Diving Deep into Context-Aware Neural Machine Translation

Jingjing Huo[1,2] **Christian Herold**[2] **Yingbo Gao**[2] **Leonard Dahlmann**[1]
Shahram Khadivi[1] **Hermann Ney**[2]
[1]eBay, Inc., Aachen, Germany
{jihuo, fdahlmann, skhadivi}@ebay.com
[2]Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{surname}@i6.informatik.rwth-aachen.de

Abstract

Context-aware neural machine translation (NMT) is a promising direction to improve the translation quality by making use of the additional context, e.g., document-level translation, or having meta-information. Although there exist various architectures and analyses, the effectiveness of different context-aware NMT models is not well explored yet. This paper analyzes the performance of document-level NMT models on four diverse domains with a varied amount of parallel document-level bilingual data. We conduct a comprehensive set of experiments to investigate the impact of document-level NMT. We find that there is no single best approach to document-level NMT, but rather that different architectures come out on top on different tasks. Looking at task-specific problems, such as pronoun resolution or headline translation, we find improvements in the context-aware systems, even in cases where the corpus-level metrics like BLEU show no significant improvement. We also show that document-level back-translation significantly helps to compensate for the lack of document-level bi-texts.

1 Introduction

Even though machine translation (MT) has greatly improved with the emergence of neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) and more recently the Transformer architecture (Vaswani et al., 2017), there remain challenges which can not be solved by using sentence-level NMT systems. Among other issues, this includes the problem of inter-sentential anaphora resolution (Guillou et al., 2018) or the consistent translation across a document (Läubli et al., 2018), for which the system inevitably needs document-level context information.

In recent years, many works have focused on changing existing NMT architectures to incorporate context information in the translation process (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018). However, often times results are reported only on very specific tasks (most commonly subtitle translation), making it difficult to assess the potential of the different methods in a more general setting. This, together with the fact that big improvements are typically reported on low resource tasks, gives the impression that document-level NMT mostly improves due to regularization rather than from leveraging the additional context information. In this work we want to give a more complete overview of the current state of document-level NMT by comparing various approaches on a variety of different tasks including an application-oriented E-commerce setting. We discuss both, widely used performance metrics, as well as highly task-specific observations.

Another important aspect when talking about document-level NMT is the applicability in "real life" settings. There, when faced with a low resource data scenario, back-translation is an established way of greatly improving system performance (Sennrich et al., 2016a). However, to the best of our knowledge, the effect of back-translation data obtained and used by context-aware models has never been explored before. The main contributions of this paper are summarized below:

- We explore several existing context-aware architectures on four diverse machine translation tasks, consisting of different domains and data quantities.
- We examine the usage of context aware embeddings created by pre-trained monolingual models and study to what extent these embeddings can be simplified.
- We conduct corpus studies and extensive analysis on corpus specific phenomena like pronoun resolution or headline translation to give

Proceedings of the 5th Conference on Machine Translation (WMT), pages 604–616
Online, November 19–20, 2020.

an interpretation of the potential improvements from leveraging context information.

- We study the effects of utilizing document-level monolingual data via back-translation and report significant improvements particularly for document-level NMT systems.

2 Related Works

The discourse- or document-level translation is a long-standing and unsolved topic in the machine translation community (Mitkov, 1999; Carpuat, 2009; Hardmeier, 2014). Although neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) has recently become the dominant translation paradigm that provides superior performance, the independence between sentences is still the fundamental assumption taken for granted by most NMT systems. This means, that discourse-level phenomena between sentences such as pronominal reference, consistent lexical choice, and verbal tenses, etc. can not be addressed by these sentence-level NMT systems (Läubli et al., 2018; Guillou et al., 2018). The current NMT approaches tackling inter-sentential discourse phenomena can be roughly categorized into three aspects, augmenting NMT by

- adding source-side context
- including both source- and target-side context
- utilizing source- and/or target-side document-level monolingual data

To include the source-side context, Tiedemann and Scherrer (2017) concatenate consecutive sentences as input to the NMT system, while Jean et al. (2017); Bawden et al. (2018); Zhang et al. (2018) use an additional encoder to extract contextual information from a few previous source-side sentences. These works only consider a local context, including a few previous sentences. Some researches seek to capture the global document context; Wang et al. (2017) summarize the global context from all previous sentences in a document with a pre-trained hierarchical RNN and then use it for updating decoder states. Very recently, Chen et al. (2020) proposed a discourse structure-based encoder that takes account of the discourse structure information of the input document.

For adding additional target-side context, Tiedemann and Scherrer (2017); Agrawal et al. (2018) conduct multi-sentences decoding and observe only a minor improvement. Maruf and Haffari (2018) apply cache-based models to store vector representations for both source- and target-side context. Similarly, Tu et al. (2018) augment their NMT system with an external cache to memorize the translation history. Werlen et al. (2018) integrate two hierarchical attention networks (HAN) (Yang et al., 2016) in the NMT model to take account for source and target context. Maruf et al. (2019) apply a hierarchical attention module on sentences and words in the context to select contextual information that is more relevant to the current sentence.

For incorporating document-level monolingual data from the source language, Zhu et al. (2020) use BERT (Devlin et al., 2019) to model the source-side context and integrate it with the encoder and decoder of the NMT model. Junczys-Dowmunt (2019) share the parameters of a BERT-style encoder trained on monolingual documents with the MT model.

To utilize the document-level monolingual data from the target language, Junczys-Dowmunt (2019) also submit a system that trained on the combination of real and synthetic document-parallel data obtained by back-translation. However, they do not consider document-level back-translation. Voita et al. (2019a) proposed a document-level post-editing system which is trained only using the monolingual document-level corpus.

Recently, there has been a tendency in the community to conclude that the context used in a context-aware MT model works as regularisation or noise generator. Kim et al. (2019) compare several multi-encoders methods and claim that including this additional information can improve translation performance, but it is mostly due to the regularization effect rather than the contextual information. Li et al. (2020) also compare some context-aware architectures by replacing the real context with some random signal and show that random signals can achieve the same level improvement as the real context. However, it should be taken with a grain of salt since solving this task, along with the analysis, is quite challenging. There are many impact factors from the architecture, the data at hand, to the metric being used for evaluation.

One issue that can not be ignored in all discourse-related researches is the problem of evaluation. Since some discourse-level phenomena between sentences appear less frequently, although relevant, there is doubt if the metrics like BLEU score (Papineni et al., 2002) can capture these complex re-

lationships (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010). To get more insights into the capacities dealing with discourse-level phenomena of their MT models, some researchers use more targeted evaluation scores (Wong et al., 2020), like the Accuracy of Pronoun Translation (APT) Werlen and Popescu-Belis (2017), or they evaluate their systems on some specific test suites that contain more and more complex discourse phenomena (Hardmeier et al., 2015; Guillou et al., 2018; Müller et al., 2018; Voita et al., 2019b).

3 Document-level NMT

In this section, we first describe several commonly used context-aware NMT architectures and highlight the differences among them, largely following the work by Kim et al. (2019). Afterwards, we describe one radical attempt to represent the document-level context in one single embedding vector using BERT (Devlin et al., 2019). Finally, we explain our proposed paradigm to use document-level back-translation in detail. Note that in this work, we consistently use Transformer (Vaswani et al., 2017) as our basic architecture and modify it accordingly.

3.1 Context-Aware Architectures

Given a source sentence in a document to be translated, in order to exploit the source-side context from its previous sentences in the same document, a simple and straightforward technique is to concatenate these contextual sentences with the current source sentence (Tiedemann and Scherrer, 2017; Agrawal et al., 2018). Similarly, if the previous and current target sentences are to be generated together, i.e. $\underline{\mathrm{e}}_1^I = e_1^{I_{\mathrm{pre}}}$ BREAK $e_1^{I_{\mathrm{cur}}}$, then the target-side context can also be considered by the model (Tiedemann and Scherrer, 2017). Two additional special tokens are introduced to indicate the boundary between sentences and the beginning of a document, respectively. In this case, there is no modification of the model architecture itself, as seen in Figure 1.

An alternative way to model the source-side context is via an additional encoder, as shown in Figure 2. The previous sentence f_{pre} is fed into an additional encoder to obtain the hidden representation of the source context sentence $h_{j_{\mathrm{pre}}}^{L-1}$. At the last layer of the encoder, the source representation h_j^{L-1} attends to $h_{j_{\mathrm{pre}}}^{L-1}$ and outputs the combined hidden representation c_j^L (Voita et al., 2018). Then,

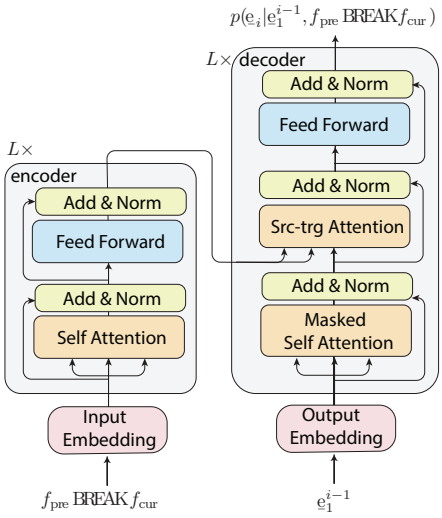

Figure 1: Single Encoder (2to2) approach only considering the one previous source sentence as context.

a gating mechanism (Bawden et al., 2018) between h_j^L and c_j^L is followed:

$$g_j = \sigma(W_g[h_j^L, c_j^L] + b_g) \quad (1)$$

$$o_j = g_i \odot W_s h_j^L + (1 - g_i) \odot W_c c_j^L \quad (2)$$

We refer to this approach as "Multi-Encoders (Out.)".

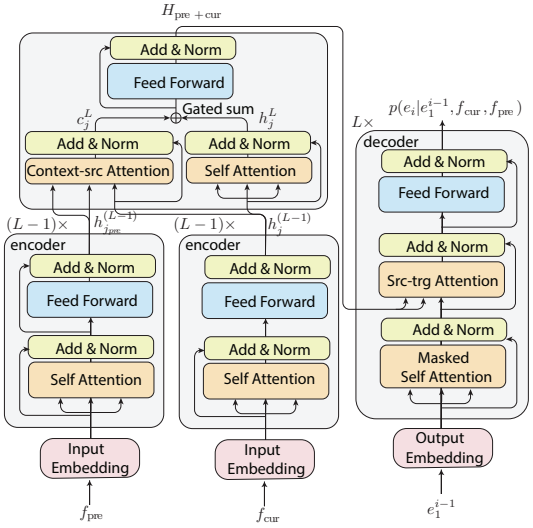

Figure 2: Multi-Encoders Out-side of decoder approach (Out.).

Another way to do the integration is to keep the representation of the current source sentence and the representation of the contexts separate and allow the decoder to have access to the context representations. Figure 3 shows a sequential integration inside of the decoder, where the decoder firstly attends to the current source representation,

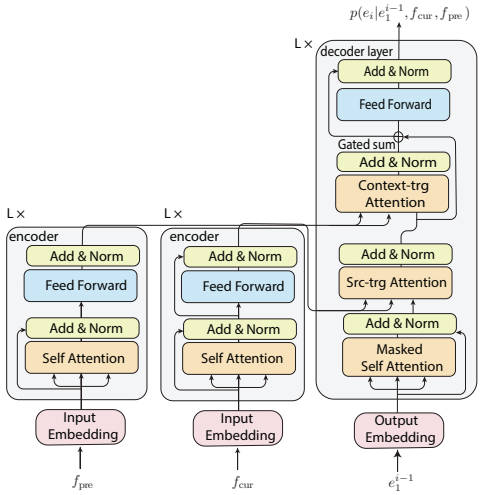

Figure 3: Multi-Encoders followed by attention components Inside of decoder Sequentially (In. Seq.).

then its output attends to the context representation (Zhang et al., 2018). The same gating mechanism as in the Multi-Encoders (Out.) approach is used between the two attention outputs. We refer to this approach that uses multi-encoders followed by attention components inside of decoder sequentially as "Multi-Encoders (In. Seq.)".

Figure 4 shows a parallel integration of the context inside of the decoder, where the decoder attends to the source and context representation in parallel and the outputs of them are combined again using a gating mechanism (Bawden et al., 2018). In this paper, we call this approach using multiple encoders followed by attention components inside the decoder in parallel "Multi-Encoders (In. Par.)".

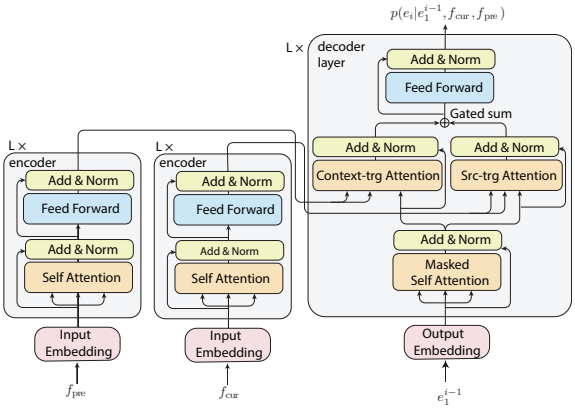

Figure 4: Multi-Encoders followed by attention components Inside of decoder in Parallel (In. Par.).

In addition, we use "WordEmb (In. Par.)" to refer to the approach that only uses word embeddings without any hidden layers to model the context and integrate it following the Multi-Encoders (In. Par.).

Considering that a pre-trained model like ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019) can capture rich representations of the input, it is apparent that one can also use it to model contextual information. Figure 5 shows the BERT-fused model proposed in Zhu et al. (2020), which uses a BERT encoder to obtain the BERT hidden representations H_B on the concatenation of the context sentence f_{pre} and the current source sentence f_{cur}. H_B is further fused into each layer of the encoder

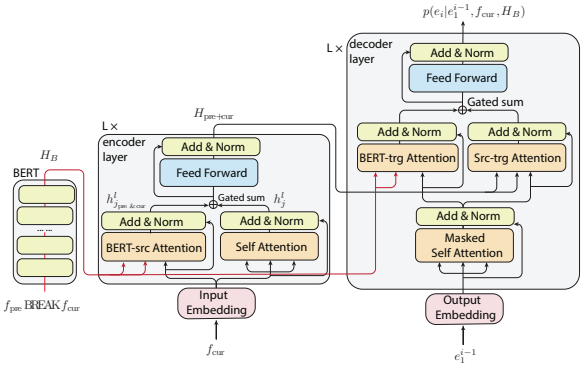

Figure 5: BERT sequence embeddings approach (Zhu et al., 2020).

and decoder of the NMT model using the attention mechanism to obtain the context representation. Instead of using the summation operation like in the original paper, we combine the context representation $h^l_{j_{\text{pre\&cur}}}$ and source representation h^l_j with a gating mechanism on the encoder side. Similar operation for the integration on the decoder side is used. This approach corresponds to the "BERT sequence embeddings (emb.)" approach in our main results in Table 3.

3.2 Single Embedding Vector as Context Representation

The introduction of additional encoders or attention components in the approaches mentioned in Section 3.1 brings a large number of parameters, which is not always ideal. Further, we propose one radical attempt to summarize the document-level context into one single embedding vector. We average the embeddings in the context representation H_B obtained by BERT to obtain one single mean-pooled embedding and then concatenate it with the word embeddings of the current source sentence along time axis (T-axis) or feature axis (F-axis). Besides, for the e-commerce dataset, we also apply a variant of BERT, which we call eBERT, that was

trained with additional e-commerce item titles as supplement in-domain data.

3.3 Document-level Back-translation

While there exist many works showing the improvements of context-aware systems, some major aspects are typically not covered - one of them being back-translation (Sennrich et al., 2016a). Back-translation is an integral part when building the strongest possible systems and is currently the best way to include monolingual data in the training of a NMT system. It uses an inverse, target-to-source, MT model to generate synthetic source sentences given target-side monolingual sentences. There exists a series of works on this topic (Hoang et al., 2018; Burlot and Yvon, 2018; Graça et al., 2019). However, the underlying inverse MT model used so far is mostly on the sentence level.

In this work, we argue that back-translation could be even more crucial when training document-level NMT systems, since even for common language pairs like German-English we have very limited amounts of parallel document-level data while having an abundance of monolingual document-level data. In addition, except for using a sentence-level inverse NMT model, we also introduce a document-level inverse MT model to generate pseudo source documents given monolingual target-side documents. The intuition behind this approach is that we expect the document-level back-translation system to keep more inter-sentential discourse-phenomena in the synthetic source documents. If the back-translation system is merely on the sentence level, some discourse-phenomena, like consistent lexical choices, might not remain in the generated source documents. Losing this potentially large amount of discourse-phenomena is not beneficial for training a context-aware model.

Since there is a large amount of document-level, monolingual in-domain data in the form of the NewsCrawl corpora, we conduct back-translation experiments on the WMT task. Here, we first train a baseline model and a context-aware model on the WMT news-commentary v14 in the reverse direction (De-En). We decide to use Multi-Encoders (In. Par.) as our inverse context-aware model, as it has the best performance on the WMT task. Then we sample 4.8M sentences pairs from news-docs2018 monolingual corpus,[1] which contains 168K documents. Next, we use the inverse NMT models to translate them applying beam search with beam size 5 and concatenate the resulting bilingual synthetic data with the real documents in the news-commentary v14 dataset (En-De). Finally, we compare the performance of a sentence-level baseline (En-De) and a context-aware model, Single Encoder (2to2), on both concatenated corpora. To our knowledge, this is the first attempt to explore the document-level back-translation data systematically (see Section 4.3.5).

4 Experiments

4.1 Datasets

We experiment with various parallel document-level datasets including IWSLT TED talk English-Italian,[2] WMT news-commentary v14 English-German,[3] OpenSubtitles (Lison and Tiedemann, 2016) v2018 English-German[4] and an additional in-house e-commerce English-Chinese dataset. The test sets for the former two are the IWSLT 2017 test set and WMT newstest2018, respectively; for the latter two, we have created the dev and test sets ourselves by doing appropriate splits to the complete dataset.[5] The data statistics of bilingual corpora used for fine-tuning context-aware models are summarized in Table 1. In the IWSLT, WMT and OpenSubtitles datasets, there exists a boundary between documents. We first take them as sentence-level corpora to train the baseline and further fine-tune the context-aware system on them.

The context-aware part of the e-commerce dataset is quite small and distinct from the other tasks: it does not contain documents or talks, but rather sentence-level item descriptions from an e-commerce website. As translation context, we provide the title of the item, instead of preceding sentences. Item descriptions and titles are user-provided, so they may contain ungrammatical sentences, spelling errors, and other noise. We give the title as context on the source-side, and we have reference translations only for the descriptions. In

[1] http://data.statmt.org/news-crawl/doc/de

[2] https://sites.google.com/site/iwsltevaluation2017

[3] https://www.statmt.org/wmt18/translation-task.html

[4] http://opus.nlpl.eu/OpenSubtitles-v2018.php

[5] We randomly sample complete documents from different years for dev and test set. The precise document IDs are: dev: {1997/517700, 2002/696617, 2007/933906, 2012/2192989, 2017/6007584}, test: {1997/708495, 2002/257044, 2007/1036109, 2012/2322334, 2017/6190628}

	IWSLT			WMT			OpenSubtitles			E-commerce data		
# Sentences	233K/	1.6K/	1.2K	338K/	2.2K/	3.0K	22.5M/	3.5K/	3.8K	36K/	478/	1K
# Running words	4.7M/	31K/	22K	8.3M/	47K/	68K	188M/	30K/	30K	596K/	12K/	26K
Avg. sentence length	20/	20/	19	25/	22/	23	8/	9/	8	17/	25/	26

Table 1: Training/development/test corpora statistics.

order to get a strong baseline, we additionally use a large sentence-level e-commerce dataset consisting of 6M sentence pairs (2.7M in-domain and 3.3M out-of-domain e-commerce) to train the baseline system, and then use it as initialization for fine-tuning on the context-aware e-commerce dataset. This dataset allows us to investigate context-aware NMT in a realistic scenario, in which the majority of training data does not have additional context.

To get a better insight into the model's performance for tackling the pronoun translation, we evaluate our models on two targeted test sets: one is ControPro for OpenSubtitles, the other is a coreference-focused test set for WMT. ControPro is introduced in Müller et al. (2018), which is a contrastive test set extracted from OpenSubtitles with previous sentences as context. The source sentence has the English pronoun *it* and three corresponding German translations containing German counterparts *es*, *sie*, *er*, i.e., one of them is correct, and the other two are incorrect. The evaluation is done by counting the decisions that models rank the correct translation higher than the incorrect translations. In addition to using it in this way, we keep the source and the corresponding correct translation to form a standard test set containing 12K sentence pairs and measure the general translation quality on it.

	ControPro	Coreference
# Sentences	12K	1.1K
# Running words	129K	28K
Avg. sentence length	11	26

Table 2: Two targeted-test sets: ControPro (Müller et al., 2018) and coreference-focused test set extracted from WMT newstest 2008-2019 using NeuralCoref.

To create a targeted test set for WMT, we use an external tool called NeuralCoref[6]. We first apply this external tool to detect the coreference resolution in two consecutive sentences from newstest2008-2019, and then only keep the sentences where the coreference is resolved intersententially. This results in a targeted test set containing more inter-sentential discourse phenomena. The detailed statistics of these two targeted test sets are given in Table 2.

All language pairs are preprocessed with the Moses tokenizer[7] except for the Chinese corpus which is preprocessed with the chinese text segmentation tool "jieba"[8]. We apply byte pair encoding (Sennrich et al., 2016b) with 32k merge operations jointly for source and target languages.

4.2 Experimental setting

All models are implemented in open-source toolkit OpenNMT (Klein et al., 2017). For the sentence-level baseline system, we follow a 6-layer base Transformer model (Vaswani et al., 2017) and set the hidden size and embedding size as 512 and the dimension of the feed-forward layer as 2048. We use 8 heads for multi-head attention. For our context-aware models, we extend baseline system to include additional encoder with the same setting. In training, we use Adam optimizer (Kingma and Ba, 2014) or its variant Lazy Adam Optimizer for optimization and follow the learning rate schedule described in (Vaswani et al., 2017). The learning rate scale factor and warm-up steps are different for different datasets. In all our experiments, we share word embeddings over the source and the context. The context encoders are also initialized by the encoder of the sentence-level baseline.

For automatic evaluation, we report case-sensitive sacreBLEU score (Post, 2018) for all corpora except for e-commerce, on which the evaluation is done in Chinese character-level with case-insensitive sacreBLEU.

4.3 Analysis

4.3.1 Performance in Terms of BLEU

Table 3 shows the corpus-level BLEU-scores of all architectures on different tasks. For the baseline as well as the "source-side-only" systems we get similar results to Kim et al. (2019) on the IWSLT

[6] https://github.com/huggingface/neuralcoref

[7] http://www.statmt.org/moses

[8] https://github.com/fxsjy/jieba

System	Type	IWSLT BLEU[%]	WMT BLEU[%]	OpenSubtitles BLEU[%]	E-commerce data BLEU[%]
Baseline	N/A	31.6	28.4	37.3	33.7
Single Encoder (2to1)	s	31.7	28.3	37.5	32.8
Single Encoder (3to1)	s	31.1	28.5	36.7	N/A
Multi-Encoders (Out.)	s	31.3	28.6	37.6	34.0
Multi-Encoders (In. Seq.)	s	31.8	29.2	37.5	**34.6**
Multi-Encoders (In. Par.)	s	32.2	**30.1**	37.5	34.2
WordEmb (In. Par.)	s	31.9	29.8	37.3	34.3
Single Encoder (2to2)	s,t	32.3	28.9	**38.2**	N/A
BERT sequence emb. (e,d)	s,m	**32.8**	29.0	37.4	34.0
BERT sequence emb. (e)	s,m	32.3	29.3	36.5	34.2
BERT sequence emb. (d)	s,m	32.1	29.7	36.6	34.3
BERT single emb. (T-axis)	s,m	31.7	28.7	37.6	34.5
eBERT single emb. (T-axis)	s,m	N/A	N/A	N/A	34.5
BERT single emb. (F-axis)	s,m	31.6	28.7	36.7	32.3

Table 3: Comparison of document-level architectures on different tasks. "Type" indicates whether the context used is from source(s)- or target(t)-side or if additional monolingual(m) data is included. "e" or "d" following the name of BERT sequence emb. approach indicates whether the context representation is fused on the encoder or decoder.

and WMT tasks, with Multi-Encoders (In. Par.) being the strongest architecture. For the e-commerce data, Multi-Encoders (In. Seq.) performs slightly better. Interestingly, with these architectures we do not see improvements on the much larger OpenSubtitles corpus. This seems to confirm the suggestion of Kim et al. (2019) that these architectures work more as a regularization which is much more important for low resource tasks.

The Single Encoder (2to2) results in an improvement on all tasks excluding the e-commerce task, for which the method is not applicable due to the lack of target translation of the context (titles). The improvements on the OpenSubtitles test set are comparable to what has been reported in the literature (Tiedemann and Scherrer, 2017) while the improvements on the other tasks are a bit smaller. We notice that with this architecture, the improvements increase with decreasing average sentence length, which makes sense since it is known that the Transformer struggles with long input sequences (Rosendahl et al., 2019). This seems also to be indicated by the deteriorating performance of the Single Encoder (3to1) system, which confirms the findings of Agrawal et al. (2018).

Including context information through BERT sequence embeddings improves the performance on IWSLT, WMT and the e-commerce tasks but not on OpenSubtitles. The pre-trained BERT brings more (monolingual) data, which should again help primarily on the low resource tasks. Contrary to the before mentioned approaches, the BERT single embedding approach does not significantly increase the number of free parameters, but it only works on the e-commerce task in our experiments. This finding as well as the discrepancy between concatenating along the time or feature axis is discussed in detail in Section 4.3.2.

While these findings are consistent with previous works, we find it to be insufficient to just rely on corpus-level BLEU scores to come to a conclusion about the usefulness of these approaches. In the subsequent sections we discuss specific aspects of the translations which might be easily overlooked. Furthermore we investigate the utilization of back-translation (Sennrich et al., 2016a) for document-level systems, in an effort to compare these architectures in a more "real-life" setting where back-translation is almost always used.

4.3.2 Including BERT

When looking at the results in Table 3, we see that using the embeddings produced by BERT yields some decent improvements on all tasks except for OpenSubtitles. This might indicate that the improvements - at least in parts - come from the usage of additional data when training the BERT model rather than from an improved context representation. A drawback when using the BERT system combination is the introduction of many additional parameters and calculations. This can be drasti-

System	IWSLT		WMT		E-commerce data	
	# tokens	BLEU[%]	# tokens	BLEU[%]	# tokens	BLEU[%]
Reference	19931	-	64276	-	40149	-
Baseline	-226	31.6	+1117	28.4	-2672	33.7
BERT single emb. (T-axis)	-66	31.7	+879	28.7	-2174	34.5
Random emb. (T-axis)	+19	31.5	+1557	28.7	-2177	34.7

Table 4: Using different vectors for context representation. For the reference, the number of tokens stands for the total number of target tokens in the reference. In all consecutive lines, the number stands for the difference in the number of tokens compared to the reference.

cally reduced when using a single vector extracted from BERT as described in Section 3.2. However, the results of this approach are not significantly outperforming the baseline system on any tasks except for the e-commerce data.

Surprisingly, the eBERT does yield no further improvement over the BERT variant and the concatenation along the F-axis leads to a significant degradation in performance. These two factors lead us to believe that the context information is not the decisive factor but something else. To investigate this, we replaced the BERT-generated context vector with a random vector and compared the resulting BLEU scores which are shown in Table 4.

Depicted in this table are the BLEU score as well as the number of tokens in the respective hypothesis for the IWSLT, WMT and e-commerce tasks. For replacing the real context vector we create the random vector by sampling from the uniform distribution. Looking at the results, we see that our assumption is correct: the variant using a random vector yields the same improvements as the real context vector on the e-commerce task - even though it inhabits no relevant context information.

The reason behind this becomes clear when comparing the number of tokens produced in the hypotheses: On the e-commerce task we have a noticeable problem with under translation. We argue that by increasing the length of the input sequence we inevitably increase the length of the output, leading to a longer hypothesis and consequently to a smaller brevity penalty when calculating BLEU. This effect is not present for the other tasks at hand, since there we do not have a significant effect of under translation. We note that similar results were obtained very recently by Li et al. (2020), who also see improvements when replacing the context signal with random noise. However, we conclude that the underlying effect is a different one, since we see no improvements when concatenating along the feature axis or when evaluating on a different task. In conclusion, we argue that the improvements seen by using the BERT-embeddings for context information rather comes from additional data and other effects discussed in this section, rather than from the usage of actual context information.

4.3.3 Better Headline Translation using Context

In this section we discuss another unexpected effect of using context information in the translation, namely giving the system additional information about the nature of the input. In the WMT task, both the train and test data consist of articles composed of a headline followed by a body of text, consisting of several sentences. This means the only time the system has no context information at hand, is when translating the headline of an article. We argue that the system can in fact use this information to distinguish whether the input sequence at hand is a headline or a real sentence and act accordingly. Since a headline has a very distinguishable style compared with a complete sentence, this should lead to improvements in the translation quality. To examine this hypothesis, we separate the WMT test set into two parts: One consisting only of headlines and the other one consisting only of body of texts. We then evaluate the baseline system and our strongest document-level system (Multi-Encoders (In. Par.) for WMT) separately on both sets, The results can be seen in Table 5.

We see that the translations of both sets are improved when using the document-level setup. However, the improvement on the headlines is much larger (+4.5% BLEU) than on the body of text (+1.7% BLEU). When manually checking the hypotheses, we find that the baseline system often times tries to translate a headline as a "complete" sentence (e.g. including a verb) while the document level system translates these in a much more consistent style. This observation coincides with the fact

System	BLEU[%]
Baseline	28.4
Doc-level	**30.1**
Baseline_headlines	19.9
Doc-level_headlines	**24.4**
Baseline_newsbody	28.5
Doc-level_newsbody	**30.2**

Table 5: System performance in terms of BLEU on headlines vs body of text for the WMT test set. The document-level system is Multi-Encoders (In. Par.).

that the baseline system shows severe signs of over-translation (on average 14.9% more tokens than the reference) and the document-level system does not (-1.2%). We note that this effect is not responsible for the overall improvement in the corpus-level BLEU, since the ratio of headlines to text is very small (3.9%). This becomes clear when comparing the improvements on the body of text vs the complete test set - which is equal. We conclude that this is another instance where the context improves the translation quality even if it is not immediately obvious.

4.3.4 Pronoun Resolution

Testing the correct translation of pronouns is an established method to compare the context-awareness of document-level machine translation systems (Guillou et al., 2016; Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018; Werlen et al., 2018; Wong et al., 2020). It can be argued that the ability of correctly translating inter-sentential pronouns not only depends on the architecture at hand but also on the data which the system is trained on. We decide to test the pronoun resolution capabilities of our systems in two different ways: First we are using an automatic metric for the accuracy of pronoun translation (APT) (Werlen and Popescu-Belis, 2017) and second we use two targeted test sets described in Section 4.1. The results on OpenSubtitles and WMT can be found in Table 6.

We calculate BLEU and APT scores on both the OpenSubtitles test set and ControPro test set (without contrastive translations) (Müller et al., 2018). Furthermore we calculate the resolution accuracy on ControPro (with contrastive translations). We compare the sentence-level baseline system with the best performing document-level system on this task - Single Encoder (2to2) as well as the Single Encoder (2to1) system. Even though the latter does not significantly improve over the baseline on the OpenSubtitles test set, we find a significant increase in pronoun translation accuracy in terms of both evaluation methods. The Single Encoder (2to2) system is even stronger in terms of pronoun translation, outperforming the baseline system by an impressive 33.9% absolute accuracy on the targeted test set. When calculating BLEU on ControPro, the gap between the baseline and the document-level systems becomes significantly larger. The BLEU scores for the Single Encoder (2to2) and the Single Encoder (2to1) systems are equal.

When looking at the APT scores on WMT test set, the context-aware system does not provide much improvement. We assume the reason is that the portion of the potential improvement regarding inter-sentential pronoun resolution is quite small, having looked through this test set. The increased gap of APT score between the baseline system and the context-aware system on the coreference-focused test set confirms this assumption, as there are more inter-sentential coreference phenomena in this targeted test set. Note that the BLEU score gaps between the baseline and context-aware systems on both test sets are almost the same.

All in all we can conclude that in this case the context information is helpful for a better translation, even though the effect might not be visible when just looking at corpus level BLEU.

4.3.5 Document-level Back-translation

When dealing with document-level monolingual data, the question arises, whether a sentence-level back-translation system is sufficient to generate the synthetic data. In this section, we investigate the effect of the sentence-level back-translation data and document-level back-translation data on the baseline system as well as a context-aware system. The sentence-level baseline and context-aware model used to generate synthetic documents have 28.3% BLEU and 29.7% BLEU on the test set, respectively. The performance of the resulting En-De systems are summarized in Table 7.

When using the synthetic data generated by the sentence-level system we see a huge increase in performance for both systems (+5.5% BLEU for the sentence-level system and +7.2% BLEU for the document-level system). A large increase in performance is to be expected since we increase the amount of data by roughly a factor of 8. The systems trained on the synthetic data generated by the document-level system show even further improvements (+1.6% BLEU for the sentence-level

System	OpenSubtitles		ControPro			WMT		Coreference test	
	BLEU	APT	BLEU	APT	corr.	BLEU	APT	BLEU	APT
Baseline	37.3	52.8	30.5	35.4	48.7	28.4	40.6	18.9	24.0
Single Encoder (2to1)	37.5	53.4	**33.1**	47.4	64.3	28.3	40.8	19.0	25.6
Single Encoder (2to2)	**38.2**	**54.2**	**33.1**	**49.5**	**82.6**	**28.9**	**41.1**	**19.7**	**26.1**

Table 6: Targeted evaluation for OpenSubtitles and WMT. All numbers are in percentage.

BT-Data used	System	BLEU[%]
-	Sent-level	28.4
	Doc-level	28.9
Sent-level	Sent-level	33.9
	Doc-level	36.1
Doc-level	Sent-level	35.5
	Doc-level	**36.5**

Table 7: Including back-translation data to the WMT task. The architecture of the document-level system is the Single Encoder (2to2) approach.

system and +0.4% BLEU for the document-level system). This might be in part due to the fact that the document-level back-translation system is stronger than the sentence-level one.

A very interesting observation is that the document-level system profits significantly more from the synthetic data in both scenarios. This contradicts the proposition that document-level architectures function mainly as a form of regularization for low resource data-settings. To the contrary we see an especially large gap in the case where we use only the sentence-level back-translation system for synthetic data generation. We argue that the reason for this is, that the document-level system is more capable in recovering from errors made during the back-translation due to the context information. For example a wrongly translated pronoun on the source side will definitely lead the sentence-level system astray, but the document-level one might still recover when the context is correct. This assumptions is also supported by the fact that the gap between sentence-level and document-level system gets smaller when using synthetic data generated by the document-level system, since we assume less such errors get made by this system.

5 Conclusion

In this work, we give a comprehensive comparison of current approaches to document-level NMT. To draw meaningful conclusions, we report results for standard NMT metrics on four diverse tasks - differing in the domain and the data size. We find that there is no single best approach to document-level NMT, but rather that different architectures work the best on various tasks. Looking at task-specific problems, such as pronoun resolution or headline translation, we find improvements in the context-aware systems, which is not visible in the corpus-level metric scores.

We also investigate methods to include document-level monolingual data on both source (using pre-trained embeddings) and target (using back-translation) sides. We argue that the performance improvements from the pre-trained encoder predominantly come from increased training data and other task-specific phenomena unrelated to actual context information utilization. Regarding back-translation, we find that document-level systems seem to benefit more from synthetically generated data than their sentence-level counterparts. We discuss that this is because document-level systems are more robust to sentence-level noise.

We plan to expand our experiments to incorporate document-level monolingual data on both source and target sides. This makes sense just by looking at the data conditions of almost every task: document-level parallel data is scarce, but document-level monolingual data is abundant.

Acknowledgments

Christian Herold and Yingbo Gao have received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1, project "CoreTec") and eBay Inc. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

References

Ruchit Rajeshkumar Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27.

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. *arXiv preprint arXiv:2006.04721*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *ACL 2019 Fourth Conference on Machine Translation*, Florence, Italy. [slides].

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the english–german MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 525–542.

Christian Hardmeier. 2014. *Discourse in statistical machine translation*. Ph.D. thesis, Acta Universitatis Upsaliensis.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

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

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In

Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 252–261.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. *arXiv preprint arXiv:2005.03393*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

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.

Sameen Maruf, André FT 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.

Ruslan Mitkov. 1999. Introduction: special issue on anaphora resolution in machine translation and multilingual NLP. *Machine translation*, pages 159–161.

Mathias Müller, Annette Rios Gonzales, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72.

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 of the Association for Computational Linguistics*, pages 311–318.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Jan Rosendahl, Viet Anh Khoa Tran, Weiyue Wang, and Hermann Ney. 2019. Analysis of positional encodings for neural machine translation. *IWSLT, Hong Kong, China*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

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, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 876–885.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212.

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

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

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25.

Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. *arXiv preprint arXiv:2004.09894*.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. *arXiv preprint arXiv:2002.06823*.