Improving Document-Level Neural Machine Translation with Domain Adaptation

Sami Ul Haq[1], **Sadaf Abdul Rauf**[2,3], **Arslan Shoukat**[1] **and Noor-e-Hira**[2]
[1] National University of Sciences and Technology, Pakistan
[2] Fatima Jinnah Women University, Pakistan
[3] LIMSI-CNRS, France
{sadaf.abdulrauf,noorehira94}@gmail.com
{sami.ulhaq,arslanshaukat}@ceme.nust.edu.pk

Abstract

Recent studies have shown that translation quality of NMT systems can be improved by providing document-level contextual information. In general sentence-based NMT models are extended to capture contextual information from large-scale document-level corpora which are difficult to acquire. Domain adaptation on the other hand promises adapting components of already developed systems by exploiting limited in-domain data. This paper presents FJWU's system submission at WNGT, we specifically participated in Document level MT task for German-English translation. Our system is based on context-aware Transformer model developed on top of original NMT architecture by integrating contextual information using attention networks. Our experimental results show that providing previous sentences as context significantly improves the BLEU score as compared to a strong NMT baseline. We also studied the impact of domain adaptation on document level translation and were able to improve results by adapting the systems according to the testing domain.

1 Introduction

In past few years, machine translation systems have witnessed remarkable growth due to increasing amount of multilingual information. Neural Machine Translation (NMT) has become one of the powerful and de-facto approaches recognized for its generality and effectiveness (Li et al., 2018). Due to better accuracy of deep neural models, it has quickly achieved state of the art performance in machine translation (Shen et al., 2015).

Standard neural machine translation model works on individual sentences and focuses on short context windows for improving translation quality while ignoring cross-sentence links and dependencies (Xiong et al., 2019). Sentence-by-sentence translation of well-formed documents may generate an incoherent target text which is unable to span the entire document. This largely limits the success of NMT, as document context is totally ignored. Intuitively, to generate coherent translation of source document, machine learning models expect cross-sentence dependencies and linkages. To this end, several models (Voita et al., 2018; Wang et al., 2017; Tu et al., 2018; Maruf and Haffari, 2017; Bawden et al., 2017; Jean et al., 2017) have been proposed for document-wide translation.

Adapting NMT models for context-aware translations has the biggest challenge of limited availability of bilingual document-level corpora. Since, only few resources are available for training, the application domain of NMT may greatly vary from domain of training data. Consequently, the performance of NMT system may quickly degrade as soon as the testing conditions deviate from training conditions.

Domain adaptation has been an active research topic in the field of machine translation to improve translation performance often for low resource settings (Koehn and Schroeder, 2007). The quality of neural machine translation heavily depends upon domain-specificity of test data and the amount of parallel training data. The demand for high quality domain specific MT systems has significantly increased over the years but the bilingual corpora for relevant languages still lack in quantity (Chu and Wang, 2018).

In this work, we aim to demonstrate performance optimization in a particular domain by training document-level models on large out-of domain parallel corpus combined with small in-domain corpus using domain adaptation techniques. Our experiments on German-English data using document-level translation model (Miculicich et al., 2018) an extension of standard NMT Transformer (Vaswani et al., 2017) reveals the importance of contextual

Proceedings of the 4th Workshop on Neural Generation and Translation (WNGT 2020), pages 225–231

www.aclweb.org/anthology/D19-56%2d

information and domain adaptation on translation quality.

By comparing the performance with standard NMT baseline models trained on bilingual data, we show that NMT models exposed to random and actual contextual information are more sensitive to translation quality. We also demonstrate the impact of domain adaptation on translation quality by adapting document-level system to the testing domain.

2 Models

We use two types of models, sentence level and document-level context aware model, both built using Transformer architecture (Vaswani et al., 2017). For our primary submission we train document-level models on parallel corpus with document boundaries. The document level model consists of hierarchical attention encoder and decoder to capture both source and target side contextual information during training and testing.

2.1 Sentence-level Models

Our baseline for sentence-level models is OpenNMT-py (Klein et al., 2017) implementation of the Transformer architecture. To be able to establish comparison with document-level models, strong sentence-level baseline is defined with the same architecture (Vaswani et al., 2017) and training data as used for document-level models. Model configurations and training/evaluation data for sentence-level models are discussed in section 3.2.

2.2 Document-level Models

The motivation behind this research work is to test document-level NMT models on sports domain for WNGT20 shared task. The standard Transformer encoder and decoder are extended to take additional sentences as contextual input (Miculicich et al., 2018). Hierarchical attention networks (Yang et al., 2016) are employed on both sides of the NMT model to capture larger context. HAN encoder and decoder can be used jointly to provide dynamic access for selecting previous sentences or predicting most appropriate words.

3 Experimental Setup

The baseline and document-level models are trained on English-German parallel data of different domains (i.e. news, press and sports) provided by WMT19[1] and WNGT20[2].

Corpus	Split	Sentences	Documents
Europarl v9	train	1.64M	109.9K
	valid	0.19M	12.93K
	test	0.09M	6.46K
Rapid	train	1.31M	42.9K
	valid	0.14M	5.04K
	test	0.07M	2.52K
News Commentary v14	train	0.28M	7.16K
	valid	0.03M	0.84K
	test	0.01M	0.42K
Rotowire	train	3.24K	242
	valid	3.32K	240
	test	3.24K	241

Table 1: Dataset statistics in terms of number of sentence pairs and documents and the corresponding train, test and development split.

3.1 Dataset

Document level models require parallel data with document boundaries for training and testing. Parallel data without document boundaries can not be directly used to train document level models, if it is imperative to use, then artificial document boundaries need to be generated. Our training corpus is also constrained to use only English-German data from the WMT19 shared task.

As mentioned earlier, one of the main constraints in training document-level models is limited availability of document-level corpora. WMT19 provides document split version of Europarl v9, News-Commentary v14 and Rapid corpus. Rotowire dataset, made available by WNGT20 DGT task also contains parallel data with document distinctions. For document-parallel data, we preserve the document boundaries during data filtering and concatenation as to get original documents back after translation. For DGT shared task submission, Rotowire test set is provided by WNGT20. Since, the models are trained on data from multiple domains, we create standard test set by selecting chunk of data from each domain to generate a more fair representation of each domain in standard test set. This was done by selecting multiple documents from a

[1] http://www.statmt.org/wmt19/
[2] https://sites.google.com/view/wngt20/

particular domain based on size of the dataset[3]. All datasets are tokenised using script provided by WNGT organizers[4]. Table 1 summarizes the corpus details.

3.1.1 In-domain data

RotoWire (Wiseman et al., 2017) is sports data consisting of article summaries about NBA basket ball games. RotoWire dataset is available in two formats, json and plain text. Both formats contain identical split for train/development and test sets. We used plain text format that contains separate files according to IDs of documents, each game summary is taken as a separate document.

3.1.2 Out-of-domain data

Major portion of training data includes out-of-domain parallel corpora taken from WMT19. We used English-German set of Rapid, News-commentary and Europarl with document boundaries. Document boundaries of Europarl v9 dataset resulted in very long documents, therefore we decided to redefine the document boundaries while keeping the same order of sentences[5]. For this, we take the average document size of Rotowire training data which gave us 14 sentences per document. After discarding original space split document boundaries from Europarl v9, we add new boundaries to keep a reasonable size of context.

3.2 Model Configuration and Training

Since our baseline and document-level MT systems use OpenNMT-py (Klein et al., 2017) implementation of Transformer model, we used similar configuration parameters are as reported in original transformer paper (Vaswani et al., 2017). Transformer model incorporates 6-hidden layers for encoder and decoder. All the hidden states have dropout of 0.1 and 512 dimensions. Model is trained with 8000 warm-up steps with a learning rate of 0.01. We checkpoint the model every 1000 steps for validation. Batch size is set to 2048 and modes are trained for 50K steps.

As in the original paper (Miculicich et al., 2018), two step process is followed for training the document-level models. In the first step, NMT model is optimized without context-aware HAN. After that, we optimize the parameter's for HAN encoder, decoder and joint model. HAN Transformer models gave best performance for 1-3 previous sentences, we use k=3 previous sentences for both source and target side context.

[3]For sentence based models, we can select sentences randomly but for document-level models entire document is considered for standard test set.

[4]https://github.com/neulab/ie-eval

[5]In the original approach for document-level NMT, they failed to obtain significant improvements when context increases beyond 3 sentences.

4 Experimental Results

We present the results of experimentation from our models on German-English translation in Tables 2, 3 and 4.

4.1 Domain adaptation: Sentence level

In our initial experiments, we investigate the impact of domain adaptation on translation results at sentence-level. Since the NMT models are adapted for sports domain (RotoWire), so following (Hira et al., 2019) we gave more weightage to RotoWire corpus by replicating the corpus twice and thrice to study the impact.

Dataset	English	German	BLEU
rap	29.3M	30.0M	5.87
roto-rap	29.5M	30.2M	9.37
roto-rap-nc	35.2M	35.8M	7.90
roto2-rap-nc	35.4M	36.0M	**16.33**
roto3-rap-nc	35.6M	36.2M	**20.01**

Table 2: Table summarizing corpora size and BLEU scores for Transformer based NMT systems.

The results in Table 2 are reported on NMT model scores for Rotowire ($roto$), Rapid (rap) and News-Commentary (nc) corpus, here $roto$ is the in-domain corpus. Adding only 0.2M of in-domain $roto$ corpus to 30M rap German corpus yields a substantial improvement of around $+4$ BLEU points (Table 2: row 2). On the other hand, addition of 5.6M German nc corpus to previous systems, gives around $+1.5$ points improvement (row 3). This is an obvious demonstration of the positive effect of domain adaptation on translation quality.

We further explore this effect by replicating twice the $roto$ corpus, this gives a big improvement of 8.43 BLEU points on $roto2-rap-nc$ (Table 2: row 4). Replicating $roto$ 3 times, however gives $+3.68$ points improvement from previous system and an overall improvements of $+12.11$ BLEU points from $roto-rap-nc$. Clearly, by adapting to the testing domain by updating the model weights, substantial improvements are achieved.

4.2 Document-level Adaptation with Context Aware Translation Experiments

For document-level models, taking a strong baseline of sentence level models, we achieved remarkable improvements by incorporating context as shown in Table 3. We report the results for 4 corpus combinations. The first column is a combination of four document segmented corpora. Starting with a baseline of 32.18 Bleu points on sentence level Transformer model on $Rotowire + Rapid + News - Commentary + Euro$ corpus, a gain of 3.79 points is achieved (32.18 $\Rightarrow$ 35.97). This is achieved by incorporation of contextual information by HAN encoder. We have used a context of 3 sentences as this was reported to be the best for capturing context by (Miculicich et al., 2018). This shows a superiority of context based models on standard NMT models. With the joint model we get a score of 38.32 BLEU points.

In columns (3, 4) of Table 3, the effect of domain adaptation to test domain is reported on document-level systems. This years test data were the documents in the test folder provided by the organisers. We experimented by building systems by replicating the Rotowire training corpus, twice and thrice in an attempt to enable the translation model to learn parameter closer to the testing domain. We can clearly see that models with replicated in-domain corpus outperformed and achieved better score than previous document-level and sentence-based models. The best score of 43.08 is obtained when Rotowire is replicated thrice i.e. $RotoWire(*3) + Rapid + NC + Euro$.

All of our document-level models preformed better than sentence-level models but most importantly encoder models gave best scores which clearly indicates that source side provides correct contextual information as compared to target side. The highest score is achieved by combining HAN encoder and HAN decoder model for corpus in second and third two columns. Joint model for last column performed poorly, this can be attributed to the fact that HAN decoder is not contributing complementary information to further improve translations. Another reason can be our selection of decoder's context, as due to limited availability of time we only use decoded states of previous sentences for target side context while other configurations (Miculicich et al., 2018) are also available.

4.3 Other Context Integration Experiments

Table 4 presents results from our different context integration experiments. We have been interested to check how much improvement is due to additional contextual information, therefore we created a similar setup (Scherrer et al., 2019) for analysis of context. For this, we create three variants of our train and test set to evaluate context aware systems:

- **Regular context:** The order of sentences in train and test set is kept same as they appear in original documents to evaluate consistent contextual setting.
- **Random context:** The train and test set is shuffled such that the document boundaries now represent inconsistent contextual sentences.
- **No context:** Document boundaries of test set is modified such that one sentence now presents one document, which means no additional context is made available during translations. We are forcing document level model to avoid context by providing single sentence document during testing.

We report Blue score for context integration experiments in Table 4. Document-level models are expected to perform better when contextual information is available. The BLEU score decreases when we move from regular context to random and no context as indicated by row 1 and column 3-5 of Table 4. Context aware models when trained on inconsistent training data, it hardly effects their performance when actual context is random or missing during testing. Model in row 2 is trained on random or inconsistent document level data, column 3-5 represent scores for regular, random and non-contextual test data. Model trained on data with random context is insensitive to context during testing.

5 Related Work

5.1 Context-aware NMT

Improving machine translation systems by developing document-level models for SMT (Garcia et al., 2015; Hardmeier et al., 2013; Gong et al., 2011) and NMT(Maruf and Haffari, 2017; Tu et al., 2018; Voita et al., 2018; Kuang et al., 2017; Wang et al., 2017) has been an important research area. These contributions are briefly discussed in this section.

Models	*BLEU Score* Roto+Rapid+NC+Euro	Roto(*2)+Rapid+NC+Euro	Roto(*3)+Rapid+NC+Euro
NMT Transformer	32.18	36.40	37.08
+ HAN encoder	35.97	39.31	43.08
+ HAN decoder	35.37	39.33	42.76
+ HAN encoder + HAN decoder	38.32	40.82	38.13

Table 3: Table summarizing HAN & NMT Transformer results (Rotowire official test) for adding document context and domain adaptation.

Models	Tokens		BLEU Score		
	EN	DE	Reg	Rand	None
HAN_reg	420M	489M	39.31	39.25	39.08
+HAN_rand	188M	224M	37.76	37.76	37.75

Table 4: BLEU for EN⇒DE translations using Regular, Random and None contextual settings of corpus.

(Miculicich et al., 2018) proposed document-level approach with Hierarchical Attention Network (HAN) to provide contextual information during translation. Two HANs are considered for integrating source and target context in NMT. HAN are believed to provide dynamic access to contextual information as compared to Hierarchical Recurrent Neural Networks (HRNN). However, the approach is restrictive for incorporating large contextual information by only considering a limited number of previous source/target sentences.

Cache-based memory approach is proposed by (Tu et al., 2018) to provide document context during translation. Memory networks keep the representation of a set of words in cache to provide contextual information to NMT in the form of words. However, the stored representations are considered irrespective of sentences in which they occur and do not provide actual context to NMT. Cache based memory models have been used in both SMT (Gong et al., 2011) and NMT to store rich representations of source and target text. (Kuang et al., 2017) use two caches, dynamic cache to capture dynamic context by storing words of translated sentence and topic cache which stores topical words of target side from entire document. Through a gating mechanism, the probability of NMT model and cache based neural model is combined to predict the next word.

Memory network-based approach presented by (Maruf and Haffari, 2017) is used to integrate global source and target context to sentence-based NMT. Keeping the source and target context in memory can be very time consuming and memory inefficient as the sentence pairs in document could be enormous. Another study by (Xiong et al., 2019) is based on deliberation networks to capture the cross-sentence context by improving the translation of baseline NMT system in the second pass. Generation of discourse coherent output is largely dependent upon the performance of the canonical NMT model.

The approach proposed by (Zhang et al., 2018) implemented with document-level context outperforms existing cache based RNN search model. Extending Transformer model has achieved better context awareness and a low computational overhead. (Voita et al., 2018) introduce a context aware NMT model in which they control and analyze the flow of information from the extended context to the translation model. They show that using the previous sentence as context their model is able to implicitly capture anaphora.

Both source and target side contextual information plays important role in document-level translation. Inspired from previous sentence-based context-aware approaches (Voita et al., 2018; Stojanovski and Fraser, 2019; Zhang et al., 2018), we are using extended Transformer model (Miculicich et al., 2018) with ability to use dynamic context for document-level experiments.

5.2 Domain Adaptation

The basic concept behind domain adaptation in NMT is utilizing large amount of available parallel data for training NMT models and adapting these to novel domains with small in-domain data (Freitag and Al-Onaizan, 2016). In the simplest approach, in-domain data can be used to fine-tune models trained on large-scale out-of-domain data. Training NMT models from scratch on combined data can take several weeks and may suffer from performance degradation on in-domain test data.

Fine-tuning is a fast and efficient method to integrate in-domain data, and does not need building systems from scratch. Fine tuning for NMT

(Dakwale and Monz, 2017; Freitag and Al-Onaizan, 2016; Luong and Manning, 2015) is achieved by further training a neural model on in-domain data which is already trained on large general domain training data. Adoption to new domain is achieved by (Sennrich et al., 2015) by using synthetic data through back-translation of target in-domain monolingual text and retraining on combined training corpus by adding new data.

For domain adaptation, we use data augmentation method similar to (Chu et al., 2017) by oversampling small in-domain corpus. This simple data augmentation approach does not require any modification in NMT architecture and forces NMT to pay equal/more attention to in-domain training data.

6 Conclusion

In this study, we present methods to improve document-level neural machine translation. Following recently reported results on the task, our experiments also reiterate the fact that incorporating context in translation helps considerably improve the quality. Taking a strong Transformer based baseline model trained on substantial corpus (a concatenation of four corpora RotoWire, Rapid, Euro and News Commentary), context aware document models result in significant improvement in BLEU points. We have also experimented with the effects of corpus replication to adapt to the domain of test corpus. We find it an effective method to improve translation quality and domain adaptation.

We have submitted results of our best model (HAN encoder) for German-English direction as reported in Table 3 , for official evaluation. Han encoder model with domain adaptation techniques achieved 43.08 BLEU score. We have computed BLEU scores using Moses $multi - blue.perl$ script.

Acknowledgments

This study is funded by Higher Education Commission of Pakistan's project: National Research Program for Universities (NRPU) (5469/Punjab/NRPU/R&D/HEC/2016).

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Praveen Dakwale and Christof Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit*, 117.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Eva Martínez Garcia, Cristina España-Bonet, and Lluís Màrquez. 2015. Document-level machine translation with word vector models. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 59–66.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 909–919. Association for Computational Linguistics.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics); 4-9 August 2013; Sofia, Bulgaria*, pages 193–198. Association for Computational Linguistics.

Noor-e Hira, Sadaf Abdul Rauf, Kiran Kiani, Ammara Zafar, and Raheel Nawaz. 2019. Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 156–163, Florence, Italy. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Modeling coherence for neural machine translation with dynamic and topic caches. *arXiv preprint arXiv:1711.11221*.

Qiang Li, Derek F Wong, Lidia S Chao, Muhua Zhu, Tong Xiao, Jingbo Zhu, and Min Zhang. 2018. Linguistic knowledge-aware neural machine translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(12):2341–2354.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Sameen Maruf and Gholamreza Haffari. 2017. Document context neural machine translation with memory networks. *arXiv preprint arXiv:1711.03688*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level nmt in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.

Dario Stojanovski and Alexander Fraser. 2019. Combining local and document-level context: The lmu munich neural machine translation system at wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 400–406.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *arXiv preprint arXiv:1704.04347*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

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. *arXiv preprint arXiv:1810.03581*.