRS_ORG#" to "#PRS_ORG#", "# Address #" to "#ADDRESS#", etc.

## 4 Results

We obtain a COMET (Rei et al., 2020) score of 0.768 and 0.907 on En-De and De-En directions. Further, we obtain chrF (Popović, 2015) scores of 0.730 and 0.729 for En-De and De-En. We obtain BLEU scores of 0.551 and 0.587 for Agent and Customer subsets. With contrastive-1 submission, we obtain similar results. For Agent subset, COMET score improved by 0.001 whereas, decreased by 0.002 for Customer subset. Similarly for contrastive-2 submission, COMET decreased by 0.003 whereas chrF and BLEU score decreased by 0.001 and 0.006 respectively for Agent subset. Without context method generated better results for Customer subset, improving BLEU and chrF by 0.005 and 0.002 respectively, whereas we observe a decrease of 0.005 on COMET metric. Thus, our experiment suggests that the usage of context played a very limited role in the submitted systems. We suggest this is due to a lower Context Window in our experimental setting. We use only one previous sentence as a context. While it has been observed...
that using one context is usually sufficient on conversational or document-level datasets, WMT22 Chat Task data contain very shorter and repetitive sentences. This includes one or two word utterances (Thanks, #EMAIL#, #NAME#, Good Bye, etc), App navigational information (Tap Settings, Tap Device information, etc), etc. These utterances has very limited information to be useful as a context. Further, appearance of duplicate utterances is a challenge during training process. However, unlike general MT, conversational datasets can not be de-duplicated easily. This is because removal of some utterance from a conversation will break its structure and might not be as meaningful.

5 Conclusion

Task translation is a challenging and important task for our society. One of the major challenges in chat translation is context-dependency. We participated in WMT22 Chat Translation Task, where we submit results obtained from multi-encoder based transformer model. We obtain COMET scores of 0.768 and 0.907 on English-to-German and German-to-English directions, respectively. We found that role of context in our experimental setting is limited. In future, we would like to explore these methods with larger window size. Further, we would like to explore data de-duplication strategies for conversations.

References


Baban Gain, Rejwanal Haque, and Asif Ekbal. 2021c. 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.


Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020.


