M2 Internship Report:
Document-level Machine Translation For Scientific Texts

Ziqian Peng [1]
Internship supervisor: François Yvon [2]
Academic tutor: Caio Corro [3]

Abstract

While neural machine translation has seen significant progress during recent years at sentence-level, translating full documents remains a challenge to efficiently incorporate document-level context. Various approaches have been proposed, but most of them consider only one to three previous source and/or target sentences as the context. This is not sufficient to faithfully translate some language phenomena, like lexical consistency and document coherence, especially in some scientific texts. In this work, we conducted experiments to include full contextual context and investigate the impact of all the past / future sentences on the source side with a context ablation study, on some abstracts from scientific publications. Our results show that future context is more influential than the past source context, and in our experiments, the Transformer architecture performs much better to translate the beginning of a long document than the end.

1. Introduction

While Neural machine translation (NMT) has experienced remarkable progress at sentence-level with the advent of the Transformer (Vaswani et al., 2017), human evaluation prefer human translation when translating full documents (Läubli et al., 2018). With only intra-sentence information, sentence-level NMT models cannot translate multiple discourse phenomena, such as anaphoric pronoun, formality, consistency and coherence, which requires long-term context (Bawden et al., 2018; Voita et al., 2019b; Maruf et al., 2019a; Herold & Ney, 2023a).

However, translating full documents remains a challenge to efficiently incorporate document-level context. Useful inter-sentence information is sparse in the document (Lupo et al., 2022), so that including global context may introduce lots of noise that distract the attention mechanism. Additionally, the attention mechanism suffers a quadratic complexity that limits its performances on long sequences. Even though a wide range of efficient transformers have been proposed to tackle this problem (Tay et al., 2022), none of them can significantly go beyond the original transformer for both quality and speed (Tay et al., 2021).

Therefore, contextualizing machine translation with inter-sentence context is necessary to boost machine translation quality, but tricky to deal with the noise in global context and the sequence length of long input. Different discourse phenomena can have diverse distributions across different languages, and they are usually sparse in data. As a result, identifying then automatically evaluating the translation of these phenomena is also complicated.

1.1. Related work

Various approaches have been proposed to perform context-aware document-level machine translations (DMT). A taxonomy (Abdul Rauf & Yvon, 2020) to classify them is

- `Single-encoder approaches`: Most methods in this category modify only the input, like the concatenation methods (Tiedemann & Scherrer, 2017), that simply concatenates the past sentences with the current one before feeding them to NMT models. Several others modified the architecture. For example, FLAT-Transformer (Ma et al., 2020) encoded the context and current sentence in a single encoder with different attention blocks. G-transformer (Bao et al., 2021) masks global context attention in lower layers, thus focusing more on current and neighboring sentences.

- `Multi-encoder approaches`: These approaches encode context information separately with current sentence, for example, using a specific encoder (Zhang et al., 2018; Bawden et al., 2018; Li et al., 2020), or adapting hierarchical attention networks (HAN) to model context information (Miculicich et al., 2018; Maruf et al., 2019a; Yin et al., 2021).

- `Cached-based approaches`: Cache-based methods store short-term memory about the recent context to boost the probabilities of recently generated target words (Maruf & Haffari, 2018; Tu et al., 2018).

- `Multi-pass approaches`: Usually, multi-pass systems refine translations of decontextualized first-pass system, with the help of document-level monolingual data (Voita et al., 2019a; Yu et al., 2020)

Some recent works fall into the categories above, but their main insights are rather improved training methods, such as the `multi-resolution training` (Sun et al., 2022) for single-encoder methods, and the `divide and rule` (Lupo et al., 2022) pretraining strategy to train the context-specific parameters in a HAN context encoder.

Regarding evaluation, traditional automatic metric such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) are widely used to report sentence-level translation quality, but they are unreliable to evaluate discourse phenomena in DMT. A common way to assess context-aware translations is contrastive evaluation (Müller et al., 2018; Bawden et al., 2018), where DMT models need to rank correct translations higher than the contrastive (incorrect) ones in the contrastive test suits. Recently, a generative test suit GenPro (Post & Junczys-Dowmunt, 2023) is proposed to assess models' generative ability instead of their discriminative ability.

Since these test suits are created for specific discourse phenomena in specific language pair, several context-aware automatic metrics are proposed to mitigate this limitation, such as the cross-mutual information (CXMI) (Fernandes et al., 2021), MuDA (Fernandes et al., 2023) and BlonDe (Jiang et al., 2022).

1.2. Machine translation with full documents

Although diverse approaches have been studied as mentioned above, most of them consider only one to three previous source and/or target sentences as the context (Maruf et al., 2019b; Abdul Rauf & Yvon, 2020). Only a few include the future source context or the full context (Macé & Servan, 2019; Bao et al., 2021; Sun et al., 2022).

However, long-term context is indispensable to translate several discourse phenomena such as lexical consistency and document coherence. When translating a scientific document that defines a new term at the beginning, this term should be consistently translated across the whole document. It is therefore worth taking the whole context into account, in spite of some technical issues to overcome.

When we process full documents in NMT, since the input sequences become much longer than one sentence, the computational complexity increases. Since the larger the size of each example, the smaller the batch size, the training process has fewer gradient update steps than training with only a few contexts. In addition, the attention weights of the current target token are spread out throughout the full document rather than within the current sentence. As useful information in global contexts is sparse, some attention is wasted to focus on noise instead of what we need to assist the translation. When generating the target texts, the beam search procedure has to consider longer branches as well (Herold & Ney, 2023b). If the past and future context of current sentences are not efficiently incorporated, this may lead to problems related to the search error and label bias (Stahlberg & Byrne, 2019).

Table 1. Amount of parallel segments in each dataset

Data set	SciPar	THE_doc	THE_sent
training set	1,116,325	2858	22949
validation set	3000	101	957
test set	3000	100	1007

Thus, in the framework of the MaTOS[1] project, we conducted experiments to explore some characteristics of DMT with full documents. In this preliminary research, we investigated the impact of the past and future source context by a context ablation study, on EN-FR scientific texts in Natural Language Processing (NLP) field. In particular, we are interested in how the attention distribution changes when translating full documents, and how to better integrate context information. Previous works (Agrawal et al., 2018; Zhang et al., 2020) conducted similar experiments, but only concentrated on local context within one to three past and future sentences.

Due to lack of document-level parallel data in the NLP domain, we created a document-level dataset with parallel EN-FR abstracts crawled from the web, and applied a two-stage training, i.e. fine-tuning from a pretrained sentence-level baseline. This training strategy is also suggested for better document-level translation quality (Liu et al., 2020). The resulting analysis shows that future context is more influential than the past source context, and the transformer does much better when translating the beginning of a long document than its end.

More details about the data preparation, model description and evaluation methods are reported in section 2, followed by the systematic result analysis in section 3. Then we conclude our study with potential future work in section 4.

2. Material and Methods

Since most open-source parallel corpus have only sentence-level metadata, we decided to pretrain a sentence-level base-

[1] The MaTOS project aims to develop machine translation techniques for open science, more details are at `https://anr-matos.github.io/`

Table 2. Average length of segments in each dataset (SciPar, THE_doc, THE_sent) by sentence pieces in English (left) and French (right), rounded to integer

Dataset	SciPar	THE_doc	THE_sent
training set	32 / 38	236 / 276	29 / 34
validation set	32 / 38	285 / 332	29 / 34
test set	32 / 38	294 / 344	28 / 33

Table 3. Amount of parallel pairs in other test sets (TAL, TED, IWSLT2023) and average length of segments by sentence pieces in English (left) and French (right), rounded to integer

Test set	Count (in sentence)	Average length (EN/FR)
TAL_doc	246	116 / 137
TAL_sent	1015	27 / 32
TED	6059	23 / 27
IWSLT2023	468	23 / 26

line, then train the DMT systems via fine-tuning.

2.1. Data

Pretraining is based on the SciPar (Roussis et al., 2022) dataset, which is a multilingual collection of parallel abstracts from openly published bachelor theses, master theses and doctoral dissertations across various fields. We focused on the EN-FR language pair and randomly selected 3000 parallel sentences for each of the validation set and the test set.

For fine-tuning, we took 1701 EN-FR parallel abstracts crawled from theses.fr[2] related to the NLP field and created a parallel document-level corpora denoted as THE. We randomly split 101 documents as validation set and 100 documents as test set, and we further extended this dataset with parallel EN-FR abstracts[3] from ISTEX.[4] With the raw parallel documents in hand, we first segmented both the English and French version into sentences with Trankit (Nguyen et al., 2021), which identify better sentences containing list of citations. Then we constructed parallel sentences using hunalign[5] (Halácsy et al., 2007). We concatenated the parallel sentences to form parallel documents, and we marked the sentence boundaries by a <sep> tag. Thus, we named the sentence-level data as THE_sent and the document-level one as THE_doc. To ensure the data quality, especially for the validation set and the test set, we evaluated the test set with TransQuest (Ranasinghe et al., 2020) that estimated each EN-FR pair to measure that to which extent they were parallel. Subsequently, we manually cleaned document pairs with score less than 0.3.

In addition to the THE dataset, we created a test set with parallel abstracts in NLP from the TAL journal,[6] denoted TAL. After the same data processing procedure as THE, we obtained a document-level test set TAL_doc and a sentence-level test set TAL_sent. Meanwhile, we tested also our EN-FR NMT systems on the TED talks from the IWSLT2016 (Cettolo et al., 2012) test set, which are talks related to scientific topics, and the IWSLT2023 development data (Salesky et al., 2023), that contain transcriptions of presentations in NLP field.

Since SciPar was also created with open source abstracts, we checked and removed sentence pairs that are duplicated in our other datasets from SciPar. All these corpora are encoded with unigram language model (Kudo, 2018) using sentencepiece (Kudo & Richardson, 2018)[7] with 32k joint vocabulary. The amount of parallel segments in each dataset, and the average segment length in sentence pieces are reported in Table 1,2 and 3.

2.2. Models

Generally, machine translation models are constructed to optimize the probability of appropriate translations given the source text. Let x_i be the i^{th} sentence in the source document and y_i be its corresponding target sentence. Translating a document of T sentences (i.e. $x_1 \cdots x_T$) can be represented as such:

$$\begin{aligned} & P(y_1 \cdots y_T | \boldsymbol{x_1 \cdots x_T}) \\ & = \prod_{l=1}^{\sum_{t=1}^{T} L_t} P(y_l | \boldsymbol{y}_{<l}, \boldsymbol{x_1 \cdots x_T}) \\ & = \prod_{t=1}^{T} \prod_{l=1}^{L_t} P(y_{f(t,l)} | \boldsymbol{y}_{<f(t,l)}, \boldsymbol{x_1 \cdots x_T}) \end{aligned} \tag{1}$$

where L_t is the length of sentence y_t, and $f(t,l)$ is

$$f(t,l) = (\sum_{i=0}^{t-1} L_i) + l, \text{ with } L_0 = 0 \tag{2}$$

the index of current token in the document.

The sentence-level NMT systems assume a conditional independence between sentences, to simplify the problem as:

$$P(y_1 \cdots y_T | x_1 \cdots x_T) = \prod_{t=1}^{T} \prod_{l=1}^{L_t} P(y_{t,l} | \boldsymbol{y}_{<l}, \boldsymbol{x_t}) \tag{3}$$

[2] The raw data were collected by Maxime Bouthors in 2022

[3] The ISTEX raw data was collected by Mathilde Huguin in the course of MaTOS project

[4] https://www.istex.fr/

[5] https://github.com/danielvarga/hunalign

[6] https://www.atala.org/node/16

[7] https://github.com/google/sentencepiece

while in reality, sentences in a document are usually connected in some way to convey a coherent message. Translating multiple discourse phenomena requires also contextual information from surrounding sentences or the full document.

To explore the machine translation of full scientific documents, the following models are compared:

- **Baseline**: a sentence-level NMT model trained with SciPar.
- **FTsent**: a sentence-level system fine-tuned from the baseline with `THE_sent` dataset.
- **FTdoc**: a document-level system fine-tuned from the baseline with `THE_doc` dataset.
- **FTdoc_MR**: the same as FTdoc but fine-tuned on augmented `THE_doc` training set using multi-resolution training (Sun et al., 2022). The augmented set contains 35376 segment pairs instead of 2858, and the average sentence length is reduced to 74 / 87 for EN/FR instead of 236 / 276 tokens.
- **FTdoc_maskPast**, **FTdoc_maskFuture** and **FTdoc_maskAll**: the same as FTdoc, but the past, future, and all source contexts are masked respectively.

We applied the original Transformer (Vaswani et al., 2017) architecture, to train the baseline, FTsent, FTdoc and FTdoc_MR. Consistently, we applied masks to this vanilla Transformer architecture for the ablation context study.

Furthermore, we implemented an attention factor matrix, that equally decreased the attention weights of past, future or all source context, to compare with the context mask that fully removes the corresponding context attention.

2.2.1. Transformer with context mask

To disable a part of the source context, we modified the cross-attention using a context mask M_c to set the attention weights of corresponding context as inf before applying the softmax:

$$Attention_M(Q, K, V) = softmax(\frac{QK^T}{\sqrt(d_k)} + M_c)V \tag{4}$$

Thus, for FTdoc_maskPast, we applied a context mask as shown in Figure 1, so that we optimized the translation without information about the past source:

$$\begin{aligned} &P(y_1 \cdots y_T | x_1 \cdots x_T) \\ &= \prod_{t=1}^{T} \prod_{l=1}^{L_t} P(y_{f(t,l)} | y_{<f(t,l)}, \boldsymbol{x_t} \cdots \boldsymbol{x_T}) \end{aligned} \tag{5}$$

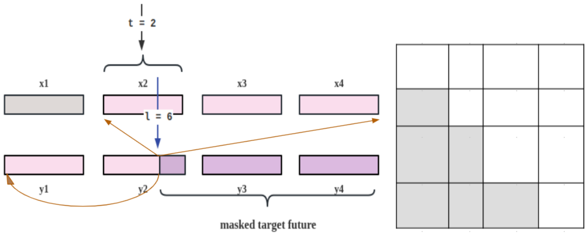

Figure 1. Decoder-side attention represented by brown arrows (left) and the context mask (right) for FTdoc_maskPast

Similarly, FTdoc_maskFuture and FTdoc_maskAll masked the future and all source context as Figure 2 to constrain the source context as $\boldsymbol{x_1} \cdots \boldsymbol{x_t}$ and $\boldsymbol{x_t}$ respectively. When masking future, the model could not converge during fine-tuning as it tended to finish the inference after decoding the first sentence.[8] Therefore, we forced the generation when an EOS tag was selected while the generated text comprised less `<sep>` tags than that of the corresponding source document if the max length is not reached in fairseq.

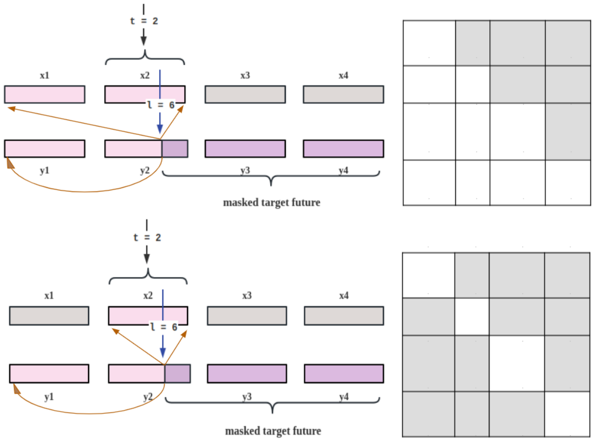

Figure 2. Decoder-side attention represented by brown arrows (left) and the context mask (right) for FTdoc_maskFuture (above) and FTdoc_maskAll (bottom)

2.2.2. Transformer with attention factor

Instead of totally excluding the past, future or all source context, we also implemented a smoother way to mask the context that decreased equally the attention weights by a

[8] As we fine-tuned `FTdoc_maskFuture` from a sentence-level baseline, during the validation step, the model tends to select the EOS tag to finish the generation after translating the first few sentences of a whole document. Therefore, the system can never generate full documents in this way.

factor α between 0 and 1 before the softmax:

$$Attention_F(Q,K,V) = softmax(\frac{QK^T}{\sqrt{(d_k)}} \odot M_\alpha)V \quad (6)$$

where $\odot$ indicates element-wise multiplication.

These matrices are similar to the context masks, only replacing the $\inf$ value as a fixed value for α, thus we can observe the differences brought by different amount of context.

2.3. Evaluation methods

2.3.1. General quality

We evaluated BLEU (Papineni et al., 2002; Post, 2018)[9] and COMET (Rei et al., 2020)[10] to assess the general translation quality of each system.

As most publications report BLEU score evaluated at sentence-level, we aligned the translated documents with its references using edlib,[11] that aligns sequences according to their edit (Levenshtein) distance (Levenshtein, 1965). Then we segmented the parallel texts with the aid of <sep> tag. To test the effectiveness of this method, for each system and both of the THE_doc and TAL_doc test set, we realigned the translated documents, then concatenated the resulting parallel sentences to recover the documents. Thus, we can compare the BLEU score of the original and recovered texts. Empirically, this approach of realignment can approximate the sentence-level score of generated target documents, with potential decrease of 0 to 0.2 BLEU score.

We also evaluated BLEU at the document-level, where translated sentences from THE_sent and TAL_sent are concatenated into full documents, and the <sep> tags were always excluded. The results are reported in the supplementary material for simplicity.

To our knowledge, no existing contrastive test set corresponds to our scenario of full scientific documents. While some context-aware automatic metrics (Fernandes et al., 2021; Jiang et al., 2022; Fernandes et al., 2023) are proposed, we leave the evaluation of discourse phenomena as future work due to the time limitation.

2.3.2. Attention analysis

Subsequently, we analyzed the attention distribution of FT-doc model when translating THE_doc test set. We assume that, to translate well the text, a transformer-like architecture should put more attention on the current sentences for the local context, and much less on the global context. The sentence-level NMT system, like our pretrained baseline, is a special case that focuses only on the current sentences. During fine-tuning, an amount of attention is expected to be assigned to the context.

Thus, we collected the average attention over 8 heads from the last decoder layer, and computed the amount of attention weights for the past, current, and future sentences.

Furthermore, we computed statistically the correlation between the length of documents in sentences and the average attention put on the last 5 current sentences, to explore the influence of input length for NMT with attention mechanism.

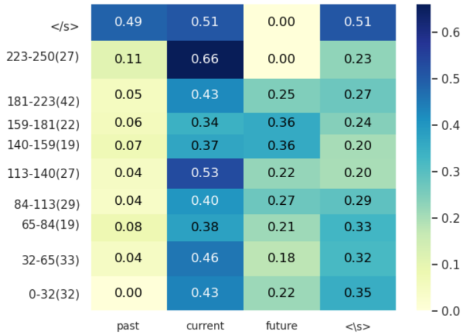

Figure 3. An example of attention distribution on past, current and future source sentences in a document with 9 sentences. The last column indicates the target attention on source EOS tag.

2.3.3. Evaluation with length constraint

In addition, we evaluated the translation output with respect to document length and sentence positions with THE_doc. We began by evaluating independently documents with different length ranges using SacreBLEU, The obtained results suggest that our models perform worse for long input especially due to a poor brevity penalty.

To explore more in detail the impact of the input length on the translation quality, we concatenated the first sentence of each document to its end, and compare the translation of the same sentences at different positions, for both THE_doc and TAL_doc. In addition, we also evaluated the translation of sentences at specific positions in the document, put in another way, we measured the translation quality of the first sentences from all documents, then all the second sentences, etc. The resulting analysis is described in section 3.

2.4. Experiments setting

Our experiments are implemented with the fairseq (Ott et al., 2019) framework. All models are based on the Transformer_base Transformer (Vaswani et al., 2017), with 6 layers, 8 attention heads, hidden size of 512 and feed forward size of 2048. The max position of each input sequence is limited to 4096. The max token size of each batch is 4096 for the baseline and 2048 for others, and we updated the

[9] https://github.com/mjpost/sacrebleu

[10] https://github.com/Unbabel/COMET

[11] https://pypi.org/project/edlib/

parameters every two batches. Our systems are all trained on NVIDIA RTX A5000, with a patience value of 5 to stop training if the most recent 5 epochs cannot improve the BLEU score on the validation set.

3. Results and analysis

3.1. Machine translation with all source context

3.1.1. Automatic metrics

Table 4 reports the BLEU score of our NMT systems with the original attention mechanism. In general, FTsent performs the best over all sentence-level test sets, except IWSLT2023, but it cannot translate a full document. FTdoc is the best to translate the full document in one pass. It is better than FTdoc_MR, suggesting that the multi-resolution training technique may not be helpful under our pretraining fine-tuning scenario. In contrast, as shown at the bottom of table 4, COMET is not informative in distinguishing the performances of our models. One reason is that it scored highly a hypothesis h of document translation even though h is only a good translation of the first sentence.

In addition, we observed that, using our DMT models FTdoc and FTdoc_MR, translating full documents in one go have equal (for TAL_doc) or worse (for THE_doc) performance than translating sentence by sentence for both n-gram precision and brevity penalty in BLEU score.

3.1.2. Attention Analysis

To find out the potential reasons of the worse performance of FTdoc compared with FTsent, we looked in details the attention distribution of each target sentences to the source sentences.

In particular, we calculated the attention weights distributed over the past, current and future sentences in THE_doc test set. We considered the future context and the EOS tag </s> separately, as this tag plays a special role in the translation procedure and attracted itself an important attention (cf. Figure 3). As reported in Table 5, future contextual sentences received more than twice as much attention than the past, and the current sentences received about a half of attention weights. We leave systematic exploration of the EOS tag for future work. Additionally, we collected the amount of attention focused on the last 5 current sentences of 90 documents that have enough sentences in THE_doc. Then we computed its correlation with the document length (cf. Figure 4). Spearman correlation[12] shows that the attention on the last 5 current sentences has a negative correlation of -0.529 with the document length. In other words, the longer the document length in sentences, the less the amount of attention concentrated on the last 5 current sentences. As our dataset contains only 90 examples, we executed a permutation test, that gave us a p-value of 0.0002, which demonstrated our hypothesis.

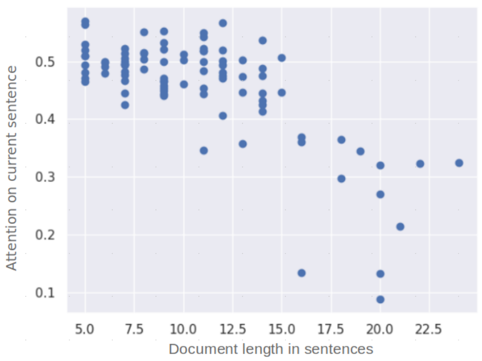

Figure 4. Average attention weights on the last 5 current sentences of 90 documents from THE_doc with respect to the document length in sentences.

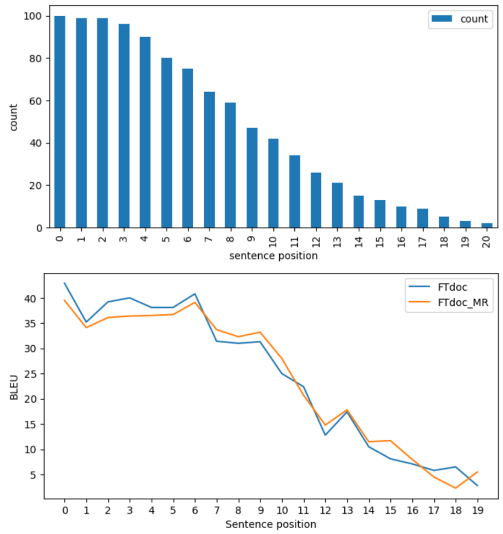

Figure 5. Evaluation of sentences in THE_doc with respect to their position in the document.

3.1.3. Performance with length constraint

Consequently, we evaluated the translation quality of our systems according to document length in sentences and in tokens, the results in supplementary material suggest that the brevity penalty gets worse with the increase of document length. We further copied the first sentence of each document to its end, and compare the translation of these identical sentences at different positions.

Table 6 shows that, for THE_doc, the translation quality degraded significantly using our DMT models, in contrast to the beginning. However, the same test was conducted

[12] We applied Spearman correlation because it does not require Gaussian distribution.

Table 4. Evaluation at sentence-level, with BLEU scores at the top, brevity penalty (BP) of BLEU in the middle and COMET at the bottom. The <sep> tags are always excluded for evaluation.

Score	Models	TED	IWSLT2023	THE_sent	THE_doc2sent	TAL_sent	TAL_doc2sent
BLEU	baseline	27.8	48.6	41.7	-	32.8	-
	FTsent	28.5	47.1	**43.2**	-	**34.3**	-
	FTdoc	24.0	41.2	40.8	**35.8**	32.7	**32.9**
	FTdoc_MR	21.2	40.2	39.6	34.4	32.3	31.7
BP	baseline	0.953	0.967	1.0	-	0.98	-
	FTsent	0.977	0.973	**1.0**	-	**0.985**	-
	FTdoc	0.984	0.984	1.0	**0.961**	0.983	**0.981**
	FTdoc_MR	0.977	0.97	1.0	0.955	0.986	0.983
COMET	baseline	0.704	0.796	0.848	-	0.818	-
	FTsent	0.696	0.796	0.847	-	0.820	-
	FTdoc	0.649	0.778	0.838	0.777	0.814	0.798
	FTdoc_MR	0.616	0.768	0.830	0.768	0.810	0.793

Table 5. Statistics of attention weights assigned to past, current, future sentences, and the EOS (</s>) tag.

Attention weights	**count**	**mean**	**std**	**min**	**max**
Past	100	0.0886	0.053	0.000	0.311
Current	100	0.4649	0.052	0.272	0.656
Future	100	0.2011	0.053	0.000	0.309
EOS	100	**0.2454**	0.056	0.112	0.468

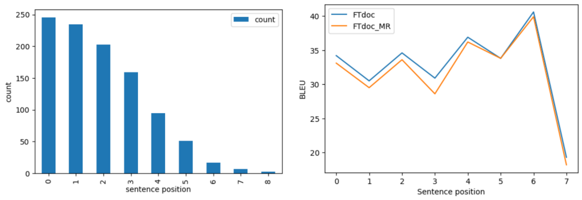

Figure 6. Evaluation of sentences in TAL_doc with respect to their position in the document.

Table 6. BLEU score for the translation of the first sentences copied to the end of each documents on THE_doc test set. BP denotes the brevity penalty.

Models (BLEU / BP)	**First**		**Last**	
FTdoc	43.1	0.996	29.4	0.857
FTdoc_MR	**39.9**	1.000	**27.1**	0.874
FTdoc_maskAll	42.8	1.000	18.0	1.000
FTdoc_maskFuture	**29.4**	1.000	**4.6**	1.000
FTdoc_maskPast	41.7	1.000	23.0	0.894

Table 7. BLEU score for the translation of the first sentences copied to the end of each documents on TAL_doc test set, where documents are much shorter than that in THE_doc

Models (BLEU / BP)	**First**		**Last**	
FTdoc	33.7	0.977	32.7	0.961
FTdoc_MR	33.1	0.980	31.8	0.956
FTdoc_maskAll	34.3	0.982	29.9	1.000
FTdoc_maskFuture	**28.4**	1.000	**9.7**	1.000
FTdoc_maskPast	34.0	0.978	32.5	0.991

on TAL_doc, and the difference between translating the beginning and the end of full documents is negligible (cf. Table 7).

Therefore, we extracted sentences at each position of the translated documents and assessed their quality for both THE_doc and TAL_doc.[13] Figure 5 illustrated the number of sentences from each collection and the BLEU score of their translation in THE_doc. The performance deteriorated with the increase of sentence location from the seventh sentence. Similarly, translation quality of TAL_doc (Figure 6) remained stable at the first part of documents, but decreased sharply at the seventh sentence using FTdoc and FTdoc_MR. We expect future experiments on larger datasets to confirm this result.

3.2. Context ablation study

In this section, we masked the past, future, or all source context during cross attention in FTdoc to measure their impact for DMT.

Table 8 shows that, FTdoc_maskPast got slightly worse than FTdoc, while masking future context resulted in loose of 10 BLEU score on THE_doc test set. This result is also reflected by scores of COMET and the evaluation with length constraint.

Instead of masking in a binary way, we also applied an attention factor that weakened the attention weights of context,

[13] In other words, given a translation with 20 sentences for a source document of length 24, we extracted 20 instead of 24 sentences.

Table 8. Evaluation at sentence-level, with BLEU scores at the top, brevity penalty (BP) of BLEU in the middle and COMET at the bottom. The <sep> tags are always excluded for evaluation.

Score	Models	TED	IWSLT2023	THE_sent	THE_doc2sent	TAL_sent	TAL_doc2sent
BLEU	FTdoc	24.0	41.2	40.8	35.8	32.7	32.9
	FTdoc_maskAll	24.4	42.8	41.3	34.2	33.7	31.9
	FTdoc_maskFuture	26.5	44.7	42.2	26.3	33.9	22.8
	FTdoc_maskPast	25.0	43.5	41.2	34.7	33.4	32.8
BP	FTdoc	0.984	0.984	1.0	0.961	0.983	0.981
	FTdoc_maskAll	0.991	0.976	1.0	1.0	0.981	0.95
	FTdoc_maskFuture	0.975	0.967	1.0	0.877	0.983	0.727
	FTdoc_maskPast	1.0	0.979	1.0	0.979	0.988	0.98
COMET	FTdoc	0.649	0.778	0.838	0.777	0.814	0.798
	FTdoc_maskAll	0.658	0.781	0.842	0.755	0.818	0.779
	FTdoc_maskFuture	0.683	0.789	0.844	0.666	0.818	0.652
	FTdoc_maskPast	0.660	0.784	0.842	0.768	0.817	0.798

Table 9. Sentence-level BLEU score of Transformer with attention factor on THE_doc test set, with factor value from 0.1 to 0.9.

Score	**Factor**	**All**	**Future**	**Past**
BLEU	0.1	34.2	29.1	35.4
	0.2	34.3	28.3	35.1
	0.3	33.8	30.1	35.6
	0.4	32.8	30.2	35.2
	0.5	33.5	31.9	35.5
	0.6	34.9	33.6	35.4
	0.7	34.4	34.5	35.6
	0.8	35	34.7	35.6
	0.9	34.9	35.6	35.6
BP	0.1	0.982	0.935	0.982
	0.2	0.983	0.926	0.981
	0.3	0.952	0.893	0.998
	0.4	0.934	0.912	0.981
	0.5	1	0.959	0.975
	0.6	0.973	0.964	0.979
	0.7	1	0.993	0.992
	0.8	0.98	0.993	0.967
	0.9	0.977	0.993	0.992

and we tested its value from 0.1 to 0.9. Evaluation results on THE_doc and TAL_doc are in Table 9 and 10.

The BLEU score only fluctuated within a small range when changing the value of attention factor of the past sentences. In contrast, it steadily increased when augmenting factor value for context of the future. Interestingly, when masking or reducing all the source context, the model performance is between the case of modifying past and future context.

These results demonstrated the importance of future sentences in contextualized document translation. The past sentences can play a role to assist the target sequence generation, while when the future is unknown, full past context gives more noise than helpful information.

Table 10. Sentence-level BLEU score of Transformer with attention factor on TAL_doc test set, with factor value from 0.1 to 0.9.

Score	**Factor**	**ALL**	**Future**	**Past**
BLEU	0.1	30.2	22.3	32.7
	0.2	30.7	23.6	32.7
	0.3	30.9	23.9	32.4
	0.4	31.3	27.2	32.3
	0.5	32.6	32.1	32.5
	0.6	32.4	32.7	32.6
	0.7	32.6	33.2	32.2
	0.8	33.3	33.2	32.7
	0.9	33	32.7	32.5
BP	0.1	0.915	0.783	0.989
	0.2	0.941	0.78	0.984
	0.3	0.934	0.743	0.987
	0.4	0.927	0.796	0.981
	0.5	0.978	0.942	0.988
	0.6	0.982	0.97	0.985
	0.7	0.986	0.985	0.984
	0.8	0.987	0.986	0.984
	0.9	0.987	0.991	0.988

4. Conclusion

Translating discourse phenomena, like consistency and coherence, needs contextual information from the full document. Nevertheless, it is delicate to efficiently incorporate long-term context. In this work, we discussed current approaches related to this topic, then explored the impact of past, future and all source contexts through an ablation study. Due to the lack of publicly released parallel documents in scientific fields, we have created a such dataset and two test sets with EN-FR abstracts crawled from the web for training and evaluation. We discovered that future context plays an important role for both the translation quality and the length, while the past context may be helpful to mitigate search error at inference time. Additionally, with our setup, the

Transformer architecture performs better at the beginning of a long document than at its end.

This exploration is quite preliminary to investigate the characteristics of DMT with full documents, more experiments need to be completed in future work. For instance, testing the results on a larger dataset and on scientific parallel corpora from other fields. Recent contextualized metrics should also be applied to estimate how the DMT models make use of contexts. We also hope to improve model architecture or training techniques to improve translation quality with these characteristics of full document translation.

References

Abdul Rauf, S. and Yvon, F. Document level contexts for neural machine translation. Research Report 2020-003, LIMSI-CNRS, December 2020. URL https://hal.science/hal-03687190.

Agrawal, R., Turchi, M., and Negri, M. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pp. 31–40, Alicante, Spain, May 2018. URL https://aclanthology.org/2018.eamt-main.1.

Bao, G., Zhang, Y., Teng, Z., Chen, B., and Luo, W. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3442–3455, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.267. URL https://aclanthology.org/2021.acl-long.267.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. 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)*, pp. 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118. URL https://aclanthology.org/N18-1118.

Cettolo, M., Girardi, C., and Federico, M. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pp. 261–268, Trento, Italy, May 28–30 2012. European Association for Machine Translation. URL https://aclanthology.org/2012.eamt-1.60.

Fernandes, P., Yin, K., Neubig, G., and Martins, A. F. T. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6467–6478, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.505. URL https://aclanthology.org/2021.acl-long.505.

Fernandes, P., Yin, K., Liu, E., Martins, A., and Neubig, G. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 606–626, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.36.

Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. *Parallel corpora for medium density languages*, pp. 247–258. 01 2007. doi: 10.1075/cilt.292.32var.

Herold, C. and Ney, H. Improving long context document-level machine translation, 2023a.

Herold, C. and Ney, H. On search strategies for document-level neural machine translation, 2023b.

Jiang, Y., Liu, T., Ma, S., Zhang, D., Yang, J., Huang, H., Sennrich, R., Cotterell, R., Sachan, M., and Zhou, M. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1550–1565, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.111. URL https://aclanthology.org/2022.naacl-main.111.

Kudo, T. Subword regularization: Improving neural network translation models with multiple subword candidates, 2018.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Läubli, S., Sennrich, R., and Volk, M. 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*, pp. 4791–4796, Brussels, Belgium, October-

November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://aclanthology.org/D18-1512.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. URL https://api.semanticscholar.org/CorpusID:60827152.

Li, B., Liu, H., Wang, Z., Jiang, Y., Xiao, T., Zhu, J., Liu, T., and Li, C. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3512–3518, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.322. URL https://aclanthology.org/2020.acl-main.322.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 11 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343.

Lupo, L., Dinarelli, M., and Besacier, L. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4557–4572, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.312. URL https://aclanthology.org/2022.acl-long.312.

Ma, S., Zhang, D., and Zhou, M. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3505–3511, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.321. URL https://aclanthology.org/2020.acl-main.321.

Macé, V. and Servan, C. Using whole document context in neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.21.

Maruf, S. and Haffari, G. 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)*, pp. 1275–1284, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1118. URL https://aclanthology.org/P18-1118.

Maruf, S., Martins, A. F. T., and Haffari, G. 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)*, pp. 3092–3102, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1313. URL https://aclanthology.org/N19-1313.

Maruf, S., Saleh, F., and Haffari, G. A survey on document-level neural machine translation: Methods and evaluation, 2019b. URL https://arxiv.org/abs/1912.08494.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2947–2954, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://aclanthology.org/D18-1325.

Müller, M., Rios, A., Voita, E., and Sennrich, R. 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*, pp. 61–72, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6307. URL https://aclanthology.org/W18-6307.

Nguyen, M. V., Lai, V. D., Pouran Ben Veyseh, A., and Nguyen, T. H. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 80–90, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.10. URL https://aclanthology.org/2021.eacl-demos.10.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Post, M. and Junczys-Dowmunt, M. Escaping the sentence-level paradigm in machine translation, 2023.

Ranasinghe, T., Orasan, C., and Mitkov, R. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5070–5081, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.445. URL https://aclanthology.org/2020.coling-main.445.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Roussis, D., Papavassiliou, V., Prokopidis, P., Piperidis, S., and Katsouros, V. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2652–2657, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.284.

Salesky, E., Darwish, K., Al-Badrashiny, M., Diab, M., and Niehues, J. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pp. 62–78, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.iwslt-1.2.

Stahlberg, F. and Byrne, B. On NMT search errors and model errors: Cat got your tongue? 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)*, pp. 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331.

Sun, Z., Wang, M., Zhou, H., Zhao, C., Huang, S., Chen, J., and Li, L. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3537–3548, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.279. URL https://aclanthology.org/2022.findings-acl.279.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey, 2022.

Tiedemann, J. and Scherrer, Y. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 82–92, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4811. URL https://aclanthology.org/W17-4811.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420, 2018. doi: 10.1162/tacl_a_00029. URL https://aclanthology.org/Q18-1029.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. https://arxiv.org/abs/1706.03762, 2017. URL https://arxiv.org/abs/1706.03762.

Voita, E., Sennrich, R., and Titov, I. 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)*, pp. 877–886, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1081. URL https://aclanthology.org/D19-1081.

Voita, E., Sennrich, R., and Titov, I. 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*, pp. 1198–1212, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116. URL https://aclanthology.org/P19-1116.

Yin, K., Fernandes, P., Pruthi, D., Chaudhary, A., Martins, A. F. T., and Neubig, G. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 788–801, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.65. URL https://aclanthology.org/2021.acl-long.65.

Yu, L., Sartran, L., Stokowiec, W., Ling, W., Kong, L., Blunsom, P., and Dyer, C. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360, 2020. doi: 10.1162/tacl_a_00319. URL https://aclanthology.org/2020.tacl-1.23.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 533–542, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1049. URL https://aclanthology.org/D18-1049.

Zhang, P., Chen, B., Ge, N., and Fan, K. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1081–1087, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.81. URL https://aclanthology.org/2020.emnlp-main.81.