The University of Maryland's Submissions to the WMT20 Chat Translation Task: Searching for More Data to Adapt Discourse-Aware Neural Machine Translation

Calvin Bao* **Yow-Ting Shiue*** **Chujun Song** **Jie S. Li** **Marine Carpuat**

Department of Computer Science, University of Maryland, College Park, MD

{csbao,ytshiue,cjsong,jli2718,marine}@cs.umd.edu

Abstract

This paper describes the University of Maryland's submissions to the WMT20 Shared Task on Chat Translation. We focus on translating agent-side utterances from English to German. We started from an off-the-shelf BPE-based standard transformer model trained with WMT17 news and fine-tuned it with the provided in-domain training data. In addition, we augment the training set with its best matches in the WMT19 news dataset. Our primary submission uses a standard Transformer, while our contrastive submissions use multi-encoder Transformers to attend to previous utterances. Our primary submission achieves 56.7 BLEU on the agent side (en→de), outperforming a baseline system provided by the task organizers by more than 13 BLEU points. Moreover, according to an evaluation on a set of carefully-designed examples, the multi-encoder architecture is able to generate more coherent translations.

1 Introduction

Recent advances have made MT a widespread tool for asynchronous consumption of text. The dream of dissolving language barriers, however, will not be fulfilled until MT enables two or more people carry on a synchronous conversation, each speaking their native languages. Building translation systems that enable seamless conversations between an English-speaking customer support agent and a German-speaking customer is the goal of WMT20's shared task of chat translation (Farajian et al., 2020). In participation of this shared task, we focused on the agent side, translating English utterances into German. Our methods are inspired by Voita et al. (2018) and Bawden et al. (2018), explicitly leveraging broader context to address coreference and cohesion to improve translation quality. We compare architectures of a standard transformer with a single encoder and a multi-encoder one with an additional transformer encoder to incorporate information from the previous utterance. In the case of blind testing or production use, since customer target utterances (English) will not be given, a separate de→en model was trained and used to back-translate customer utterances.

Additionally, given the limited training pairs, we experiment with augmenting our dataset. We selected a subset of WMT19 en-de news data that were similar to the chat training data, which we then added to the training data. The subset was constructed using a full-text search engine loaded with the entire en-de WMT19 news data, which iterated through each chat training example, querying for the two closest matches with both the source and target as search strings.

Our primary system, denoted `PRIMARY`, is a single-encoder pretrained transformer fine-tuned on WMT20 Chat data. The first contrastive system, denoted `CONTRASTIVE1`, is a multi-encoder transformer that pre-warms, using WMT19 news data, the weights of an additional encoder after loading the pretrained transformer. The second contrastive system, denoted `CONTRASTIVE2`, is a multi-encoder transformer that fine-tunes the pretrained transformer on a combination of WMT19 news data and WMT20 chat data.

2 Related Work

One of the main challenges for translating discourse arises from ambiguities of sentences when they are taken out of context, as MT models often do (Yamashita et al., 2009). Especially in dialogue, sentences tend to reference entities in previous sentences, which necessitates using cross-sentential information to translate a given sentence. Individual words can be translated in different ways,

*Equal contribution.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 456–461
Online, November 19–20, 2020.

significantly varying the meaning of the resulting sentence in a larger context (Gao et al., 2015). In addition, dialogue in the customer support domain is a distinctive and spontaneous category of text, with colloquialisms, errors and minimal revisions. All of these deviations can accumulate error throughout the course of a conversation.

Dialogue Translation: Specific interests in translating dialogues can be found as early as Lee and Kim (1997)'s work on Korean-English dialogue translation based on syntactic patterns and *n*-grams. Though their model parses sentences into speech acts instead of generating full-sentence translations, they have pointed out the importance of context (previous sentence) in interpreting the current sentence properly. The most relevant recent work is (Maruf et al., 2018), in which contexts for both source-side and target-side are utilized as additional generation conditions for the decoder in their NMT model. Several variants of the model architecture and the attention mechanism are explored. However, their experiments are conducted on Europarl and OpenSubtitles. The former is formal language and the latter scripted conversations of movies and TV. Here, in contrast, chat data is informal unscripted real-world language.

Context-Aware Machine Translation: Chat translation can be regarded as a special case of context-aware translation. Jean et al. (2017) extends the vanilla attention-based neural MT model (Bahdanau et al., 2015) by conditioning the decoder on the previous sentence via attention over its words. Wang et al. (2017) propose a cross-sentence context-aware model. They integrate the historical representation into NMT with two strategies: a warm-start of encoder and decoder states, and an auxiliary context source for updating decoder. Bawden et al. (2018) use multi-encoder NMT models to exploit context from the previous source and target sentence. Voita et al. (2018) propose a context-aware model based on the Transformer. Their model controls the flow of information from the extended context and improves on pronoun translation.

NMT Facilitated with Retrieved Translations: There is a line of NMT research inspired by example-based translation systems that aims to generate better translations by retrieving and referencing additional translation pairs. Gu et al. (2018) utilize an off-the-shelf search engine to retrieve training sentence pairs whose source side is similar to a given source sentence and incorporate them as additional input to the decoder. Zhang et al. (2018) use the retrieved examples at prediction time to upweight outputs whose constituents match retrieved *n*-gram translation candidates. In a similar vein, but at training time rather than prediction time, we use a retrieval system to select similar examples from a larger dataset to augment the smaller in-domain training set.

3 Data Preparation

3.1 Preprocessing

We used the Moses toolkit (Koehn et al., 2007) to preprocess our data. The training corpus was tokenized and cleaned. After that, we applied byte pair encoding (BPE) (Sennrich et al., 2016) on the data with the BPE model learned on the data of the pretrained model (Section 4.1.1). Following the pre-trained model, we use its shared vocabulary for both target and source sides. The size of the vocabulary, which is the union of English and German tokens, is 36,628.

3.2 Retrieval-Based Training Data Augmentation

There were only 13.85k utterances in the provided parallel WMT20 Chat training data. Given the limited data, we start off with a model pretrained on the WMT17 en-de news data, and additionally augment our training data with a filtered set of 4.75k lines of WMT19 en-de news data. We adopted Elasticsearch[1] to build a fast full-text search engine on the entire WMT19 en-de news set, and then iterated through each (source, reference) pair in the Chat training data. With each pair, we used the search engine to find the top two matches with the current source and target as search strings. We truncated this set to 4.75k training samples to limit the possibility of overwhelming our fine-tuning set and denote it *Chat-Similar News*. This technique brings the total training set to 18.6K parallel utterances.

4 Experiments

We conducted varied experiments in the English-German direction. We included the English reference of the customer utterances as training data for the scope of these experiments, even though this would not be available in a production setting. This was a strategy to provide more training pairs to our

[1] https://www.elastic.co/

models, knowing that the English references for customers is natural language, according to task organizers.

4.1 Systems Overview

We base all our systems off the Transformer architecture. Our implementation is based on the JoeyNMT toolkit (**?**). We kept hyperparameters common throughout. We experimented with the following settings:

1. Trained a standard single-encoder Transformer model.

2. Introduced a second encoder into our NMT architecture to process the preceding sentence, using context-target attention along with source-target attention to compute the final encoder hidden state, on a combination of *Chat-Similar News* and Chat.

3. As in item 2, introduced a second encoder and pre-warmed that encoder's weights on *Chat-Similar News*, before fine-tuning on Chat.

4.1.1 Off-the-shelf Pretrained Model

We found that an existing model trained on a different domain can generalize to this smaller dataset. We downloaded model weights for WMT17 en-de, provided by JoeyNMT Transformer[2]. This model was able to adapt to the chat domain, so the pretrained model was used for all experimental settings.

4.1.2 Common Hyperparameters

We kept hyperparameters consistent across models we tested, with some exceptions to account for slight differences in architecture. All models had embedding and hidden layers with 512 units, and feed-forward layers with 2048 units. A dropout rate of 0.1 was used on both the encoder and decoder layers. Training was performed with the Adam optimizer and in minibatches of 2048 tokens, with cross-entropy loss, an initial learning rate of 0.0002, and a patience of 8 validation cycles. All models were trained for a maximum of 65 epochs. The checkpoint with lowest validation perplexity is selected as the final model. For all validation cycles, greedy decoding is adopted. For testing, we used beam search decoding with a beam width of 5.

[2] https://www.cl.uni-heidelberg.de/statnlpgroup/joeynmt/wmt_ende_transformer.tar.gz

4.2 Single-encoder Implementation: `PRIMARY`

We trained a discourse-agnostic Transformer model with self-attention. This model had 6 layers for the both the encoder and decoder, each with 8 attention heads. A single-encoder implementation fine-tuned only on the Chat data was used to produce the primary submission results. We selected this model due to its slightly higher Chat validation BLEU (Table 1). It also achieves the highest test BLEU but with only minor differences with the contrastive systems. However, the gaps between it and the two contrastive multi-encoder implementations are not wide as can be seen.

4.3 Multi-encoder Implementation

Two context-aware models that are partial extensions of that described in Voita et al. (2018) were produced for the contrastive submissions. Voita et al. (2018)'s context-aware model encodes a source sentence and a context sentence independently and applies a gating function to produce a context-aware representation of the source sentence. We explored this combination idea by implementing a trainable gating function, *à la* (Voita et al., 2018), that takes the independently encoded source-side context and independently encoded source-side sentence as inputs to generate a representation for the decoder. Each layer retained 8 attention heads. We used 6 layers in each encoder. The total number of trainable parameters can be seen in Table 2.

4.3.1 Incremental Domain Adaptation: `CONTRASTIVE1`

This system has two steps: we pre-warm the context encoder of a multi-encoder implementation by fine-tuning on *Chat-Similar News* and validating on a subset of the Chat training data. We then fine-tune this intermediate model on the Chat training data, validating against Chat validation data. We consider this an incremental domain adaptation technique because we prewarm the trainable parameters of a new encoder with similar data, before finally tuning on the Chat data. Compared to a multi-encoder baseline implementation trained strictly on Chat, we achieve a 1.79 validation BLEU point improvement. Compared to `CONTRASTIVE2`, a model that fine-tunes on a mixture of Chat and *Chat-Similar News* in one step, we achieve a 0.59 validation BLEU point improvement.

System	Architecture	Domain Adaptation	Dev. BLEU	Test BLEU	Human Score
BASELINE	Standard Transformer	Chat	-	43.4	-
PRIMARY	Standard Transformer	Chat	**58.54**	56.7	79.29
Vanilla Chat	Multi-encoder	Chat	57.52	-	-
CONTRASTIVE1	Multi-encoder	Similar News → Chat	58.31	55.6	-
CONTRASTIVE2	Multi-encoder	Chat + Similar News	57.72	56.4	-
WMT20-CHAT-BEST	-	-	-	**60.1**	**88.21**

Table 1: Agent-side (en→de) performance of submitted systems on the official development and test sets of the WMT20 chat translation task. BASELINE was the best performing model in the WMT19 News task, PRIMARY is our primary submission, and WMT20-CHAT-BEST produced the best Agent-side outputs, according to human evaluation.

Model	# Params	# Samples
PRIMARY	63M	13845
CONTRASTIVE1	83M	(1303, 13845)
CONTRASTIVE2	83M	18624

Table 2: # Trainable parameters and # Training samples per model. Values within tuples indicate the number of training samples available to a corresponding, intermediate model.

4.3.2 Same-time Training of Chat-Similar News: CONTRASTIVE2

This system fine-tunes on the multi-encoder architecture with the combined *Chat-Similar News* and Chat training data in one shot. We see only a 0.2 validation BLEU point improvement here over the multi-encoder fine-tuned only with Chat (Vanilla Chat in Table 1).

5 Evaluation

5.1 Official Evaluation

BLEU scores on the development and test sets, and official human evaluation results are shown in Table 1. The PRIMARY system achieves the best validation and test BLEU. While CONTRASTIVE1 has a slightly higher validation BLEU, it turns out that CONTRASTIVE2 performs better at test time, showing that the same-time training technique may be less prone to overfitting.

5.2 Coherence Evaluation with Hand-crafted Examples

The official evaluation results seem to suggest the context-aware multi-encoder architecture (contrastive systems) is not superior to the standard Transformer which has no access to contextual information. We manually examined the training data, and noted that between two people interacting with each other on the phone or through their computer screens, there are not many indirect pronouns, possibly because there is no associated real-life gesturing necessitating expressions such as "that one" or "those ones". Seemingly, in the provided datasets, the need to be clear over the phone/internet means key words are often repeated for clarity, **especially** on the agent side ("I would like to order a pizza"; **"how can I help you with ordering a pizza"**). Inspired by (Bawden et al., 2018), we carefully evaluate performance of the systems on a hand-crafted dataset consisting of coreference and cohesion test instances. Example instances can be seen in Tables 4 and 5 respectively.

A contributor fluent in both English and German produced two versions of a dataset of 103 source-target pairs[3] based loosely off the provided validation set, following the spirit of (Bawden et al., 2018), in which a current utterance will require the previous utterance in order to make a disambiguating translation in the current. One version has the reference translation set to the correct coreference or cohesion resolution, while the other version can be a potentially correct translation viewing the source sentence in isolation but is incorrect with the additional context. The source side remains unchanged in both versions. We benchmarked each of our submitted models by producing hypotheses using each model given the source sentence, and then computing BLEU scores on the reference from both versions of this dataset.

In Table 3, we show the results of each model against the two versions. We used greedy decoding to generate the hypotheses. We observe that the contrastive multi-encoder systems, though performing worse in BLEU than the single-encoder primary system on the provided validation dataset, actually score higher in the specifically crafted correct coreference/cohesion dataset. By contrast, PRIMARY scores higher for the incorrect coreference/cohesion dataset. Furthermore, the difference in BLEU points between the correct and incorrect

[3] https://github.com/SongChujun/joeynmt/blob/master/chatnmt/coher/manual_coher.json

System	BLEU with Correct Ref.	BLEU with Incorrect Ref.	Diff
PRIMARY	50.44	**49.54**	-0.90
CONTRASTIVE1	50.64	48.84	-1.80
CONTRASTIVE2	**51.39**	49.23	-2.16

Table 3: Agent-side (en→de) performance of submitted systems on our coherence dataset.

Context utterance	Nein. Ich weiss nicht wo **sie** ist. (*No, I do not know where it is.*)
Source utterance	It's 200 meters north of City Center.
Correct reference	**Sie** ist zweihundert Meter nordlich vom Stadtzentrum.
Incorrect reference	**Es** ist zweihundert Meter nordlich vom Stadtzentrum.

Table 4: Example of sentence requiring anaphoric pronoun resolution. A better translation should bias to the correct pronoun based on context as 'sie' and not as 'er' or 'es' (for masculine and neuter nouns respectively).

Context utterance	Ist 30% in Ordnung als Trinkgeld? (*Is 30% alright as tip?*)
Source utterance	Yes, that's more than **generous**.
Correct reference	Ja, das ist mehr als **grosszuegig**.
Incorrect reference	Ja, das ist mehr als **wohlwollend**.

Table 5: Example of sentence requiring lexical disambiguation. Given the context of giving a "tip", a system should bias the translation of "generous" more towards "grosszuegig" (someone is free with money) and away from "wohlwollend" (more in the altruistic, do-gooder sense), which is inappropriate here.

coherence datasets is more significant in the contrastive systems, suggesting that the contrastive models are recovering more of the correct coreference and cohesion, as opposed to retrieving vocabulary words in other areas of the reference.

6 Discussion

Our results largely agree with those of (Voita et al., 2018), chiefly that combining knowledge from a previous "context" sentence can improve the model's ability to improve translation quality when measured against sentences whose translations require anaphora considerations. To accommodate this, we produced one set of sentences which require coreference and cohesion resolution, and one set of sentences that have invalid resolution. We found that each submitted system scored worse on the invalid set compared to the valid set, but the difference was more staggering (Table 3) in the context-aware contrastive systems, lending evidence that these models are able to resolve this type of anaphora.

Our work and submission to the shared task can be viewed with several caveats in mind, which may explain the sub-optimal performance of the contrastive systems compared to the primary system. First, we used hyperparameters consistent with a context-agnostic pretrained model in order to have a fair comparison for evaluation and because these presumably have been well-tuned *for the original model*. It may be the case that different hyperparameters would work better for this particular data and the slightly larger architectures used for the contrastive submissions. It would be worth strategizing with better hyperparameter optimization.

Secondly, we use the provided target sides of the de→en direction to provide context to our en-de data as if it were back-translated. Since both the agent and customer sides of this datasets were actually produced in English (the latter being translated with human-corrected machine translation), these additional utterances are likely higher quality than we would get from back-translating in a real test setting.

7 Conclusions

In this paper, we discussed our methods for training and submitting the outputs of three models for the WMT20 shared task of chat translation. Each system was based off a transformer model pretrained on WMT17 en-de news to provide better fluency. Our best system achieves a test BLEU of 56.7, improving over the provided baseline by more than 13 BLEU points, and less than 4 points behind the best shared task submission. Though we were unable to show that a context-aware model produced better translation quality than the context-agnostic model on the given dataset, our coherence evaluations indicated that it can produce better translations when measured against references needing context for

coreference and cohesion resolution. This was validated both in terms of BLEU and by model scoring of references.

Acknowledgments

This work was partially supported by AWS Cloud Credits provided by an Amazon Machine Learning Research Award. We also thank the Natural Language Processing Laboratory, National Taiwan University[4], where a contributor was previously affiliated, for partially supporting us in computational resources.

References

Dzmitry Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and B. Haddow. 2018. Evaluating discourse phenomena in neural machine translation. *ArXiv*, abs/1711.00513.

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*.

Ge Gao, Bin Xu, David C Hau, Zheng Yao, Dan Cosley, and Susan R Fussell. 2015. Two is better than one: improving multilingual collaboration by giving two machine translation outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 852–863.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

S. Jean, Stanislas Lauly, Orhan Firat, and K. Cho. 2017. Does neural machine translation benefit from larger context? *ArXiv*, abs/1704.05135.

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. Association for Computational Linguistics.

Jae-won Lee and Gil Chang Kim. 1997. A dialogue analysis model with statistical speech act processing for dialogue machine translation. In *Spoken Language Translation*.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. 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. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 679–688.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

[4] http://nlg.csie.ntu.edu.tw/