Document Level NMT of Low-Resource Languages with Backtranslation

Sami Ul Haq[1], **Sadaf Abdul Rauf**[2,3], **Arslan Shoukat**[1] **and Abdullah Saeed**[4]
[1] National University of Sciences and Technology, Pakistan
[2] Fatima Jinnah Women University, Pakistan
[3] LIMSI-CNRS, France
[4] COMSATS University, Pakistan
{sadaf.abdulrauf,abdullahsaeed98}@gmail.com
{sami.ulhaq,arslanshaukat}@ceme.nust.edu.pk

Abstract

This paper describes our system submission to WMT20 shared task on similar language translation. We examined the use of document-level neural machine translation (NMT) systems for low-resource, similar language pair Marathi−Hindi. Our system is an extension of state-of-the-art Transformer architecture with hierarchical attention networks to incorporate contextual information. Since, NMT requires large amount of parallel data which is not available for this task, our approach is focused on utilizing monolingual data with back translation to train our models. Our experiments reveal that document-level NMT can be a reasonable alternative to sentence-level NMT for improving translation quality of low resourced languages even when used with synthetic data.

1 Introduction

With the widespread use of MT systems in commercial and research community, there is an increased attention to train NMT models for direct translation between language pairs other than English Barrault et al. (2019). This is because of the growing need to translate between pairs of similar languages without considering English as pivot language. The task is to overcome the challenge of limited availability of parallel data by exploiting the advantages of similarity between languages when building machine translation models. Similar languages have the advantage of having some magnitude of common information such as lexical and semantic structures. A number of research studies have been published to exploit commonalities when translating text between close language pairs Pourdamghani and Knight (2017); Lakew et al. (2018); Costa-jussà (2017).

This paper describes our system submission at WMT shared Similar Language Translation task[1] which focuses on improving translation quality of similar languages in low-resource setting, the detail of task is provided in Barrault et al. (2019). This year's task includes five pairs of languages from three different language families i.e. Indo-Aryan, Romance and South-Slavic languages; we participated for Hindi-Marathi language pair. Since we are using NMT which requires large bitext, we need to alleviate this specific problem of bitext shortage. Sennrich et al. (2016) introduced an approach to utilize monolingual data using back translation. This requires a machine translation system in opposite direction to generate synthetic parallel corpora from target side monolingual text.

Our work is an attempt to investigate the translation of a similar language pair (Marathi-Hindi) using document-level NMT and back translation. We participated under team name "FJWU_NUST". We submitted one constrained system i.e. we only used the parallel and monolingual data provided by WMT20[2] organizers to train and evaluate our models. We train and evaluate NMT systems in both directions (i.e. HI⇒MR and MR⇒HI) but our submission to similar language shared task comprises of MR⇒HI systems only.

The rest of the paper is structured as follows: In Section 2 we give a brief background of document-level NMT, Section 3 presents utilization of monolingual data, Section 4 and 5 present our experimental setup and results. We conclude the paper in Section 6.

2 Document-Level NMT

Standard NMT works by translating individual sentences and focuses on short context windows while ignoring cross-sentence links and dependencies Xiong et al. (2019). Document-level NMT aims to consider discourse dependencies across sentences

[1] http://www.statmt.org/wmt20/similar.html

[2] http://www.statmt.org/wmt20/.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 442–446
Online, November 19–20, 2020.

to capture document wide context. Most recently, there has been great interest in modelling larger context in standard NMT (Voita et al., 2018; Wang et al., 2017; Tu et al., 2018; Maruf and Haffari, 2017; Bawden et al., 2017; Jean et al., 2017; Chen et al., 2020). Cache based Tu et al. (2018) memory models can be used to hold rich information, can also provide the context of document 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. Kuang et al. (2017) used 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. Miculicich et al. (2018) has proposed to use hierarchical attention network (HAN) Yang et al. (2016) to provide dynamic contextual information to NMT during translation. HANs are used on both sides, encoder and decoder to integrate source and target side context in NMT. In contrast to Recurrent Neural Networks (RNN), HANs provide dynamic access to contextual information during training and evaluation.

Similarly, Maruf and Haffari (2018) used pretrained RNN encoder to attach global source and target context to sentence based NMT. Zhang et al. (2018) has shown that integration of short context (2 sentences) outperforms existing cache based RNNSearch model. Voita et al. (2018) introduce a context aware NMT model with additional multi-head attention component, in which they control and analyze the flow of information from the extended context to the translation model.

Stojanovski and Fraser (2020) studied the use of Transformer based document-level models adoptable to novel (zero-resource) domains. They have shown the implicit domain adaptation of document-level NMT models trained on multi-domain data, is capable of capturing large context. The challenge of translating single sentences efficiently while keeping models insensitive to enlarge and noisy context is addressed by Zheng et al. (2020). To make general purpose context-aware MT, both for short and long sentences, they opt for having independent global and local context integration into sentence based NMT.

3 Utilizing Monolingual Data

Large amounts of monolingual resources are generally available for a multitude of languages. Back translation is considered a well known approach to mitigate the need of large parallel corpora by automatically translating target language monolingual data to source language Sennrich et al. (2016). Back translation requires a MT system in opposite direction, where target side monolingual data is translated into source text to generate synthetic parallel training data. Several techniques exists to utilize monolingual text for improving NMT (Abdul-Rauf et al., 2016; Zhang and Zong, 2016; Currey et al., 2017; Domhan and Hieber, 2017).

Document-level models require parallel data with document boundaries for training and evaluation. As compared to sentence-level systems, data for building robust document-level models is significantly low resourced Liu and Zhang (2020). WMT20 provides document-level distinctions for Europarl v9, New-Commentary v14 and Rapid corpus. Our training data is constrained to have only parallel and monolingual data provided by WMT20 shared task, the statistics of data are given in section 4.1. Since, our system is build in Marathi-Hindi direction, we backtranslated Hindi (News Crawl2008-2019) monolingual data into Marathi to generate bitext. This backtranslated data is than concatenated with parallel data made available by organizers, to train machine translation models.

4 Experimental Setup

For our primary submission we use document-level Miculicich et al. (2018) model, an extension of transformer with additional context attentions. For comparison with sentence-based NMT systems, a strong baseline using OpenNMT-py Klein et al. (2017) is first defined. For true comparison, the architecture and configurations of both the models are kept the same.

4.1 Dataset

Table 1 presents details of training, development and test corpus. We used all the parallel data (HI, MR) provided by WMT20 for similar language translation task. The available parallel data was insufficient to train NMT models, therefore we used monolingual "News Crawl" data for generating synthetic parallel corpus through backtranslation. NMT models are trained on backtranslated bitext combined with existing parallel corpus. Training corpus contains data of multiple domains, a self test set is created by selecting chunk of data from each domain according to size of dataset. Original bitext and backtranslated parallel training data is tokenized with Indic-NLP [3] library, which supports tokenization/de-tokenization of Hindi and Marathi.

Our document-level systems Miculicich et al. (2018) expect document boundaries in text file during training and testing. Available data for this shared task does not contains document boundaries, for this we followed the same approach used by Ul Haq et al. (2020) to generate artificial document boundaries. They have taken average document size from document-level corpora and used the same size to generate document boundaries for parallel data without document distinctions. For train and dev set, instead of splitting on sentences, they considered number of documents. We have used average of two best performing context variables for document size as reported in Table 3 of Miculicich et al. (2018).

Corpus	Sentences	Documents
News	12.3K	4.1K
PmIndia	25.9K	8.6K
IndicWordNet	11.2K	3.7K
NewsCrawl-Monolingual	0.6M	0.2M
Dev	1114	278
Test	1941	485

Table 1: Train, Dev and Test dataset statistics along with document split.

4.2 Model Configurations

As our sentence-level baseline and document-level systems are based on Transformer model, we followed similar configuration parameters for both as reported in original paper Vaswani et al. (2017). 6 hidden layers are incorporated on both encoder and decoder side of Transformer model. All the hidden states have a dropout of 0.1 and 512 dimensions. Transformer model is trained with 8000 warm-up steps with a learning rate of 0.01. We checkpoint the model every 1000 steps for validation. For all the models, batch size is set to 2048 and is trained for 150 epochs.

Two step training process is followed as described by Miculicich et al. (2018). Initially NMT models are optimized without considering contextual information, after that encoder and decoder models are optimized by using context-aware HANs. HAN Transformer models gave best performance for 1-3 previous sentences, we use k=3 previous sentences for both source and target side context.

5 Results

Table 2 shows our results for Hindi−Marathi translations. Our document-level systems for both directions HR⇒MR and MR⇒HI outperformed sentence-level baselines.

BLEU score for WMT, Dev and $Self$ test set is reported in Table 2 for all systems. BLEU score for WMT test data is provided by WMT20 organizers. We have computed BLEU scores using Moses $multi - blue.perl$ script. For submission, we used output of document-level system trained on all data in MR⇒HI direction which gave highest BLEU score (6.79) on WMT test set. Our document-level models are optimized by adding context-aware HANs on encoder side only[4]. With DL−NMT model trained on corpus containing 90% backtranslated data, a gain of 0.63 BLEU points is achieved (6.16 ⇒ 6.79) over sentence-level baseline (row 2).

In last rows (3 and 4) of Table 2, NMT models are build in opposite direction of back-translated data, depicted as $\text{NMT}_{\text{forward}}$ and and $\text{DL−NMT}_{\text{forward}}$. For forward translation models, source side is backtranslated data while target side is original monolingual data used for backtranslation. Similarly, DL−NMT models trained in forward direction of data, achieved batter score over NMT systems. Since, the large portion of training data contains synthetic data, on self test set all models performed better due to over fitting.

System	Direction	$BLEUScore$ Wmt	Dev	Self
NMT	MR⇒HI	6.16	8.08	12.50
+DL−NMT	MR⇒HI	**6.79**	9.31	14.93
+NMT_{fwd}	HI⇒MR	3.29	6.33	16.69
+$\text{DL−NMT}_{\text{fwd}}$	HI⇒MR	3.54	6.28	17.75

Table 2: Table summarizing Document-level NMT (DL-NMT) and NMT Transformer results for different test sets.

[3] https://github.com/anoopkunchukuttan/indic_nlp_library

[4] Due to limited availability of time, HAN for decoder side and HAN joint models were not used for experiments.

6 Summary

This paper presented the "FJWU_NUST" system submitted to the Similar Language Translation task at WMT20. The limited and out-of-domain parallel training data provided by organizers, emerged as a challenging task to train NMT models, whose quality is dependent on large data.

We have utilized monolingual data with back-translation along with available parallel data for training NMT system which incorporated context-aware HANs on encoder side. Our document-level systems outperformed sentence-level NMT systems, even in the absence of document-level corpora. This showed that document-level machine translation can be reasonable alternative of NMT, since it can deliver good quality translation for low-resource languages without requiring document-level parallel data.

7 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

Sadaf Abdul-Rauf, Holger Schwenk, Patrik Lambert, and Mohammad Nawaz. 2016. Empirical use of information retrieval to build synthetic data for smt domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):745–754.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation.

Marta R. Costa-jussà. 2017. Why Catalan-Spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62, Valencia, Spain. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. 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*.

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*.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Siyou Liu and Xiaojun Zhang. 2020. Corpora for document-level neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France. European Language Resources Association.

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

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

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

Nima Pourdamghani and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2020. Addressing zero-resource domains using document-level context in neural machine translation.

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.

Sami Ul Haq, Sadaf Abdul Rauf, Arslan Shoukat, and Noor-e Hira. 2020. Improving document-level neural machine translation with domain adaptation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 225–231, Online. Association for Computational Linguistics.

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*.

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*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Toward making the most of context in neural machine translation.