Document Context Neural Machine Translation with Memory Networks

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

Abstract

We present a document-level neural machine translation model which takes both source and target document context into account using memory networks. We model the problem as a structured prediction problem with interdependencies among the observed and hidden variables, i.e., the source sentences and their unobserved target translations in the document. The resulting structured prediction problem is tackled with a neural translation model equipped with two memory components, one each for the source and target side, to capture the documental interdependencies. We train the model end-to-end, and propose an iterative decoding algorithm based on block coordinate descent. Experimental results of English translations from French, German, and Estonian documents show that our model is effective in exploiting both source and target document context, and statistically significantly outperforms the previous work in terms of BLEU and METEOR.

1 Introduction

Neural machine translation (NMT) has proven to be powerful (Sutskever et al., 2014; Bahdanau et al., 2015). It is on-par, and in some cases, even surpasses the traditional statistical MT (Luong et al., 2015) while enjoying more flexibility and significantly less manual effort for feature engineering. Despite their flexibility, most neural MT models translate sentences independently. Discourse phenomenon such as pronominal anaphora and lexical consistency, may depend on long-range dependency going farther than a few previous sentences, are neglected in sentence-based translation (Bawden et al., 2017).

There are only a handful of attempts to document-wide machine translation in statistical and neural MT camps. Hardmeier and Federico (2010); Gong et al. (2011); Garcia et al. (2014) propose document translation models based on statistical MT but are restrictive in the way they incorporate the document-level information and fail to gain significant improvements. More recently, there have been a few attempts to incorporate source side context into neural MT (Jean et al., 2017; Wang et al., 2017; Bawden et al., 2017); however, these works only consider a very local context including a few previous source/target sentences, ignoring the global source and target documental contexts. The latter two report deteriorated performance when using the target-side context.

In this paper, we present a document-level machine translation model which combines sentence-based NMT (Bahdanau et al., 2015) with memory networks (Sukhbaatar et al., 2015). We capture the global source and target document context with two memory components, one each for the source and target side, and incorporate it into the sentence-based NMT by changing the decoder to condition on it as the sentence translation is generated. We conduct experiments on three language pairs: French-English, German-English and Estonian-English. The experimental results and analysis demonstrate that our model is effective in exploiting both source and target document context, and statistically significantly outperforms the previous work in terms of BLEU and METEOR.

2 Background

2.1 Neural Machine Translation (NMT)

Our document NMT model is grounded on sentence-based NMT model (Bahdanau et al.,

2015) which contains an encoder to *read* the source sentence as well as an attentional decoder to *generate* the target translation.

Encoder It is a bidirectional RNN consisting of two RNNs running in opposite directions over the source sentence:

$$\overrightarrow{\boldsymbol{h}_i} = \overrightarrow{\text{RNN}}(\overrightarrow{\boldsymbol{h}}_{i-1}, \boldsymbol{E}_S[x_i]), \overleftarrow{\boldsymbol{h}}_i = \overleftarrow{\text{RNN}}(\overleftarrow{\boldsymbol{h}}_{i+1}, \boldsymbol{E}_S[x_i])$$

where $\boldsymbol{E}_S[x_i]$ is embedding of the word x_i from the embedding table $\boldsymbol{E}_S$ of the source language, and $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$ are the hidden states of the forward and backward RNNs which can be based on the LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014) units. Each word in the source sentence is then represented by the concatenation of the corresponding bidirectional hidden states, $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i]$.

Decoder The generation of each word y_j is conditioned on all of the previously generated words $\boldsymbol{y}_{<j}$ via the state of the RNN decoder $\boldsymbol{s}_j$, and the source sentence via a *dynamic* context vector $\boldsymbol{c}_j$:

$$\begin{aligned} y_j &\sim \text{softmax}(\boldsymbol{W}_y \cdot \boldsymbol{r}_j + \boldsymbol{b}_r) \\ \boldsymbol{r}_j &= \tanh(\boldsymbol{s}_j + \boldsymbol{W}_{rc} \cdot \boldsymbol{c}_j + \boldsymbol{W}_{rj} \cdot \boldsymbol{E}_T[y_{j-1}]) \\ \boldsymbol{s}_j &= \tanh(\boldsymbol{W}_s \cdot \boldsymbol{s}_{j-1} + \boldsymbol{W}_{sj} \cdot \boldsymbol{E}_T[y_{j-1}] + \boldsymbol{W}_{sc} \cdot \boldsymbol{c}_j) \end{aligned}$$

where $\boldsymbol{E}_T[y_j]$ is embedding of the word y_j from the embedding table $\boldsymbol{E}_T$ of the target language, and $\boldsymbol{W}$ matrices and $\boldsymbol{b}_r$ vector are the parameters. The dynamic context vector $\boldsymbol{c}_j$ is computed via $\boldsymbol{c}_j = \sum_i \alpha_{ji} \boldsymbol{h}_i$, where

$$\begin{aligned} \boldsymbol{\alpha}_j &= \text{softmax}(\boldsymbol{a}_j) \\ a_{ji} &= \boldsymbol{v} \cdot \tanh(\boldsymbol{W}_{ae} \cdot \boldsymbol{h}_i + \mathbf{W}_{at} \cdot \boldsymbol{s}_{j-1}) \end{aligned}$$

This is known as the *attention* mechanism which dynamically attends to relevant parts of the source necessary for generating the next target word.

2.2 Memory Networks (MemNets)

Memory Networks (Weston et al., 2015) are a class of neural models that use external memories to perform inference based on long-range dependencies. A memory is a collection of vectors $\boldsymbol{M} = \{\boldsymbol{m}_1, .., \boldsymbol{m}_K\}$ constituting the memory cells, where each cell $\boldsymbol{m}_k$ may potentially correspond to a discrete object $\boldsymbol{x}_k$. The memory is equipped with a *read* and optionally a *write* operation. Given a query vector $\boldsymbol{q}$, the output vector generated by reading from the memory is $\sum_{i=1}^{|M|} p_i \boldsymbol{m}_i$, where p_i represents the relevance of the query to the i-th memory cell $\boldsymbol{p} = \text{softmax}(\boldsymbol{q}^T \cdot \boldsymbol{M})$. For the rest of the paper, we denote the read operation by $\text{MemNet}(\boldsymbol{M}, \boldsymbol{q})$.

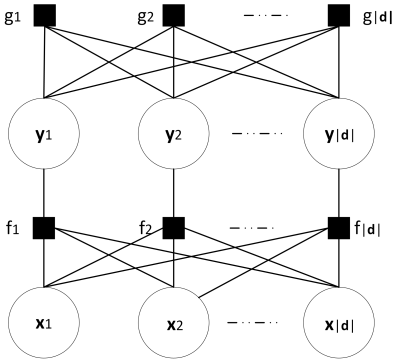


Figure 1: Factor graph for document-level MT

3 Document NMT as Structured Prediction

We formulate document-wide machine translation as a *structured* prediction problem. Given a set of sentences $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|\boldsymbol{d}|}\}$ in a source document $\boldsymbol{d}$, we are interested in generating the collection of their translations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{|\boldsymbol{d}|}\}$ taking into account *interdependencies* among them imposed by the document. We achieve this by the factor graph in Figure 1 to model the probability of the target document given the source document. Our model has two types of factors:

- $f_{\boldsymbol{\theta}}(\boldsymbol{y}_t; \boldsymbol{x}_t, \boldsymbol{x}_{-t})$ to capture the interdependencies between the translation $\boldsymbol{y}_t$, the corresponding source sentence $\boldsymbol{x}_t$ and all the other sentences in the source document $\boldsymbol{x}_{-t}$, and
- $g_{\boldsymbol{\theta}}(\boldsymbol{y}_t; \boldsymbol{y}_{-t})$ to capture the interdependencies between the translation $\boldsymbol{y}_t$ and all the other translations in the document $\boldsymbol{y}_{-t}$.

Hence, the probability of a document translation given the source document is

$$P(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{|\boldsymbol{d}|} | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|\boldsymbol{d}|}) \propto \exp\Big(\sum_t f_{\boldsymbol{\theta}}(\boldsymbol{y}_t; \boldsymbol{x}_t, \boldsymbol{x}_{-t}) + g_{\boldsymbol{\theta}}(\boldsymbol{y}_t; \boldsymbol{y}_{-t})\Big).$$

The factors $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\theta}}$ are realised by neural architectures whose parameters are collectively denoted by $\boldsymbol{\theta}$.

Training It is challenging to train the model parameters by maximising the (regularised) likelihood since computing the partition function is hard. This is due to the enormity of factors

$g_{\boldsymbol{\theta}}(\boldsymbol{y}_t; \boldsymbol{y}_{-t})$ over a large number of translation variables $\boldsymbol{y}_t$'s (i.e., the number of sentences in the document) as well as their unbounded domain (i.e., all sentences in the target language). Thus, we resort to maximising the *pseudo-likelihood* (Besag, 1975) for training the parameters:

$$\arg\max_{\boldsymbol{\theta}} \prod_{\boldsymbol{d}\in\mathcal{D}} \prod_{t=1}^{|\boldsymbol{d}|} P_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t}) \quad (1)$$

where $\mathcal{D}$ is the set of bilingual training documents, and $|\boldsymbol{d}|$ denotes the number of (bilingual) sentences in the document $\boldsymbol{d} = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}_{t=1}^{|\boldsymbol{d}|}$. We directly model the document-conditioned NMT model $P_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t})$ using a neural architecture which subsumes both the $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\theta}}$ factors (covered in the next section).

Decoding To generate the best translation for a document according to our model, we need to solve the following optimisation problem:

$$\arg\max_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{|\boldsymbol{d}|}} \prod_{t=1}^{|\boldsymbol{d}|} P_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t})$$

which is hard (due to similar reasons as mentioned earlier). We hence resort to a block coordinate descent optimisation algorithm. More specifically, we initialise the translation of each sentence using the base neural MT model $P(\boldsymbol{y}_t|\boldsymbol{x}_t)$. We then repeatedly visit each sentence in the document, and update its translation using our document-context dependent NMT model $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t})$ while the translations of other sentences are kept fixed.

4 Context Dependent NMT with MemNets

We augment the sentence-level attentional NMT model by incorporating the document context (both source and target) using memory networks when generating the translation of a sentence, as shown in Figure 2.

Our model generates the target translation word-by-word from left to right, similar to the vanilla attentional neural translation model. However, it conditions the generation of a target word not only on the previously generated words and the current source sentence (as in the vanilla NMT model), but also on all the other source sentences of the document and their translations. That is, the generation process is as follows:

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t}) = \prod_{j=1}^{|\boldsymbol{y}_t|} P_{\boldsymbol{\theta}}(y_{t,j}|\boldsymbol{y}_{t,<j}, \boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t}) \quad (2)$$

where $y_{t,j}$ is the j-th word of the t-th target sentence, $\boldsymbol{y}_{t,<j}$ are the previously generated words, and $\boldsymbol{x}_{-t}$ and $\boldsymbol{y}_{-t}$ are as introduced previously.

Our model represents the source and target document contexts as external memories, and *attends* to relevant parts of these external memories when generating the translation of a sentence. Let $\boldsymbol{M}[\boldsymbol{x}_{-t}]$ and $\boldsymbol{M}[\boldsymbol{y}_{-t}]$ denote external memories representing the source and target document context, respectively. These contain memory cells corresponding to all sentences in the document except the t-th sentence (described shortly). Let $\boldsymbol{h}_t$ and $\boldsymbol{s}_t$ be representations of the t-th source sentence and its current translation, from the encoder and decoder respectively. We make use of $\boldsymbol{h}_t$ as the query to get the relevant *context* from the source external memory:

$$\boldsymbol{c}_t^{src} = \text{MemNet}(\boldsymbol{M}[\boldsymbol{x}_{-t}], \boldsymbol{h}_t)$$

Furthermore, for the t-th sentence, we get the relevant information from the target context:

$$\boldsymbol{c}_t^{trg} = \text{MemNet}(\boldsymbol{M}[\boldsymbol{y}_{-t}], \boldsymbol{s}_t + \boldsymbol{W}_{at} \cdot \boldsymbol{h}_t)$$

where the query consists of the representation of the translation $\boldsymbol{s}_t$ from the decoder endowed with that of the source sentence $\boldsymbol{h}_t$ from the encoder to make the query robust to potential noises in the current translation and circumvent error propagation, and $\boldsymbol{W}_{at}$ projects the source representation into the hidden state space.

Now that we have representations of the relevant source and target document contexts, Eq. 2 can be re-written as:

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{y}_{-t}, \boldsymbol{x}_{-t}) = \prod_{j=1}^{|\boldsymbol{y}_t|} P_{\boldsymbol{\theta}}(y_{t,j}|\boldsymbol{y}_{t,<j}, \boldsymbol{x}_t, \boldsymbol{c}_t^{trg}, \boldsymbol{c}_t^{src}) \quad (3)$$

More specifically, the memory contexts $\boldsymbol{c}_t^{src}$ and $\boldsymbol{c}_t^{trg}$ are incorporated into the NMT decoder as:

- **Memory-to-Context** in which the memory contexts are incorporated when computing the next decoder hidden state:

$$\mathbf{s}_{t,j} = \tanh(\boldsymbol{W}_s \cdot \mathbf{s}_{t,j-1} + \boldsymbol{W}_{sj} \cdot \boldsymbol{E}_T[y_{t,j}] + \boldsymbol{W}_{sc} \cdot \boldsymbol{c}_{t,j} + \boldsymbol{W}_{sm} \cdot \boldsymbol{c}_t^{src} + \boldsymbol{W}_{st} \cdot \boldsymbol{c}_t^{trg})$$

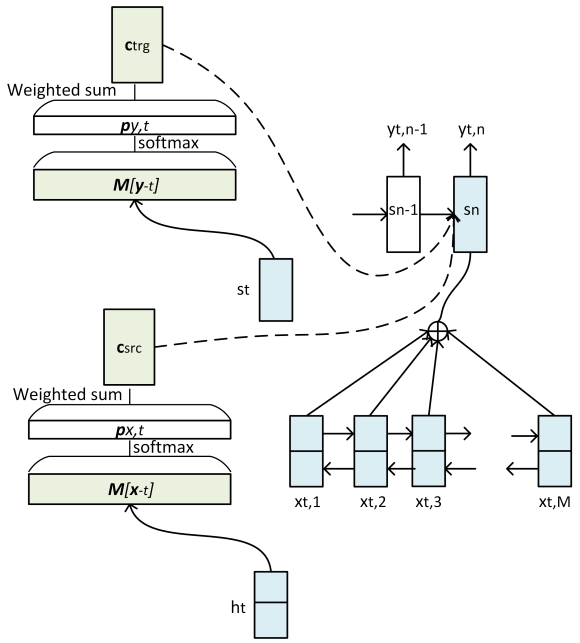


Figure 2: Our Memory-to-Context document-NMT model consisting of sentence-based NMT model with source and target external memories.

- **Memory-to-Output** in which the memory contexts are incorporated in the output layer:

$$y_{t,j} \sim \text{softmax}(\boldsymbol{W}_y \cdot \mathbf{r}_{t,j} + \boldsymbol{W}_{ym} \cdot \boldsymbol{c}_t^{src} + \boldsymbol{W}_{yt} \cdot \boldsymbol{c}_t^{trg} + \mathbf{b}_r)$$

where $\boldsymbol{W}_{sm}$, $\boldsymbol{W}_{st}$, $\boldsymbol{W}_{ym}$, and $\boldsymbol{W}_{yt}$ are the new parameter matrices. We use only the source, only the target, or both external memories as the additional conditioning contexts. Furthermore, we use either the Memory-to-Context or Memory-to-Output architectures for incorporating the document contexts. In the experiments, we will explore these different options to investigate the most effective combination. We now turn our attention to the construction of the external memories for the source and target sides of a document.

The Source Memory We make use of a hierarchical 2-level RNN architecture to construct the external memory of the source document. More specifically, we pass each sentence of the document through a sentence-level bidirectional RNN to get the representation of the sentence (by concatenating the last hidden states of the forward and backward RNNs). We then pass the sentence representations through a document-level bidirectional RNN to propagate sentences' information across the document. We take the hidden states of the document-level bidirectional RNNs as the memory cells of the source external memory.

The source external memory is built once for each minibatch, and does not change throughout the document translation. To be able to fit the computational graph of the document NMT model within GPU memory limits, we pre-train the sentence-level bidirectional RNN using the language modelling training objective. However, the document-level bidirectional RNN is trained together with other parameters of the document NMT model by back-propagating the document translation training objective.

The Target Memory The memory cells of the target external memory represent the current translations of the document. Recall from the previous section that we use coordinate descent iteratively to update these translations. Let $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{|\boldsymbol{d}|}\}$ be the current translations, and let $\{\boldsymbol{s}_{|\boldsymbol{y}_1|}, \ldots, \boldsymbol{s}_{|\boldsymbol{y}_{|\boldsymbol{d}|}|}\}$ be the last states of the decoder when these translations were generated. We use these last decoder states as the cells of the external target memory. We could make use of hierarchical sentence-document RNNs to transform the document translations into memory cells (similar to what we do for the source memory); however, it would have been computationally expensive and may have resulted in error propagation. We will show in the experiments that our efficient target memory construction is indeed effective.

5 Experiments and Analysis

Datasets. We conducted experiments on three language pairs: French-English, German-English and Estonian-English. Table 1 shows the statistics of the datasets used in our experiments. The French-English dataset is based on the TED Talks corpus[1] (Cettolo et al., 2012) where each talk is considered a document. The Estonian-English data comes from the Europarl v7 corpus[2] (Koehn, 2005). Following Smith et al. (2013), we split the speeches based on the `SPEAKER` tag and treat them as documents. The French-English and Estonian-English corpora were randomly split into train/dev/test sets. For German-English, we use the News Commentary v9 corpus[3] for training, `news-dev2009` for development,

[1] https://wit3.fbk.eu/

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

[3] http://statmt.org/wmt14/news-commentary-v9-by-document.tgz

	# docs	# sents	doc len	src/tgt vocab
Fr-En	10/1.2/1.5	123/15/19	123/128/124	25.1/21
Et-En	150/10/18	209/14/25	14/14/14	48.6/24.9
De-En	49/.9/1.1/1.6	191/2/3/3	39/23/27/19	45.1/34.7

Table 1: Training/dev/test corpora statistics: number of documents ($\times$100) and sentences ($\times$1000), average document length (in sentences) and source/target vocabulary size ($\times$1000). For De-En, we report statistics of the two test sets `news-test2011` and `news-test2016`.

and `news-test2011` and `news-test2016` as the test sets. The news-commentary corpus has document boundaries already provided.

We pre-processed all corpora to remove very short documents and those with missing translations. Out-of-vocabulary and rare words (frequency less than 5) are replaced by the `<UNK>` token, following Cohn et al. (2016).[4]

Evaluation Measures We use BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores to measure the quality of the generated translations. We use bootstrap resampling (Clark et al., 2011) to measure statistical significance, $p < 0.05$, comparing to the baselines.

Implementation and Hyperparameters We implement our document-level neural machine translation model in C++ using the DyNet library (Neubig et al., 2017), on top of the basic sentence-level NMT implementation in `mantis` (Cohn et al., 2016). For the source memory, the sentence and document-level bidirectional RNNs use LSTM and GRU units, respectively. The translation model uses GRU units for the bidirectional RNN encoder and the 2-layer RNN decoder. GRUs are used instead of LSTMs to reduce the number of parameters in the main model. The RNN hidden dimensions and word embedding sizes are set to 512 in the translation and memory components, and the alignment dimension is set to 256 in the translation model.

Training We use a stage-wise method to train the variants of our document context NMT model. Firstly, we pre-train the Memory-to-Context/Memory-to-Output models, setting their *readings* from the source and target memories to the zero vector. This effectively learns parameters associated with the underlying sentence-based NMT model, which is then used as initialisation when training *all* parameters in the second stage (including the ones from the first stage). For the first stage, we make use of stochastic gradient descent (SGD)[5] with initial learning rate of 0.1 and a decay factor of 0.5 after the fourth epoch for a total of ten epochs. The convergence occurs in 6-8 epochs. For the second stage, we use SGD with an initial learning rate of 0.08 and a decay factor of 0.9 after the first epoch for a total of 15 epochs[6]. The best model is picked based on the dev-set perplexity. To avoid overfitting, we employ dropout with the rate 0.2 for the single memory model. For the dual memory model, we set dropout for Document RNN to 0.2 and for the encoder and decoder to 0.5. Mini-batching is used in both stages to speed up training. For the largest dataset, the document NMT model takes about 4.5 hours per epoch to train on a single P100 GPU, while the sentence-level model takes about 3 hours per epoch for the same settings.

When training the document NMT model in the second stage, we need the target memory. One option would be to use the ground truth translations for building the memory. However, this may result in inferior training, since at the test time, the decoder iteratively updates the translation of sentences based on the noisy translations of other sentences (accessed via the target memory). Hence, while training the document NMT model, we construct the target memory from the translations *generated* by the pre-trained sentence-level model[7]. This effectively exposes the model to its potential test-time mistakes during the training time, resulting in more robust learned parameters.

5.1 Main Results

We have three variants of our model, using: (i) only the source memory (*S-NMT+src mem*), (ii) only the target memory (*S-NMT+trg mem*), or

[4]We do not split words into subwords using BPE (Sennrich et al., 2016) as that increases sentence lengths resulting in removing long documents due to GPU memory limitations, which would heavily reduce the amount of data that we have.

[5]In our initial experiments, we found SGD to be more effective than Adam/Adagrad; an observation also made by Bahar et al. (2017).

[6]For the document NMT model training, we did some preliminary experiments using different learning rates and used the scheme which converged to the best perplexity in the least number of epochs while for sentence-level training we follow Cohn et al. (2016).

[7]We report results for two-pass decoding, i.e., we only update the translations once using the initial translations generated from the base model. We tried multiple passes of decoding at test-time but it was not helpful.

	Memory-to-Context								Memory-to-Output							
	BLEU				METEOR				BLEU				METEOR			
	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En
		NC-11	NC-16			NC-11	NC-16			NC-11	NC-16			NC-11	NC-16	
S-NMT	20.85	5.24	9.18	20.42	23.27	10.90	14.35	24.65	20.85	5.24	9.18	20.42	23.27	10.90	14.35	24.65
+src	21.91†	6.26†	10.20†	22.10†	24.04†	11.52†	15.45†	25.92†	**21.80**†	6.10†	9.98†	21.50†	23.99†	11.53†	15.29†	25.44†
+trg	21.74†	6.24†	9.97†	21.94†	23.98†	11.58†	15.32†	25.89†	21.76†	**6.31**†	10.04†	21.82†	24.06†	**12.10**†	15.75†	25.93†
+both	**22.00**†	**6.57**†	**10.54**†	**22.32**†	**24.40**†	**12.24**†	**16.18**†	**26.34**†	21.77†	6.20†	**10.23**†	**22.20**†	**24.27**†	11.84†	**15.82**†	**26.10**†

Table 2: BLEU and METEOR scores for the sentence-level baseline (S-NMT) vs. variants of our Document NMT model. **bold**: Best performance, †: Statistically significantly better than the baseline.

Lang. Pair	Memory-to-Context			Memory-to-Output		
	Fr→En	De→En	Et→En	Fr→En	De→En	Et→En
S-NMT	42.5	66.8	58.4	42.5	66.8	58.5
+src mem	48.8	73.1	64.8	68.7	107.1	88.7
+trg mem	43.8	68.1	59.8	53.8	85.1	71.8
+both mems	50.1	74.4	66.1	80	125.4	102

Table 3: Number of model parameters (millions).

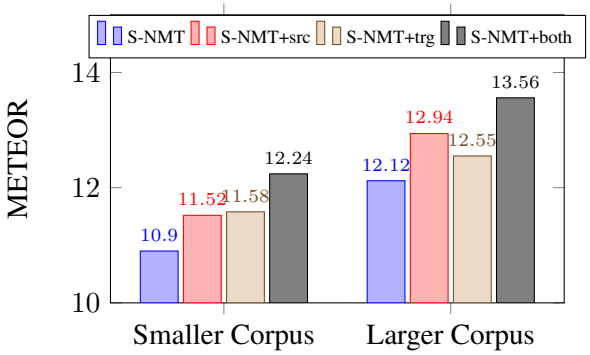
(iii) both the source and target memories (*S-NMT+both mems*). We compare these variants against the standard sentence-level NMT model (*S-NMT*). We also compare the source memory variants of our model to the local context-NMT models[8] of Jean et al. (2017) and Wang et al. (2017), which use a few previous source sentences as context, added to the decoder hidden state (similar to our Memory-to-Context model).

Memory-to-Context We consistently observe +1.15/+1.13 BLEU/METEOR score improvements across the three language pairs upon comparing our best model to *S-NMT* (see Table 2). Overall, our document NMT model with both memories has been the most effective variant for all of the three language pairs.

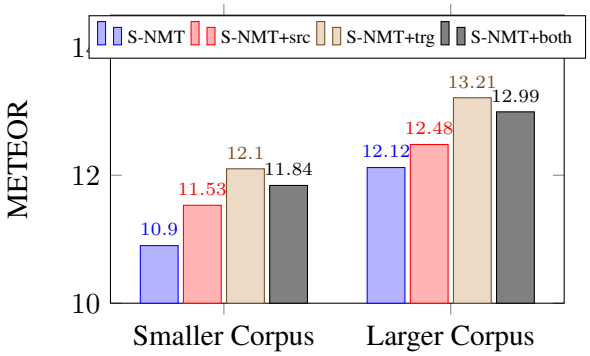
We further experiment to train the target memory variants using *gold* translations instead of the generated ones for German-English. This led to -0.16 and -0.25 decrease[9] in the BLEU scores for the target-only and both-memory variants, which confirms the intuition of constructing the target memory by exposing the model to its noises during training time.

Memory-to-Output From Table 2, we consistently see +.95/+1.00 BLEU/METEOR improvements between the best variants of our model and the sentence-level baseline across the three language pairs. For French→English, all variants of document NMT model show comparable performance when using BLEU; however, when evaluated using METEOR, the dual memory model is the best. For German→English, the target memory variants give comparable results, whereas for Estonian→English, the dual memory variant proves to be the best. Overall, the Memory-to-Context model variants perform better than their Memory-to-Output counterparts. We attribute this to the large number of parameters in the latter architecture (Table 3) and limited amount of data.

We further experiment with more data for train-



(a) Memory-to-Context model



(b) Memory-to-Output model

Figure 3: METEOR scores on De→En (NC-11) while training S-NMT with smaller vs. larger corpus.

[8]We implemented and trained the baseline local context models using the same hyperparameters and training procedure that we used for training our memory models.

[9]Latter is statistically significant decrease w.r.t. the both memory model trained on generated target translations.

	BLEU				METEOR			
	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En
		NC-11	NC-16			NC-11	NC-16	
Jean et al. (2017)	21.95	6.04	10.26	21.67	24.10	11.61	15.56	25.77
Wang et al. (2017)	21.87	5.49	10.14	22.06	24.13	11.05	15.20	26.00
S-NMT	20.85	5.24	9.18	20.42	23.27	10.90	14.35	24.65
+src mem	21.91†	6.26$^{\clubsuit}$	10.20	22.10$^{\spadesuit}$	24.04†	11.52$^{\clubsuit}$	15.45$^{\clubsuit}$	25.92$^{\spadesuit}$
+both mems	**22.00**†	**6.57**$^{\diamondsuit}$	**10.54**$^{\clubsuit}$	**22.32**$^{\diamondsuit}$	**24.40**$^{\diamondsuit}$	**12.24**$^{\diamondsuit}$	**16.18**$^{\diamondsuit}$	**26.34**$^{\diamondsuit}$

Table 4: Our Memory-to-Context Source Memory NMT variants vs. S-NMT and Source context NMT baselines. **bold**: Best performance, †, ♠, ♣, ♢: Statistically significantly better than only S-NMT, S-NMT & Jean et al. (2017), S-NMT & Wang et al. (2017), all baselines, respectively.

	BLEU-1			
	Fr→En	De→En		Et→En
		NC-11	NC-16	
Jean et al. (2017)	52.8	30.6	39.2	51.9
Wang et al. (2017)	52.6	28.2	38.3	52.3
S-NMT	51.4	28.7	36.9	50.4
+src mem	53.0	30.5	39.1	52.6
+both mems	**53.5**	**33.1**	**41.3**	**53.2**

Table 5: Unigram BLEU for our Memory-to-Context Document NMT models vs. S-NMT and Source context NMT baselines. **bold**: Best performance.

ing the sentence-based NMT to investigate the extent to which document context is useful in this setting. We randomly choose an additional 300K German-English sentence pairs from WMT'14 data to train the base NMT model in stage 1. In stage 2, we use the same document corpus as before to train the document-level models. As seen from Figure 3, the document MT variants still benefit from the document context even when the base model is trained on a larger bilingual corpus. For the Memory-to-Context model, we see massive improvements of $+0.72$ and $+1.44$ METEOR scores for the source memory and dual memory model respectively, when compared to the baseline. On the other hand, for the Memory-to-Output model, the target memory model's METEOR score increases significantly by $+1.09$ compared to the baseline, slightly differing from the corresponding model using the smaller corpus ($+1.2$).

Local Source Context Models Table 4 shows comparison of our Memory-to-Context model variants to local source context-NMT models (Jean et al., 2017; Wang et al., 2017). For French→English, our source memory model is comparable to both baselines. For German→English, our *S-NMT+src mem* model is comparable to Jean et al. (2017) but outperforms Wang et al. (2017) for one test set according to BLEU, and for both test sets according to METEOR. For Estonian→English, our model outperforms Jean et al. (2017). Our global source context model has only surface-level sentence information, and is oblivious to the individual words in the context since we do an offline training to get the sentence representations (as previously mentioned). However, the other two context baselines have access to that information, yet our model's performance is either better or quite close to those models. We also look into the unigram BLEU scores to see how much our global source memory variants lead to improvement at the word-level. From Table 5, it can be seen that our model's performance is better than the baselines for majority of the cases. The *S-NMT+both mems* model gives the best results for all three language pairs, showing that leveraging both source and target document context is indeed beneficial for improving MT performance.

5.2 Analysis

Using Global/Local Target Context We first investigate whether using a local target context would have been equally sufficient in comparison to our global target memory model for the three datasets. We condition the decoder on the previous target sentence representation (obtained from the last hidden state of the decoder) by adding it as an additional input to all decoder states (*PrevTrg*) similar to our Memory-to-Context model. From Table 6, we observe that for French→English and Estonian→English, using all sentences in the target context or just the previous target sentence gives comparable results. We may attribute this to these specific datasets, that is documents from TED talks or European Parliament Proceedings may depend more on the local than on the global context. However, for German→English (NC-11), the target memory model performs the best show-

	BLEU			METEOR		
Lang. Pair	Fr→En	De→En	Et→En	Fr→En	De→En	Et→En
S-NMT	20.85	5.24	20.42	23.27	10.90	24.65
+prev trg	**21.75**	5.93	**22.08**	**24.03**	11.40	**25.94**
+trg mem	21.74	**6.24**	21.94	23.98	**11.58**	25.89

Table 6: Analysis of target context model.

ing that for documents with richer context (e.g. news articles) we do need the global target document context to improve MT performance.

Output Analysis To better understand the dual memory model, we look at the first sentence example in Table 7. It can be seen that the source sentence has the noun "Qimonda" but the sentence-level NMT model fails to attend to it when generating the translation. On the other hand, the single memory models are better in delivering some, if not all, of the underlying information in the source sentence but the dual memory model's translation quality surpasses them. This is because the word "Qimonda" was being repeated in this specific document, providing a strong contextual signal to our global document context model while the local context model by Wang et al. (2017) is still unable to correctly translate the noun even when it has access to the word-level information of previous sentences.

We resort to manual evaluation as there is no standard metric which evaluates document-level discourse information like consistency or pronominal anaphora. By manual inspection, we observe that our models can identify nouns in the source sentence to resolve coreferent pronouns, as shown in the second example of Table 7. Here the topic of the sentence is "*the country under the dictatorship of Lukashenko*" and our target and dual memory models are able to generate the appropriate pronoun/determiner as well as accurately translate the word '*diktatuur*', hence producing much better translation as compared to both baselines. Apart from these improvements, our models are better in improving the readability of sentences by generating more context appropriate grammatical structures such as verbs and adverbs.

Furthermore, to validate that our model improves the consistency of translations, we look at five documents (roughly 70 sentences) from the test set of Estonian-English, each of which had a word being repeated in the gold translation. Our model is able to resolve the consistency in 22 out of 32 cases as compared to the sentence-based model which only accurately translates 16 of those. Following Wang et al. (2017), we also investigate the extent to which our model can correct errors made by the baseline system. We randomly choose five documents from the test set. Out of the 20 words/phrases which were incorrectly translated by the sentence-based model, our model corrects 85% of them while also generating 10% new errors.

Source	qimonda täidab lissaboni strateegia eesmärke.
Target	qimonda meets the objectives of the lisbon strategy.
S-NMT	<UNK> is the objectives of the lisbon strategy.
+Src Mem	the millennium development goals are fulfilling the millennium goals of the lisbon strategy.
+Trg Mem	in writing. - (ro) the lisbon strategy is fulfilling the objectives of the lisbon strategy.
+Both Mems	qimonda fulfils the aims of the lisbon strategy.
Wang et al. (2017)	<UNK> fulfils the objectives of the lisbon strategy.
Source	... et riigis kehtib endiselt lukašenka diktatuur, mis rikub inim- ning etnilise vähemuse õigusi.
Target	... this country is still under the dictatorship of lukashenko, breaching human rights and the rights of ethnic minorities.
S-NMT	... the country still remains in a position of lukashenko to violate human rights and ethnic minorities.
+Src Mem	... the country still applies to the brutal dictatorship of human and ethnic minority rights.
+Trg Mem	... the country still keeps the <UNK> dictatorship that violates human rights and ethnic rights.
+Both Mems	... the country still persists in lukashenko's dictatorship that violate human rights and ethnic minority rights.
Wang et al. (2017)	... there is still a regime in the country that is violating the rights of human and ethnic minority in the country.

Table 7: Example Et→En sentence translations (Memory-to-Context) from two test documents.

6 Related Work

Document-level Statistical MT There have been a few SMT-based attempts to document MT, but they are either restrictive or do not lead to significant improvements. Hardmeier and Federico (2010) identify links among words in the source document using a word-dependency model to improve translation of anaphoric pronouns. Gong et al. (2011) make use of a cache-based system to save relevant information from the previously generated translations and use that to enhance document-level translation. Garcia et al. (2014) propose a two-pass approach to improve the translations already obtained by a sentence-level model.

Docent is an SMT-based document-level decoder (Hardmeier et al., 2012, 2013), which tries to modify the initial translation generated by the Moses decoder (Koehn et al., 2007) through stochastic local search and hill-climbing. Garcia et al. (2015) make use of neural-based continuous word representations to incorporate distributional semantics into Docent. In another work, Garcia et al. (2017) incorporate new word embedding features into Docent to improve the lexical consistency of translations. The proposed methods fail to yield improvements upon automatic evaluation.

Larger Context Neural MT Jean et al. (2017)

extend the vanilla attention-based neural MT model (Bahdanau et al., 2015) by conditioning the decoder on the previous sentence via attention over its words. Extending their model to consider the global source document context would be challenging due to the large size of computation graph over all the words in the source document. Wang et al. (2017) employ a 2-level hierarichal RNN to summarise three previous source sentences, which is then used as an additional input to the decoder hidden state. Bawden et al. (2017) use multi-encoder NMT models to exploit context from the previous source and target sentence. They highlight the importance of target-side context but report deteriorated BLEU scores when using it. All these works consider a very local source/target context and completely ignore the global source and target document contexts.

7 Conclusion

We have proposed a document-level neural MT model that captures global source and target document context. Our model augments the vanilla sentence-based NMT model with external memories to incorporate documental interdependencies on both source and target sides. We show statistically significant improvements of the translation quality on three language pairs. For future work, we intend to investigate models which incorporate specific discourse-level phenomena.

Acknowledgments

The authors are grateful to André Martins and the anonymous reviewers for their helpful comments and corrections. This work was supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) (www.massive.org.au), and partially supported by a Google Faculty Award to GH and the Australian Research Council through DP160102686.

References

Parnia Bahar, Tamer Alkhouli, Jan-Thorsten Peter, Christopher Jan-Steffen Brix, and Hermann Ney. 2017. Empirical investigation of optimization algorithms in neural machine translation. In *Conference of the European Association for Machine Translation*. Prague, Czech Republic, pages 13–26.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

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

Julian Besag. 1975. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)* 24(3):179–195.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*. pages 261–268.

Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

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)*. Association for Computational Linguistics, pages 176–181. http://www.aclweb.org/anthology/P11-2031.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 876–885. http://www.aclweb.org/anthology/N16-1102.

Eva Martínez Garcia, Carles Creus, Cristina España-Bonet, and Lluís Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics* 108:85–96.

Eva Martínez Garcia, Cristina España-Bonet, and Lluís Màrquez. 2014. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural* 53:103–110.

Eva Martínez Garcia, Cristina España-Bonet, and Lluís Màrquez. 2015. Document-level machine translation with word vector models. In *Proceedings of the18th 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*. Association for

Computational Linguistics, pages 909–919. http://dl.acm.org/citation.cfm?id=2145432.2145532.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *International Workshop on Spoken Language Translation*. pages 283–289.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1179–1190. http://www.aclweb.org/anthology/D12-1108.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pages 193–198. http://www.aclweb.org/anthology/P13-4033.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the 10th Machine Translation Summit*. AAMT, pages 79–86.

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*. Association for Computational Linguistics, pages 177–180. http://www.aclweb.org/anthology/P07-2045.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, StatMT '07, pages 228–231. http://dl.acm.org/citation.cfm?id=1626355.1626389.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1412–1421. http://aclweb.org/anthology/D15-1166.

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*. Association for Computational Linguistics, pages 311–318. https://doi.org/10.3115/1073083.1073135.

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. http://www.aclweb.org/anthology/P16-1162.

Jason R. Smith, Herve Saint-Amand, Chris Callison-Burch, Magdalena Plamada, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the Conference of the Association for Computational Linguistics*. http://aclweb.org/anthology//P/P13/P13-1135.pdf.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. MIT Press, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. MIT Press, pages 3104–3112.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2816–2821. http://aclweb.org/anthology/D17-1300.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *Proceedings of the International Conference on Learning Representations*.