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

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 949 - 954
December 7-8, 2022

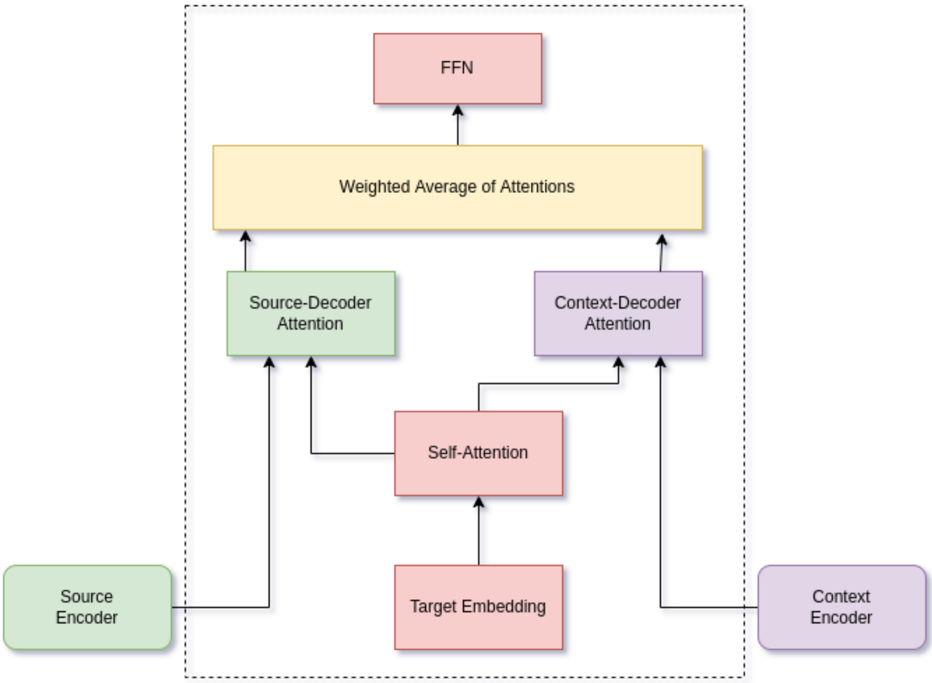

Figure 1: Diagram of our model; The weight is determined by a FFN from concatenated represenations of the attentions

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 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 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 [1]. 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 enocder-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

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wmt19/README.md

Models	En-De (agent)			De-En (customer)		
	COMET	chrF	BLEU	COMET	chrF	BLEU
Baselines						
Baseline without context	0.403	0.550	0.325	0.588	0.621	0.472
Baseline with context (N=2)	0.376	0.537	0.308	0.680	0.642	0.493
Primary submissions						
BJTU-WeChat	0.810	0.735	0.555	0.946	0.775	0.649
Unbabel-IST	0.774	0.733	0.555	0.915	0.737	0.612
Our Submission	0.768	0.730	0.551	0.907	0.729	0.587
HW-TSC	0.704	0.725	0.552	0.918	0.766	0.642
Contrastive submissions						
BJTU-WeChat, C1	0.804	0.731	0.550	0.948	0.780	0.650
BJTU-WeChat, C2	0.805	0.738	0.560	0.951	0.778	0.652
Unbabel-IST, C1	0.780	0.737	0.558	0.924	0.741	0.616
Unbabel-IST, C2	0.778	0.734	0.554	0.925	0.743	0.615
Our Submission (C1)	0.769	0.730	0.551	0.905	0.729	0.587
Our Submission (C2)	0.765	0.729	0.545	0.902	0.731	0.592
HW-TSC, C1	0.649	0.670	0.473	0.909	0.755	0.618
HW-TSC, C2	0.726	0.732	0.559	0.929	0.767	0.641

Table 1: Results of submissions at WMT22 Chat task for En–De; C1: contrastive-1 submission; C2: contrastive-2 submission

submission, we use the same model as on Primary submission, but we ignore the context during testing.

	Context Encoder		Parameters	
Submission	Training	Testing	Training	Testing
Primary	Yes	Yes	359M	359M
contrastive-1	Yes	No	359M	313M
contrastive-2	No	No	313M	313M

Table 2: Comparison of methodologies for our submissions

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 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

Calvin Bao, Yow-Ting Shiue, Chujun Song, Jie Li, and Marine Carpuat. 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.

Alexandre Berard, Ioan Calapodescu, Vassilina Nikoulina, and Jerin Philip. 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.

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

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

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online. Association for Computational Linguistics.

Baban Gain, Ramakrishna Appicharla, Soumya Chennabasavraj, Nikesh Garera, Asif Ekbal, and Muthusamy Chelliah. 2022. Low resource chat translation: A benchmark for Hindi–English language pair. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 83–96, Orlando, USA. Association for Machine Translation in the Americas.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021a. Experiences of adapting multimodal machine translation techniques for Hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44, Online (Virtual Mode). INCOMA Ltd.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021b. IITP at WAT 2021: System description for English-Hindi multimodal translation task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 161–165, Online. Association for Computational Linguistics.

Baban Gain, Rejwanul 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.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

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

Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 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.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 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.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 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.

Yunlong Liang, Chulun Zhou, Fandong Meng, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 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.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea.

Nikita Moghe, Christian Hardmeier, and Rachel Bawden. 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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation*, pages 314–319, Florence, Italy.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. Tencent AI Lab Machine Translation Systems for WMT20 Chat Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 483–491.

Tao Wang, Chengqi Zhao, Mingxuan Wang, Lei Li, and Deyi Xiong. 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.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.