Monash University's Submissions to the WNGT 2019 Document Translation Task

Sameen Maruf and **Gholamreza Haffari**
Faculty of Information Technology, Monash University, Australia
{firstname.lastname}@monash.edu

Abstract

We describe the work of Monash University for the shared task of Rotowire document translation organised by the 3rd Workshop on Neural Generation and Translation (WNGT 2019). We submitted systems for both directions of the English-German language pair. Our main focus is on employing an established document-level neural machine translation model for this task. We achieve a BLEU score of 39.83 (41.46 BLEU per WNGT evaluation) for En-De and 45.06 (47.39 BLEU per WNGT evaluation) for De-En translation directions on the Rotowire test set. All experiments conducted in the process are also described.

1 Introduction

This paper describes the work of Monash University for the shared task of Rotowire document translation organised by the 3rd Workshop on Neural Generation and Translation (WNGT 2019). Despite the boom of work on document-level machine translation in the past two years, we have witnessed a lack of the application of the proposed approaches to MT shared tasks. Thus, our main focus in this work is on employing an established document-level neural machine translation model for this task.

We first explore a strong sentence-level baseline, trained on large-scale parallel data made available by WMT 2019 for their news translation task.[1] We use this system as the initialisation of the document-level models, first proposed by Maruf et al. (2019), making use of the complete document (both past and future sentences) as the conditioning context when translating a sentence. Given the task of translating Rotowire basketball articles, we leverage the document-delimited data provided by the organisers of WMT 2019 to train the document-level models. Due to resource constraints, we do not use any monolingual data nor any sort of pre-trained embeddings for training the baseline or our document-level models. We ensemble 3 independent runs of all models using two strategies of ensemble decoding. We have conducted experiments for both directions of the English-German language pair. Our submissions achieve a BLEU score of 39.83 (41.46 BLEU per WNGT evaluation) for En→De and 45.06 (47.39 BLEU per WNGT evaluation) for De→En translation directions on the Rotowire test set (Hayashi et al., 2019).

2 Sentence-level Model

As in the original paper, our document-level models are based on the state-of-the-art Transformer architecture (Vaswani et al., 2017). In the remainder of this section, we will describe how we prepare the data to train our sentence-level model and the training setup.

2.1 Data Preparation

To train our sentence-level model, we want to use the maximum allowable high-quality data from the English-German news task in WMT 2019. This would produce a fair baseline for comparing with our document-level models. Upon considering the task of translating basketball-related articles, we have decided to utilise parallel data from Europarl v9, Common Crawl, News Commentary v14 and the Rapid corpus.[2]

Before proceeding to the pre-processing, we remove repetitive sentences[3] from Rapid corpus occurring at the start and end of the documents.

[1] http://www.statmt.org/wmt19/translation-task.html

[2] Given the limited time and resources at our disposal, we did not use the ParaCrawl corpus.

[3] "European Commission - Announcement", "Related Links", "Audiovisual material", etc.

Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019), pages 256–261
Hong Kong, China, November 4, 2019.
www.aclweb.org/anthology/D19-56%2d

Corpus	#Sentence-Pairs
Europarl v9	1.79M
Common Crawl	2.37M
News Commentary v14	0.33M
Rapid	1.46M
Rotowire	3247

Table 1: Sentence-parallel training corpora statistics.

From all corpora, we also remove sentences with length greater than 75 tokens after tokenisation.[4,5] Table 1 summarises the number of sentences of each corpus in the pre-processed sentence-parallel dataset. We further apply truecasing using Moses (Koehn et al., 2007) followed by joint byte-pair encoding (BPE) with 50K merge operations (Sennrich et al., 2016).

2.2 Model and Training

We use the DyNet toolkit (Neubig et al., 2017) for all of our experiments; the implementation of the sentence-level system is in DyNet, namely *Transformer-DyNet*.[6] Our experiments are run on a single V100 GPU, so we use a rather small mini-batch size of 900 tokens. Furthermore, we have filtered sentences with length greater than 85 tokens to fit the computation graph in GPU memory. The hyper-parameter settings are the same as in the Transformer-base model except the number of layers which is set to 4. We also employ all four types of dropouts as in the original Transformer with a rate of 0.1. We use the default Adam optimiser (Kingma and Ba, 2014) with an initial learning rate of 0.0001 and employ early stopping.

3 Document-level Models

The motivation behind our participation in the shared task is to test the document-level MT models (Maruf et al., 2019) on real world tasks. Here we will briefly describe the models, data preprocessing and training/decoding setup.

3.1 Model Description

There are two ways to incorporate context into the sentence-level model: (i) integrate monolingual context into the encoder, or (ii) integrate the bilingual context into the decoder. For both, the document-level context representation is combined with the deep representation of either the source or target word (output from the last layer of the Transformer) using a gating mechanism (Tu et al., 2018).

The document-level context representation is itself computed in two ways: (i) a single-level flat attention over all sentences in the document-context, or (ii) a hierarchical attention which has the ability to identify the key sentences in the document-context and then attend to the key words within those sentences. For the former, we use a soft attention over the sentences, while for the latter we use a sparse attention over the sentences and a soft or sparse attention over the words in the sentences. For more details of how the document-level context representations are computed, we refer the reader to the original paper (Maruf et al., 2019).

3.2 Data Preparation

Since our document-level model requires document boundaries, we are unable to use the sentence-parallel corpus as is. Out of all the WMT19 corpora, News Commentary and Rapid corpus are document delimited.[7] The Rotowire dataset also has document boundaries provided. Thus, we decided to combine these three corpora to construct the document-parallel training corpus.[8] Furthermore, we remove documents from this document-parallel corpus which have sentence lengths greater than 75 (after tokenisation). We also remove short documents with number of sentences less than 5, and long documents with number of sentences greater than 145. The filtered document-parallel corpus comprises 49K documents making up approximately 1.36M sentence-pairs. The corpus is then truecased using the truecaser model trained on the sentence-level corpus followed by BPE. Since we filtered out sentences with lengths greater than 85 while training the baseline, we also filter those from the document-level corpus. However, removing individual sentences from documents could introduce noise in the training process, hence we remove entire such documents. Finally, we use 48K doc-

[4] The threshold was carefully chosen based upon the maximum length of sentences in the Rotowire training set so that we do not remove any of the sentences from the shared task corpus.

[5] Tokenisation script was provided by WNGT organisers.

[6] https://github.com/duyvuleo/Transformer-DyNet

[7] Europarl v9 also had document boundaries but these resulted in very long documents and thus we decided against using it for the document-level training.

[8] The training corpus used for training the document-level models is a subset of the training corpus used for training the baseline sentence-level model.

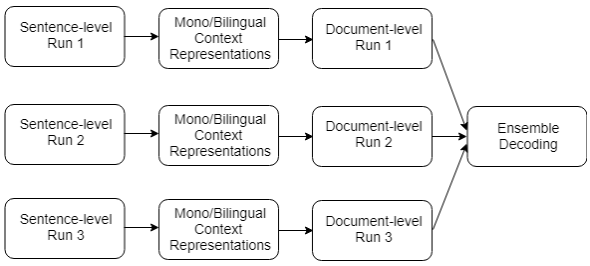


Figure 1: Ensemble decoding.

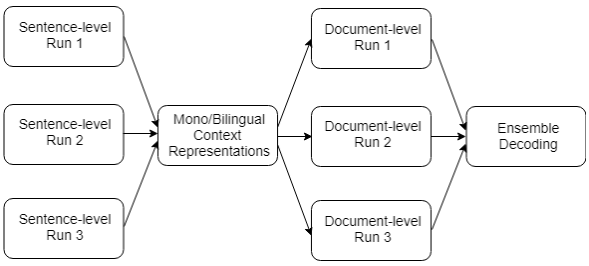


Figure 2: Ensemble-Avg decoding.

uments comprising 1.21M sentences for training our document-level models.

3.3 Training

We use a stage-wise method to train the variants of the document-context NMT model. We pre-train the sentence-level model described in the previous section and then use it to compute the monolingual and the bilingual context representations. These are then used to compute the document-level context representation in our models. The pre-trained sentence-level model is also used to initialise our document-level model and is further fine-tuned alongwith training the document-level model in the second stage. Here we also employ all four types of dropouts[9] but with a higher rate of 0.2 to avoid overfitting. Since the documents on average have lengths much shorter than 900 tokens, we update the model parameters after processing 5 documents instead of a single document.

3.4 Decoding

For the models using the source monolingual context, we do an initial pass over the source documents to compute the initial context representations, which are then used by the document-level model to perform a greedy decoding to obtain the target translations. For the models using the bilingual context, we need an initial bilingual context which is computed by generating initial translations from the sentence-level NMT model. This is followed by a second pass of decoding, where the translation for each sentence is updated using the document-context NMT model while fixing the translations of the other sentences. This is what we refer to as two-pass iterative decoding (Maruf and Haffari, 2018). It should also be mentioned that since decoding is a computationally expensive process, we perform greedy decoding in both passes.

[9] *input dropout* - dropout applied to the sum of token embeddings and position encodings, *residual dropout* - dropout applied to the output of each sublayer before adding to the sublayer input, *relu dropout* - dropout applied to the inner layer output after ReLU activation in each feed-forward sublayer, and *attentiondropout* - dropout applied to attention weight in each attention sublayer

4 Experimental Evaluation

4.1 Setup

We have 3 independent runs of the sentence-level Transformer architecture. For each of these runs, we train the variants of the document-level models: (i) the flat attention over sentences in the context, and (ii) the hierarchical attention with sparse attention over sentences and soft/sparse attention over words in the sentences, using the two types of context. Results are reported on the Rotowire test set for the single best model obtained through early stopping on the Rotowire development set.

We also decode the test set with an ensemble of the systems for the 3 independent runs. This is done in two ways:

- *Ensemble*. This is the traditional way of ensembling where the different models are combined by averaging the target probability distributions when computing the softmax (Figure 1).
- *Ensemble-Avg*. Apart from combining the probability distributions at the softmax level, we also average the context representations from each run, i.e., we use the same initial context representations for the different runs of a document-level model (Figure 2).

For evaluation, BLEU (Papineni et al., 2002) is reported on the detruecased translations (with original tokenisation) and is calculated using the MultEval toolkit (Clark et al., 2011).

4.2 English→German

It can be seen from Table 2 that for all runs, the document-level models outperform the sentence-level baseline trained with 4 times the data. The

System	Transformer	Integration into Encoder			Integration into Decoder		
		Attention *soft*	H-Attention *sparse-soft*	H-Attention *sparse-sparse*	Attention *soft*	H-Attention *sparse-soft*	H-Attention *sparse-sparse*
Run 1	34.70	37.93	**38.28**	37.07	38.23	38.13	**38.43**
Run 2	34.45	38.40	**38.72**	38.27	37.42	38.20	**39.02**
Run 3	33.15	37.43	**38.64**	38.25	38.64	**38.65**	37.97
Ensemble	36.10	39.36	**39.83**	39.28	39.33	39.51	**39.71**
Ensemble-Avg	36.10	39.25	**39.79**	39.22	39.42	39.54	**39.71**

Table 2: BLEU scores for the Transformer vs. variants of our document-level NMT model for English→German. **bold**: Best performance.

System	Transformer	Integration into Encoder			Integration into Decoder		
		Attention *soft*	H-Attention *sparse-soft*	H-Attention *sparse-sparse*	Attention *soft*	H-Attention *sparse-soft*	H-Attention *sparse-sparse*
Run 1	37.75	42.58	**43.47**	42.58	**44.42**	44.30	42.96
Run 2	37.86	43.27	42.37	**43.81**	43.47	43.42	**44.05**
Run 3	37.35	43.75	44.08	**44.11**	43.53	**44.16**	43.81
Ensemble	39.33	44.23	44.52	**44.56**	44.94	**45.06**	44.66
Ensemble-Avg	39.33	43.66	43.85	**43.96**	44.83	**44.99**	44.62

Table 3: BLEU scores for the Transformer vs. variants of our document-level NMT model for German→English. **bold**: Best performance.

hierarchical attention model with soft attention over words is the best when using monolingual context (atleast +3.58 BLEU for all runs), while the hierarchical attention model with sparse attention over the words is the best in majority cases when using the bilingual context (atleast +3.73 BLEU for all runs). Among the two types of context, the bilingual context yields better BLEU scores in majority cases.

For traditional ensemble decoding, we get upto +3.73 BLEU improvement for our best hierarchical attention model over the sentence-level model. For ensemble-avg decoding, we see improvements almost equivalent to ensemble decoding. When it comes to speed, there is negligible difference between the two approaches.

4.3 German→English

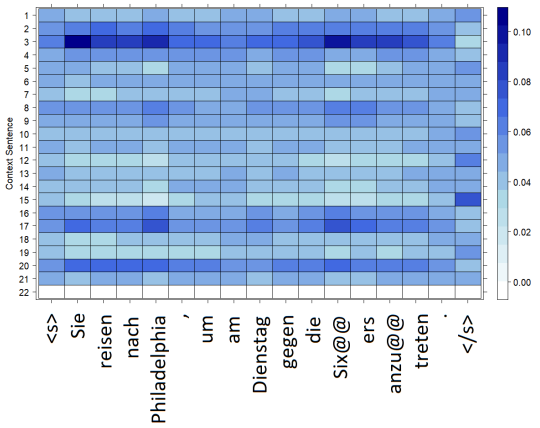
From Table 3, we see that the document-level models again outperform the sentence-level Transformer baseline for all runs by a wider margin than for English→German. For the monolingual context, the hierarchical attention model with sparse attention over words is the best in majority cases (atleast +5.95 BLEU), while for the bilingual context, there does not seem to be a clear winner (atleast +6.19 BLEU). Again, using the bilingual context yields better performance than using monolingual context in terms of BLEU.

For ensemble decoding, we get upto +5.73 BLEU improvement for our best hierarchical attention model when using the bilingual context over the sentence-level baseline. However, for the ensemble-avg decoding, we see the performance decrease in comparison to simple ensemble counterparts when using the monolingual context. The context representations that we averaged for the ensemble-avg decoding were coming from independent models (not checkpoints from the same model) and we believe this to be the reason we observe either deteriorating performance or no improvements for the ensemble-avg decoding in comparison to the ensemble decoding.

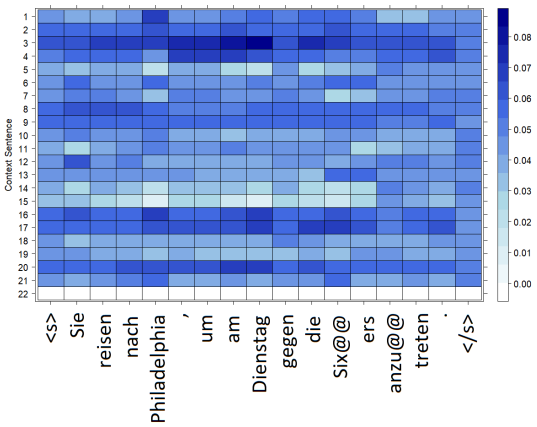
4.4 Analysis

Figure 3 illustrates the attention matrices[10] for an example German sentence as inferred by the flat and hierarchical attention models. The sentence-level attention component of the hierarchical attention model (Figure 3b) appears to be more distributed than its counterpart in the flat attention model (Figure 3a). For the word 'Sie' in the Ger-

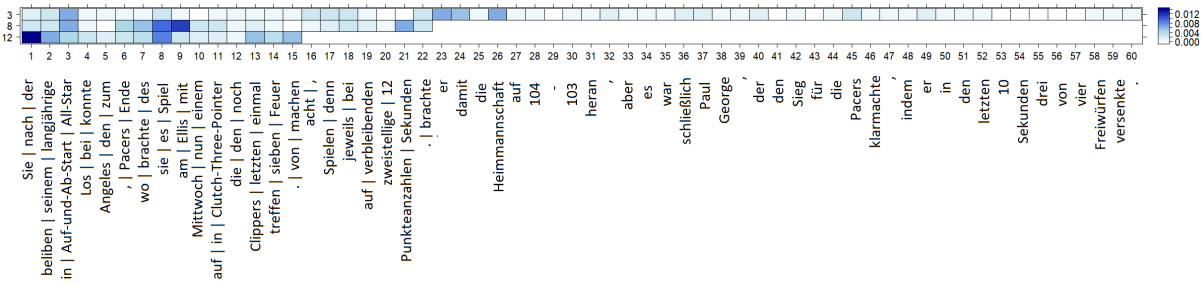
[10]For this analysis, the attention weights are an average over heads of per-head attention weights.



(a) Flat Attention over Sentences



(b) Attention over sentences for the Hierarchical Attention



(c) Scaled Attention over words for the Hierarchical Attention

Figure 3: Attention maps for the Flat Attention and Hierarchical Attention (sparse-sparse) models for a source sentence and source-side context (all in German). The current German sentence (position 22) has been masked.

man sentence, Figure 3c shows the scaled word-level attention map (scaled with the sentence-level attention weights) for the top three sentences, as observed in Figure 3b. *Sie* is an ambiguous pronoun in German and can be translated to *she, they* (sie) and even *you* in the polite form (Sie). It is even more ambiguous when used at the start of the sentence since the capitalisation removes this distinction. It can be seen from Figure 3c that the words given the highest attention weights while encoding this word are mostly other mentions of the same pronoun (Sie, sie). It should also be mentioned that in the 12-th sentence, both occurrences of the pronoun 'sie' also translate to 'they' as in the current sentence.

4.5 Submissions

We have submitted our best ensemble models, one for each translation direction, as reported in Tables 2 and 3, for the official evaluation. As mentioned before, we computed BLEU scores via MultEval toolkit on tokenised and cased Rotowire test set. Table 4 shows the scores as provided by the WNGT organisers. Surprisingly, the scores have increased further. We have been interested in exploring the effectiveness of NMT under constrained resource conditions, i.e., without back-translation on large monolingual data and pre-trained contextualised embeddings. We believe these enhancements could further improve upon the reported results.

Lang. Pair	Our Scores	WNGT Scores
En→De	39.83	41.46
De→En	45.06	47.39

Table 4: BLEU scores for submitted systems.

Acknowledgments

The authors are grateful to the anonymous reviewers for their helpful comments and to Philip Arthur for discussion. This work was supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) (www.massive.org.au). G. H. is supported by a Google Faculty Research Award.

References

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Short Papers)*, pages 176–181. Association for Computational Linguistics.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Constas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. Findings of the third workshop on neural generation and translation. In *Proceedings of the Third Workshop on Neural Generation and Translation*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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 ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. 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*, pages 1715–1725.

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 (TACL)*, 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 30*, pages 5998–6008. Curran Associates, Inc.