Investigating Length Issues in Document-level Machine Translation

Ziqian Peng[1,2], Rachel Bawden[2], François Yvon[1],

[1]Sorbonne Université & CNRS, ISIR, Paris, France,
[2]Inria, Paris, France,
{ziqian.peng,francois.yvon}@isir.upmc.fr rachel.bawden@inria.fr

Abstract

Transformer architectures are increasingly effective at processing and generating very long chunks of texts, opening new perspectives for document-level machine translation (MT). In this work, we challenge the ability of MT systems to handle texts comprising up to several thousands of tokens. We design and implement a new approach designed to precisely measure the effect of length increments on MT outputs. Our experiments with two representative architectures unambiguously show that (a) translation performance decreases with the length of the input text; (b) the position of sentences within the document matters, and translation quality is higher for sentences occurring earlier in a document. We further show that manipulating the distribution of document lengths and of positional embeddings only marginally mitigates such problems. Our results suggest that even though document-level MT is computationally feasible, it does not yet match the performance of sentence-based MT.

1 Introduction

Statistical and neural machine translation (MT) architectures (Koehn, 2020) have been designed to process isolated sentences, limiting their ability to properly handle discourse phenomena, such as coherence and cohesion, the modelling of which requires longer contexts (Fernandes et al., 2023). A first step to address this shortcoming has been to augment the source and/or the target side with a couple of preceding sentences (Tiedemann and Scherrer, 2017). Multiple approaches to encode and fully exploit such extended contexts have been proposed (Popescu-Belis, 2019; Maruf et al., 2021; Castilho and Knowles, 2024) and have been shown to improve the ability of MT engines to preserve local discourse coherence and cohesiveness through word-sense disambiguation or the resolution of anaphoric references (Bawden et al., 2018; Voita et al., 2018). Most of these approaches continue to process texts on a per-sentence basis with an extended context, even though attempts have also been made to process continuous chunks of texts comprising several sentences (Scherrer et al., 2019; Lopes et al., 2020; Ma et al., 2020, 2021; Lupo et al., 2022a; Wu et al., 2023).

The ability of today's neural MT models—relying on encoder-decoder or decoder-only architectures—to handle large context lengths, up to thousands of tokens (Peng et al., 2024), opens new perspectives to go beyond *context-augmented MT* and develop fully-fledged *document-level MT*, where the entire document context is available at once, and where the target text is generated in a single pass.[1] Two main technical novelties have made this possible: (a) more efficient computation in the attention layers (Tay et al., 2022) and (b) changes in the design of positional encodings (PEs). In particular, replacing the sinusoidal absolute PEs (APEs) of (Vaswani et al., 2017) with methods like ALIBI (Press et al., 2022) and RoPE (Su et al., 2024), which lend themselves well to length extrapolation (Sun et al., 2023; Zhao et al., 2024), seems to make today's transformers amenable to the processing of arbitrarily long contexts (Mohtashami and Jaggi, 2023; Han et al., 2024).

In this work, we challenge the ability of contemporary MT models to effectively handle long spans of texts. For this, we develop a new methodology for assessing the impact of length variations on MT performance. We perform a series of controlled experiments with two representative neural MT systems, where the same documents are processed by chunks of increasing lengths in a document-level manner and show that (a) MT performance tends

[1]These perspectives are, for instance, explored in the latest edition of the WMT shared task on General Machine Translation (Kocmi et al., 2024).

to degrade with the length of the source document, (b) length issues happen even for in-distribution lengths and get worse when extrapolating to unseen document lengths, and (c) most of the degradation happens in the final parts of the translation. Hypothesising that this may be due to a mismatch between the distribution of train and test PEs, we explore a possible mitigation, which flattens the distribution of PEs during training. We observe a consistent improvement of automatic metric scores for the APE-based vanilla encoder-decoder model, while the RoPE based decoder-only model remains mostly unaffected. In summary, our main contributions are: (a) a new approach to the detection and diagnosis of length issues in document-level MT, (b) a new variant of the SHAPE (Kiyono et al., 2021) method, which improves the distribution of PEs during training, and (c) a confirmation that (perhaps for lack of an appropriate document-level evaluation tool) sentence-level MT remains a strong baseline in most settings.

2 Related work

2.1 Document-level MT

Previous attempts to incorporate more contextual information in MT models can be roughly categorized into two categories: context-augmented MT, also called *Doc2Sent* in (Sun et al., 2022), and document-level MT, also called *Doc2Doc*. Recent surveys of this field include (Popescu-Belis, 2019; Maruf et al., 2021; Castilho and Knowles, 2024).

Translation of discourse phenomena, such as lexical consistency, reference, and word sense disambiguation, requires inter-sentential context (Bawden et al., 2018; Wong et al., 2020; Fernandes et al., 2023). This has motivated the integration of extended (local) contexts in *Doc2Sent* models. Such approaches include concatenation-based methods (Tiedemann and Scherrer, 2017); architecture adaptations to process context in different components of the same encoder (Ma et al., 2020; Wu et al., 2023), in a dedicated encoder (Voita et al., 2018; Zhang et al., 2018), or via hierarchical attention networks (Miculicich et al., 2018; Maruf et al., 2019; Yin et al., 2021); cache-based methods using a short-term MT memory (Maruf and Haffari, 2018; Tu et al., 2018; Yang et al., 2019; Dobreva et al., 2020) and multi-pass decoding algorithms (Voita et al., 2019; Yu et al., 2020; Kang et al., 2020).

Translating sentence by sentence, even with augmented contexts, still fails to capture phenomena related to coherence and consistency (Fernandes et al., 2023), motivating *Doc2Doc* approaches to process documents as a whole. This can be done with concatenation-based methods (Tiedemann and Scherrer, 2017; Sun et al., 2022; Karpinska and Iyyer, 2023), along with sliding window attention (Zhuocheng et al., 2023; Liu et al., 2023) and group attention (Bao et al., 2021) to address the issue of quadratic complexity. Other strategies include focusing on improving training through data augmentation with a balanced length distribution (Sun et al., 2022) and richer context-dependent phenomena (Lupo et al., 2022a; Wu et al., 2024), or on better training strategies with multilingual denoising pretraining (Lee et al., 2022), adapted loss functions (Lupo et al., 2022b), and enriched positional encodings (Li et al., 2023; Lupo et al., 2023). Multiple methods have recently emerged for large language models (LLMs) (Wang et al., 2023), which also show a decline in translation quality as input length increases (Wang et al., 2024).[2] These include a two-stage training recipe with the use of a monolingual corpus and high-quality parallel documents (Xu et al., 2024; Alves et al., 2024), and applying LLMs as post-editors (Koneru et al., 2024).

2.2 Extrapolating PEs

Since self-attention is position-agnostic, PEs are used to provide position information in Transformer models. PEs embed the absolute token position (APEs) (Vaswani et al., 2017), or the relative distance between tokens (RPEs) (Shaw et al., 2018; Raffel et al., 2020; Press et al., 2022), with RoPE (Su et al., 2024) being the go-to approach in recent LLMs such as Llama2 (Touvron et al., 2023). Despite RPEs yielding better length extrapolation ability than APEs, both of them struggle to efficiently extrapolate input lengths beyond the predefined maximum training length (Dai et al., 2019; Chen et al., 2023; Peng et al., 2024; Zhao et al., 2024), motivating the development of input extension methods for PEs.

For APEs, SHAPE (Kiyono et al., 2021) offsets all indices in a sequence by some random values. Its authors show that this simple technique mimics the computation of RPEs at a much smaller cost and helps to improve the interpolation abilities of a vanilla encoder-decoder model, as measured by BLEU (Papineni et al., 2002) with long pseudo-documents. Our experiments confirm that

[2] They also confirmed the effectiveness of training LLMs on documents of varied sizes (similar to Sun et al. (2022)).

this technique is effective using actual document contexts and a sounder experimental methodology, based on paired tests, and using COMET (Rei et al., 2020). Sinha et al.'s (2022) experiments adopt a setting similar to ours, offsetting the absolute value of APEs' input to evaluate their ability to capture relative distances between tokens. Their results, like ours, illustrate the lack of robustness of APEs and suggest that they overfit their training data.

For RPEs, especially RoPE, both position interpolation (PI) and position extrapolation methods have been proposed. PI methods interpolate positions to extrapolate context length directly during inference or through fine-tuning (Chen et al., 2023; Peng et al., 2024). The position extrapolation methods aim to extend context using documents that are shorter than the predefined maximum length. For example, RandPos (Ruoss et al., 2023) randomly maps position indices to a much larger interval with the original word order, and PoSE (Zhu et al., 2024) divides each training sequence into N chunks and adjusts the position indices of every chunk except the first one by adding a uniformly sampled offset, within the scope of a predefined maximal length.

3 Methods and Metrics

3.1 Holistic Document-Level MT

Compared to sentence-based MT, holistic document-Level MT (Doc2Doc) possesses several appealing features, as it gives access to all the available textual context. This should enable the MT system to improve on global aspects pertaining for instance to coherence and cohesion. However, Doc2Doc also introduces several new challenges compared to the Sent2Sent scenario:

1. in Doc2Doc, input texts are longer, causing a computational overhead due to the quadratic complexity of attention (Tay et al., 2022).
2. for longer inputs, attention weights are spread over a larger number of tokens (Herold and Ney, 2023); however, at each decoding step, most attention needs to remain concentrated on the corresponding local source context (Bao et al., 2021). This is in contrast with Doc2Sent, where sentence alignment is readily available.
3. decoding longer sequences increases the impact of search errors and of exposure bias (Ranzato et al., 2016). Beam search also becomes more difficult due to the input length.
4. output sentences may not always stand in one-to-one correspondence with source sentences, which complicates the computation of automatic metrics, which are designed to evaluate one-to-one mappings between hypotheses and references.

These differences motivate our main research questions, which we rephrase as: (a) For existing models, does *Doc2Doc* bring more benefits than disadvantages compared with *Sent2Sent*? (b) How do these results vary with the input document length? (c) Which methods and metrics can we use to automatically evaluate the impact of length differences?

3.2 Shades of BLEU

Answering such questions requires metrics for comparing holistic translations with sentence-based translations: as the number of segments produced by the former may differ from the number of source segments, a basic requirement is that they allow the evaluation of translation hypotheses with more (or fewer) sentences than the source (for quality estimation scores) and/or the reference (for reference-based metrics). However, most existing document-level MT approaches still rely on BLEU (Papineni et al., 2002), despite its well-documented shortcomings (Callison-Burch et al., 2006; Reiter, 2018; Mathur et al., 2020; Dahan et al., 2024); or rather a variant dubbed *d-BLEU* by Liu et al. (2020).[3] We accordingly focus on BLEU in this section, noting that the same questions would need to be addressed with any metric relying on sentence-based surface comparison (e.g., METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), BertSCORE (Zhang et al., 2020), PRISM (Thompson and Post, 2020), COMET (Rei et al., 2020), and many others).[4]

BLEU is computed by counting, sentence by sentence, the number of n-grams (for $n \in [1:4]$) shared by each translation hypothesis and its human reference. These counts are aggregated and turned into frequencies, then averaged (geometrically) at the corpus level. Finally, a length penalty

[3] Hendy et al. (2023) also consider a variant of COMET (Rei et al., 2022) while Zhuocheng et al. (2023) introduce d-ChrF, a document-level version of ChrF (Popović, 2015).

[4] We choose to evaluate using the standard metrics, BLEU and COMET, rather than evaluation approaches specifically designed to test the use of increased context. This choice is motivated by the fact that the score differences we observe reveal a significant degradation in translation quality for longer documents, indicating greater problems than those targeted by finer-grained evaluation techniques.

is applied to degrade the score when the cumulated length of the hypotheses is shorter than that of the references. BLEU is a corpus-level score that depends on sentence alignments. d-BLEU is also a global score but counts common n-grams at the document level. As a consequence, d-BLEU, which records matches for larger spans than BLEU, delivers higher scores, as the opportunities to match n-grams are greater for a wider window.[5] These two scores cannot be compared, and we contend that their shortcomings make them inappropriate for analysing length-related issues in MT.

An alternative to d-BLEU is to perform evaluation at the document level, rather than the corpus level. This can be implemented either as (a) calculating one BLEU score (with realignment) per document, then averaging at the corpus level or (b) calculating the equivalent of sentence-level BLEU scores (Lin and Och, 2004) but where each segment is a concatenated full document rather than a sentence. However, (a) counts matches at the sentence-level, which requires a realignment between translated and reference sentences and may introduce some measurement noise. Therefore, our experiments use method (b) to compute document-level scores, hereafter referred to as **ds-BLEU** scores.

3.3 Evaluating Length Issues in MT

Another recurring methodological caveat with length-related evaluation is related to the way scores are compared. For instance, in (Sun et al., 2022, Figure 1) BLEU scores are reported for buckets of sentences of varying lengths in a plot which suggests that performance increases with length (up to a certain extent). Such vizualisations are misleading, as global BLEU scores should only be compared when measured with the same corpus.

What we propose instead is to compare matching automatic translation scores for a set of inputs $\mathcal{S} = \{s_1 \dots s_T\}$, systematically varying the translation models M in $\{M_1 \dots M_N\}$ and the length of the translation window $W \in \{W_1 \dots W_K\}$. For each pair of settings, we can perform a paired t-test for the average score difference and decide whether two configurations (M_i, W_k) and (M_j, W_l), each associating a system and a length, are statistically different, and if so, which of the two is the best.

[5]This effect is well known, e.g. in (Koehn and Knowles, 2017, Figure 7), where BLEU increases when considering sentence groups of increasing lengths (at least for a certain length range), where we would expect a decrease, as the length is often linked to syntactic complexity and therefore to translation difficulty. We reproduce this observation in Figure 2.

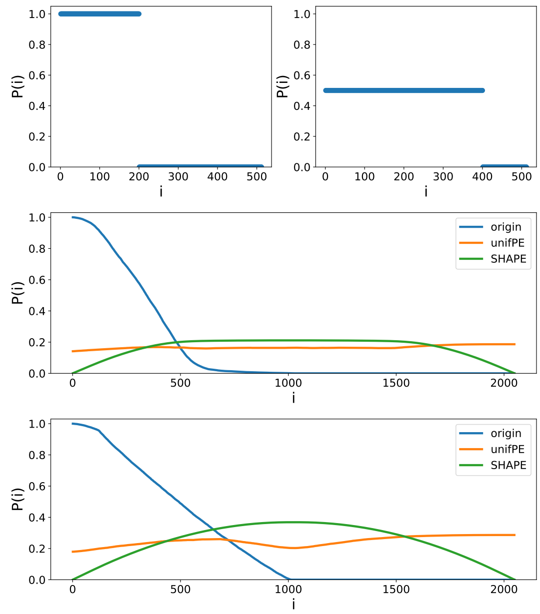

Figure 1: Top: probability of observing training position i ($P(i)$) for a sentence of length $l = 200$, with standard training ($k_i = 0$, left) and with our uniform sampling scheme (right) for $M = 512$. Middle: original, UNIFPE, and SHAPE $P(i)$ for training set **TED-G** and $M = 2048$. Bottom: original, UNIFPE and SHAPE $P(i)$ for **TED-U** and $M = 2048$.

In our experiments, we consider two ways of presenting $\mathcal{S}$: (a) at the document level, where each s_i is a document and the evaluation is the ds-BLEU score introduced in Section 3.2,[6] and (b) at the sentence level, where each s_i is a sentence and the associated metric is COMET (Rei et al., 2020).[7] For (b) we need to realign translation hypotheses with their references. This can be performed with the method of Wicks and Post (2022),[8] or with that of Matusov et al. (2005),[9] which has long been used for evaluating speech translation systems, and which we adopt.[10] Variations in configurations (M, W) are obtained by changing the translation engine and the length of input source texts. In all cases, score comparisons are performed on identical source texts.

[6]We use SacreBLEU (Post, 2018) with signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0; the parameter *eff* is set to yes for ds-BLEU.

[7]Using the library `https://github.com/Unbabel/COMET` with the default model `wmt22-comet-da`.

[8]Junczys-Dowmunt (2019)'s approach includes a set of tags that constrain input and output to have the same number of sentences, see also (Li et al., 2023).

[9]`https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz/`.

[10]The per-sentence COMET scores are averaged at the document level to be associated with document lengths, or at the corpus level to assess global translation quality.

The same technique is also used to measure the impact of the position within a document on translation quality. The question we study is whether quality remains constant across a document, or whether it tends to decrease when sentences are processed at higher position indices. For this, we consider groups of sentences translated at varying starting positions with multiple systems and compare the differences between COMET scores with a paired difference test. Details regarding the corpus and window sizes are given in Section 5.1.

4 Manipulating the Distribution of PEs

A basic requirement for document-level systems is that they should be trained, or at least fine-tuned, with long text inputs, ideally with complete documents. Using the empirical document length distribution may however not be ideal, as it yields very skewed distributions of PEs where small position indices are over-represented. We discuss two approaches to obtain more balanced distributions.

4.1 Distribution of PEs

A training sequence of length l yields examples for all indices in $\{1, \ldots, l\}$. For a complete corpus, position index 1 will be observed for all inputs, while the last index of the longest sequence will likely only be observed once. Training with the "natural" distribution of document lengths is therefore likely to overfit to smaller position indices while underfitting to larger ones, hindering the ability to handle long texts or extrapolate to lengths unseen in training (Peng et al., 2024; Zhu et al., 2024).

A first way to improve the distribution of token positions seen in training is to increase the representation of long documents in the training data while keeping a good balance with shorter ones (Bao et al., 2021; Sun et al., 2022). This is easy to do in our controlled setting (see Section 5.1). As our experiments show, this significantly improves automatic scores for the context lengths seen during training. An alternative, which allows us to better study the effect of PE distributions in training, is to directly manipulate the indices (for a fixed length distribution). The UNIFPE algorithm, introduced below, is one way to achieve this.

4.2 Uniform SHAPEs (UNIFPE)

We assume a training set of texts $s_1 \ldots s_N$ of respective lengths $l_1 \ldots l_N$, and a maximum model length of M, with $\forall i, M > l_i$. Training with text s_i creates training samples for positions i in $[1\!:\!l_i]$. For the whole corpus, positions from 1 (observed N times) to $l_{max} = \max_{i=1 \ldots N}(l_i)$ are observed, with larger indexes being less trained than smaller ones. Positions indices in $[l_{max} : M]$ are never observed. We wish to make the training PE distribution more even, so that all positions in $[1 : M]$ are equally well-trained, which should also help to extrapolate PEs for indices larger than l_{max}.

This can be achieved by shifting the starting index of every s_i by some offset k_i, making it possible to train with PEs in $[1 + k_i : l_i + k_i]$. How should k_i be chosen? Randomly choosing $k_i = 0$ or $k_i = l_i$ with probability $1/2$ makes the probability of observing any index in $[1\!:\!2l_i]$ equal to $1/2$. This can be generalised to choose k_i with uniform probability $1/m$ among $\{0, l_i, \ldots, (m-1) * l_i\}$, with $m = \lfloor M/l_i \rfloor$. However, doing so implies that indices in $[m * l_i : M]$ are never observed. We compensate for this as follows: before sampling k_i, we modify the set of possible shifts by adding $r_i = M - m * l_i$ to all values larger than a random index $l \in [1 : m]$. In other words, k_i is sampled from $\{j * l_i + r'_{i,j}, j = 0 \ldots m-1\}$, with $r'_{ij} = 0$ if $j < l$ and r_i otherwise. Sampling k_i independently for each text s_i in each training batch ensures that all indices are uniformly represented. A formal description of UNIFPE is given in Algorithm 1. Figure 1 illustrates the difference between always starting at position 1 ($\forall i, k_i = 0$) and using our UNIFPE strategy.

This approach is reminiscent of SHAPE (Kiyono et al., 2021); while SHAPE chooses the offset k_i uniformly at random in a fixed interval to simulate relative PEs, which reduces the frequency of small position indices, we sample k_i non-uniformly to ensure that all indexes are equally represented in training.

5 Experimental Settings

5.1 Datasets

For our experiments, we prepare multiple sets of parallel pseudo-documents based on the EN–FR part of the TEDtalks corpus (Cettolo et al., 2012).

Training and validation sets Our training set consists of pseudo-documents from both the training and validation splits of **IWSLT-2016**.[11] Our goal is to simulate real corpora of parallel documents with source documents shorter than a certain

[11] https://wit3.fbk.eu/2016-01

	TED-full		TED-G		TED-U	
	train	dev	train	dev	train	dev
Count	1831	19	15625	160	10582	106
Length	2915	2861	341	339	504	512

	sent	256	512	768	1024	1200	1600	2048	doc
Count	5103	503	261	184	142	123	100	80	52
Length	23	233	450	638	827	955	1175	1468	2259

Table 1: Left: Statistics of the TED talks training and dev sets. Right: Statistics of the TED talks test sets from IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017**. 'Count' denotes the number of parallel pseudo-documents, 'Length' denotes the average length of source (i.e. English) pseudo-documents (in NLLB tokens).

	l_{max}	2014	2015	2016	2017
NLLB	sent	45.1 (0.97)	43.9 (0.98)	41.7 (1.00)	41.8 (1.00)
	256	33.9 (0.82)	35.4 (0.84)	33.3 (0.86)	33.5 (0.87)
	512	14.6 (0.44)	16.0 (0.56)	15.2 (0.52)	13.8 (0.49)
	768	7.3 (0.27)	7.9 (0.32)	10.0 (0.46)	6.7 (0.27)
	1024	8.8 (0.56)	7.4 (0.51)	7.5 (0.50)	6.5 (0.48)
TowerBase	sent	43.4 (0.98)	42.9 (0.99)	39.7 (1.00)	38.7 (1.00)
	256	44.0 (0.96)	42.8 (0.98)	40.9 (1.00)	39.4 (1.00)
	512	42.9 (0.96)	39.8 (0.98)	39.9 (1.00)	40.6 (1.00)
	768	39.6 (0.98)	39.0 (0.97)	38.1 (0.99)	39.9 (1.00)
	1024	38.5 (0.98)	33.1 (0.99)	35.4 (1.00)	35.4 (0.98)
	1200	37.4 (0.92)	35.5 (0.98)	36.2 (1.00)	35.6 (0.98)
	1600	33.3 (0.96)	34.9 (0.96)	26.7 (0.94)	31.0 (0.97)
	2048	24.0 (0.97)	27.7 (0.95)	27.2 (0.96)	23.5 (0.87)

Table 2: ds-BLEU scores (and brevity penalty) for NLLB200-DISTILLED-600M and TOWERBASE-7B.

length l_{max} – using $l_{max} = 1024$. We split all document pairs whose source side is longer than 1024 tokens into fragments.[12] For each document pair, we iterate the following procedure: (1) sample a maximum pseudo-document length l'_i following the same Gaussian-like length distribution as the full TED talks with $l'_i < l_{max}$, (2) concatenate consecutive sentence pairs up to l'_i to form a training pseudo-document s_i. The resulting distribution of document lengths is displayed in Figure 3 in Appendix A.3. The development set is built similarly, using document pairs from IWSLT **tst2010** and **tst2011**. We denote these training datasets as **TED-G** (G for Gaussian). As discussed in Section 4, we consider another dataset generation strategy, which produces a more balanced length distribution, for which we do as above but we sample uniformly: $l'_i \sim U(128, l_{max})$.[13] Fine-tuning with the resulting **TED-U** corpus allows us to contrast two distributions with differences in document length.

Test sets To evaluate MT systems for their ability to handle documents of varying sizes and extrapolate beyond the training samples, we build a series of test sets of increasing document lengths. For each document in IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017**, we accumulate consecutive sentence pairs into parallel pseudo-documents such that all resulting source texts have a length close to l_{max}, with $l_{max} \in \{256, 512, 1024, 1200, 1600, 2048\}$.[14] Contrarily to training sets, test sets are homogeneous in length. Statistics are in Table 1 with more details in Appendix A.3. Evaluation is always performed with complete original talks, after concatenating and aligning all the corresponding parts.

5.2 Models

We used the UNIFPE algorithm to fine-tune two pre-trained MT systems that were not trained with TED talks. As UNIFPE is designed for APEs, we considered NLLB200-DISTILLED-600M[15] or NLLB for short (Costa-jussà et al., 2024) as a representative encoder-decoder model based on APEs. NLLB is a 12-layer encoder-decoder multilingual MT model pre-trained on 200 languages. We used the HuggingFace implementation, which relies on sinusoidal APEs (Vaswani et al., 2017). We also perform fine-tuning with SHAPE for comparison. We refer to the specific MT systems with respect to their fine-tuning method (FT, UNIFPE or SHAPE), backbone model (e.g. NLLB) and training corpus (U for **TED-U** or G for **TED-G**. More precisely, we denote MT systems trained on **TED-U** (resp. **TED-G**) as FT-NLLB-U (resp. FT-NLLB-G), UNIF-NLLB-U (resp. UNIF-NLLB-G) when fine-tuning with UNIFPE, and SHAPE-NLLB-U (resp. SHAPE-NLLB-G).

We also experiment with an LLM-based architecture, TOWERBASE-7B[16] (Alves et al., 2024) (TOWERBASE for short), derived from Llama2 (Touvron et al., 2023) using translation-related

[12] All statistics counted in tokens use the tokeniser of NLLB (Costa-jussà et al., 2024).

[13] Short pseudo-documents continue to be slightly over-represented, because the last pseudo-document in any given talk is often strictly shorter than the desired length l'_i.

[14] At the end of each talk, we concatenate the last parallel sentences into the last pseudo-document if they are shorter than 50 to avoid exceedingly short parallel sequences.

[15] https://huggingface.co/facebook/nllb-200-distilled-600M

[16] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

	NLLB	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
sent-256	9.2	0.8	-2.1	-2.5	0.4	-0.7	-1.7
256-512	19.1	-	-0.4	1.4	-	-	2.1
512-768	6.9	-	-0.6	-	5.9	2.5	5.3
768-1024	-	0.5	-	-	7.2	3.0	3.7
1024-1200	2.2	3.5	1.9	4.1	4.0	3.3	4.1
1200-1600	-	6.8	6.5	5.4	5.8	5.5	4.9
1600-2048	1.9	5.2	4.5	3.1	4.2	5.7	6.0
sent-256	16.7	3.5	1.7	1.3	2.7	2.1	1.9
256-512	20.7	-	-0.4	2.4	0.6	-	4.2
512-768	5.6	-	-	-	11.6	7.2	8.1
768-1024	5.2	2.3	0.8	-	10.4	7.6	7.8
1024-1200	-	7.4	4.3	6.9	3.8	4.7	4.6
1200-1600	6.1	9.6	13.4	8.3	5.4	5.5	5.6
1600-2048	-	5.1	5.0	5.9	3.9	5.6	5.0

	TOWER	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
sent-256	-	-	-	-	-	-	-
256-512	-	0.9	0.8	0.8	0.6	0.6	0.5
512-768	-	-	-	-	0.6	-	-
768-1024	3.4	-	1.0	1.2	1.7	1.2	2.1
1024-1200	-	-	-	-	-	-	-
1200-1600	4.7	1.7	2.1	-	1.6	2.0	1.5
1600-2048	5.9	7.5	6.5	7.3	8.1	7.8	7.3
sent-256	3.9	2.3	2.4	2.3	2.3	2.3	2.2
256-512	-	0.4	-	-	0.2	0.3	0.3
512-768	-	-	-	0.5	0.3	-	-
768-1024	2.9	1.0	0.5	-	1.1	0.8	1.2
1024-1200	-	-	1.0	0.9	-	-	-
1200-1600	6.2	1.7	1.8	-	-	1.9	1.8
1600-2048	8.7	10.0	8.9	9.1	11.0	10.2	9.2

Table 3: Average differences evaluated on **ds-BLEU** (top) of full TED talks and on $100\times$**COMET** (bottom) of realigned parallel sentences, between translations in increasing context size, for NLLB (left) and TOWERBASE (right) models. U and G respectively denote **TED-U** and **TED-G**. A positive value means that shorter segments result in higher scores than longer ones. Text in olive for p-values > 0.01. - for p-values > 0.05.

tasks. TOWERBASE uses RoPE (Su et al., 2024) to encode RPEs. As mentioned by Peng et al. (2024), they nonetheless encode some form of APE signal in some dimensions, and may therefore be also mildly impacted by the PE training distribution. We refer to the models based on TowerBase as FT-TOWER-U (resp. FT-TOWER-G), UNIF-TOWER-U (resp. UNIF-TOWER-G) and SHAPE-TOWER-U (resp. SHAPE-TOWER-G) the model fine-tuned on **TED-U** (resp. **TED-G**) with original PEs, UNIFPE or SHAPE.

Both backbone models were pretrained with large amounts of EN–FR data; we focus exclusively on the EN into FR direction. Details on fine-tuning and decoding parameters can be found in Appendix A.4.

6 Results and Analyses

6.1 Length Issues

We report the ds-BLEU (Table 2) and COMET (Appendix, Table 7) scores of the pretrained models NLLB and TOWERBASE for multiple test sets, varying the average input segment lengths from one sentence to the maximum input length used in training.[17] For NLLB, we observe a drop of around 10 ds-BLEU points and about 0.2 COMET points when translating test sets of $l_{max} = 256$ instead of isolated sentences. Scores and their associated brevity penalties (BPs) only get worse with larger context lengths. For TOWERBASE, the decrease in BLEU is more progressive, with a sharp decline for all test sets for $l_{max} > 1024$. The related COMET scores plummet immediately with a context size of 256. Even though TOWERBASE is based on Llama2, which accepts inputs up to 4096 tokens, the continued pretraining that was used mostly uses isolated sentences, which introduces an inductive bias affecting its ability to translate long texts.

As expected, document-level fine-tuning (DLFT) has a strong positive impact (see Appendix, Table 11). However, the length issues remain.

Length Bias We performed paired comparisons for the translation of our test sets with increasing text lengths for each MT system as presented in Section 3.3. Results are given in Table 3, where a positive difference (e.g. 9.2 for NLLB in line "sent-256") means that the translation of shorter segments (here: sentences) yield better scores than that of longer ones (256 tokens). Scores in the same column are comparable. Except for a handful of configurations, translating longer texts is never better than translating short ones. We conclude that in our experimental settings, the disadvantages associated with long inputs (Section 3.1) overwhelm the benefits of a complete context. These length issues result in large score degradations and are not easily fixed by simple manipulation of PEs. We also observe that results obtained with COMET and ds-BLEU sometimes disagree. These cases are rare, though, suggesting that our results are robust.

Document-level Tuning with UNIFPE Again using the paired comparison methodology, we compare the performance of DLFT with original PEs, UNIFPE and SHAPE. As shown in the left and middle parts of Table 4, fine-tuning using UNIFPE leads to steady improvements in translation scores for all test lengths, especially for systems fine-

[17] As explained in Section 3.2, these COMET scores require the realignment of target sentences with the reference.

	TED-U		TED-G		FT	Unif	SHAPE
	FT vs Unif	FT vs SHAPE	FT vs Unif	FT vs SHAPE	U vs G	U vs G	U vs G
sent	3.3 (0.00)	4.0 (0.00)	1.2 (0.00)	2.4 (0.00)	-	-2.1 (0.00)	-1.6 (0.00)
256	-	0.7 (0.00)	-	-	-	-0.6 (0.01)	-0.7 (0.00)
512	-0.5 (0.01)	1.7 (0.01)	-	2.3 (0.00)	-0.7 (0.00)	**-0.4 (0.01)**	-
768	-0.8 (0.00)	2.7 (0.00)	-3.7 (0.00)	-	5.5 (0.00)	2.6 (0.00)	4.5 (0.00)
1024	-	1.6 (0.00)	-7.8 (0.00)	-1.8 (0.04)	12.2 (0.00)	5.1 (0.00)	8.8 (0.00)
1200	-2.3 (0.00)	2.3 (0.02)	-8.5 (0.00)	-	12.7 (0.00)	6.5 (0.00)	8.8 (0.00)
1600	**-2.6 (0.01)**	-	-8.8 (0.00)	**-2.6 (0.01)**	11.8 (0.00)	5.6 (0.00)	8.3 (0.00)
2048	**-3.3 (0.00)**	-	-7.3 (0.00)	-	10.7 (0.00)	6.8 (0.00)	11.3 (0.00)
sent	1.9 (0.00)	2.9 (0.00)	0.9 (0.00)	1.4 (0.00)	-	-0.9 (0.00)	-1.4 (0.00)
256	-	0.8 (0.00)	**0.4 (0.00)**	**0.7 (0.00)**	**-0.8 (0.00)**	-0.5 (0.01)	-0.9 (0.00)
512	-0.6 (0.02)	2.9 (0.00)	-	4.3 (0.00)	-0.4 (0.03)	-	-
768	-0.7 (0.00)	4.7 (0.00)	-4.7 (0.00)	-	11.2 (0.00)	7.2 (0.00)	7.3 (0.00)
1024	**-2.2 (0.00)**	3.3 (0.00)	-7.5 (0.00)	-1.8 (0.02)	19.3 (0.00)	14.0 (0.00)	14.2 (0.00)
1200	-5.3 (0.00)	2.7 (0.02)	-6.5 (0.00)	-	15.7 (0.00)	14.5 (0.00)	11.9 (0.00)
1600	-	-	-6.4 (0.00)	-	11.5 (0.00)	6.6 (0.00)	9.2 (0.00)
2048	-	**2.1 (0.04)**	-4.7 (0.00)	-	10.4 (0.00)	7.2 (0.00)	8.3 (0.00)

Table 4: Average difference (and p-values) in **ds-BLEU** (top) evaluated on full TED talks and 100×**COMET** (bottom) evaluated on realigned sentences for NLLB. Left and middle: paired comparison between fine-tuning with the original PEs (FT), UNIFPE (Unif) and SHAPE on **TED-U** and **TED-G** respectively. Right: differences between fine-tuning on **TED-U** (U) and **TED-G** (G). - for p-values > 0.05. Bold values when the two metrics disagree on significativity.

	NLLB	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
p_0-p_1	12.7	-	-	4.0	1.6	3.0	5.0
p_1-p_2	7.2	-	-	-2.0	1.9	2.4	-
p_2-p_3	2.9	1.0	-1.3	-	2.4	-	-2.1
p_3-p_4	7.2	5.5	4.6	7.3	26.1	10.7	13.9
p_4-p_5	-	3.9	-	-	8.3	4.7	4.3
p_5-p_6	3.3	31.6	27.1	19.5	6.1	15.4	15.3

	TOWER	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
p_0-p_1	1.2	-	-	-	-	-	-
p_1-p_2	-	-	-	-	0.6	0.6	-
p_2-p_3	-	-	-	-	-	-	-
p_3-p_4	4.9	1.1	0.6	1.4	1.1	1.3	1.1
p_4-p_5	1.7	1.4	2.1	1.7	2.0	1.8	2.2
p_5-p_6	26.3	24.5	26.0	25.2	27.1	26.4	27.0

Table 5: Average difference of 100×COMET-score evaluated on 794 sentence pairs, translated at different positions (e.g. p_0 and p_1 with $p_0 < p_1$) by NLLB-based systems (top) and TOWERBASE-based systems (bottom). Olive text for p-values > 0.01. - for p-values > 0.05.

tuned with the unbalanced corpus (**TED-G**). The only exception is for sentence-level translations, which remain marginally better using standard DLFT than with UNIFPE. In contrast, SHAPE only improves DLFT performance on the **TED-G** corpus and for translation windows greater than 1024 tokens, due to the under-representation of small position indices during training, as shown in Figure 1. As Tables 11 to 15 show, these improvements remain moderate, and the length issues continue to strongly impact translation scores, especially for test documents of 1024 tokens or more. For TOWERBASE, UNIFPE does not yield any significant difference with standard DLFT, and SHAPE occasionally delivers slight improvements (see Appendix, Table 16), likely because this model relies on RPEs. From these comparisons, we conclude that UNIFPE partly resolves length issues for NLLB, but hardly changes the situation for TOWERBASE.

Impact of Data Distribution In the right part of Table 4, we evaluate the impact of the length distribution during fine-tuning for NLLB: the balanced distribution (**TED-U**) slightly but consistently underperforms the use of **TED-G** for short documents (fewer than 512 tokens), a trend that is reversed for longer documents with strong improvement (over 768 tokens). Manipulating the distribution of PEs with UNIFPE reduces the gap between the two fine-tuning corpora and makes the model more robust to document lengths rarely observed (or even unobserved) during fine-tuning. This analysis again reveals small differences between using ds-BLEU and COMET scores: in nine cases out of 56 comparisons (marked in bold), one metric detects a difference that is non-significant for the other.

6.2 Position Bias

To investigate potential translation issues related to large position indices, we collected the 794 sentences that come from the final part of long talks and for which varying the window length also varied the position index. For each of them, we have seven translations, corresponding to positions $\{p_0^j, \ldots, p_6^j\}, j \in \{1, \ldots, 794\}$. The av-

	256	512	768	1024	1200	1600	2048
NLLB	0.04	0.35	0.49	0.66	0.64	0.74	0.81
FT-U	0.01	0.03	0.08	0.09	0.13	0.26	0.44
Unif-U	0.01	0.03	0.07	0.10	0.20	0.34	0.32
SHAPE-U	0.03	0.08	0.11	0.20	0.36	0.39	0.46
FT-G	0.01	0.03	0.11	0.31	0.40	0.57	0.69
Unif-G	0.01	0.04	0.16	0.20	0.27	0.25	0.36
SHAPE-G	0.02	0.05	0.12	0.17	0.21	0.24	0.28

	256	512	768	1024	1200	1600	2048
TOWER	0.01	0.05	0.13	0.30	0.29	0.45	0.64
FT-U	0.01	0.05	0.10	0.15	0.13	0.23	0.59
Unif-U	0.02	0.05	0.09	0.15	0.16	0.28	0.61
SHAPE-U	0.02	0.05	0.08	0.15	0.17	0.21	0.59
FT-G	0.02	0.05	0.10	0.15	0.19	0.26	0.62
Unif-G	0.02	0.07	0.10	0.16	0.20	0.28	0.64
SHAPE-G	0.02	0.05	0.07	0.17	0.17	0.27	0.62

Table 6: Percentage of pseudo-documents among IWSLT **tst2014-2017** in which 10-gram repetition is detected in the translation given by NLLB-based (top) and TOWERBASE-based models (bottom).

erage values for $\{p_0^j, \dots, p_6^j\}$ are $\{p_0, \dots, p_6\} = [66, 173, 262, 335, 585, 779, 1477]$. For this subset of sentences, we performed a paired t-test to compare the impact of the position on the translation score (using COMET as the only metric). We observe in Table 5 that in almost all comparisons but three, a small position index is preferable to a larger one. This suggests that one of the main challenges faced by *Doc2Doc* with large context lengths is to control the quality degradation for the final parts of the input text. Here again, UNIFPE slightly mitigates the problem for NLLB models compared with original PEs and SHAPE, but no such improvement is observed for TOWERBASE.

6.3 Repeated n-grams in Translation

One obvious problem with holistic translations produced by NLLB is the generation of outputs that are too short. A closer look at translation outputs also reveals that outputs contain many instances of repeated texts, usually occurring in the final part of the translation. To quantify this problem, we compute the percentage of translations of pseudo-documents in which the repetition of a long n-gram (with $n \geq 10$) is detected. Detailed results are given in Table 6. For all systems and fine-tuning strategies, the percentage of repetitions increases with the length, a problem that seems (for large text lengths) slightly more severe for TOWERBASE, which has a much better BP, than NLLB.

7 Conclusion

In this work, we have studied the ability of current MT architectures to handle long input texts, ideally entire documents, and to translate them holistically. Our analyses are based on systematic comparisons of translation outputs computed with varying input lengths, which are then evaluated with two automatic metrics. They consistently show that, even when the test document lengths match that of the training set and remain within the model limits, the translation scores tend to decrease with the source length, a degradation that mostly impacts sentences occurring far from the beginning of the document. We also show that manipulating the training distribution of lengths or PEs has a positive effect for APE-based models, which vanishes in RoPE-based models like TOWERBASE. These results finally confirm the robustness of sentence-level baselines. They hint at the need to improve existing models to truly benefit from the potential of document-level MT, for instance by constraining the attention mechanism to simulate a form of sentence alignment, by improving the memorization capacities of existing architectures, or by ensuring that the generation algorithm does not eventually get trapped in repetition loops. These are some of the directions we wish to explore in future work.

8 Limitations

The empirical observations reported in this paper are based on one single language direction, and one domain (TEDtalks). This experimental design is motivated by (a) the fact that French-English is considered an easy pair for MT, with large sets of parallel training data; (b) TEDtalks data are a standard benchmark for document-level MT, and crucially contain very long parallel documents, allowing us to implement our evaluation methodology on a large range of length values. Furthermore, these datasets are not included in the training data of our models. We contend that the length issues observed in these favorable conditions for two representative systems would only be worse for more difficult or less-resourced language pairs.

9 Ethics Statement

This study has been performed with standard benchmarks and open-weight models. We do not see any ethical problems with this work.

10 Carbon Impact Statement

The experiments were conducted on a private infrastructure using a single A100 SXM4 GPU, with a carbon efficiency of 0.432 kgCO_2eq/kWh. The average time required for fine-tuning and checkpoint selection was 14.21 hours for the six NLLB models, and 5.6 hours for the TOWERBASE models. The average emissions are estimated to be 2.45 kgCO_2eq for NLLB-based models and 0.97 kgCO_2eq for models derived from TOWERBASE, with no offset applied. These estimations were based on the Machine Learning Impact calculator[18] (Lacoste et al., 2019).

Acknowledgments

This work was supported by the French national agency ANR as part of the MaTOS project.[19] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors are grateful to the anonymous reviewers for their insightful comments and suggestions and to Paul Lerner for his review and feedback on a preliminary draft of this work.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, pages 65–72, Ann Arbor, Michigan.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. 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)*, pages 3442–3455, Online. Association for Computational Linguistics.

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, New Orleans, Louisiana.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, page 1–31.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *Preprint*, arXiv:2306.15595.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. ISBN: 1476-4687.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level. Technical report, Inria Paris; ISIR-CNRS .

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Radina Dobreva, Jie Zhou, and Rachel Bawden. 2020. Document sub-structure in neural machine translation. In *Proceedings of the 12th Language Resources*

[18] `https://mlco2.github.io/impact#compute`

[19] `http://anr-matos.fr/`

and Evaluation Conference, pages 3657–3667, Marseille, France. European Language Resources Association.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. 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)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. LM-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008, Mexico City, Mexico. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. 2021. SHAPE: Shifted absolute position embedding for transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3309–3321, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet. In *Proceedings of the Eighth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

P. Koehn. 2020. *Neural Machine Translation*. Cambridge University Press.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. 2022. DOCmT5: Document-level pretraining of multilingual language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 425–437, Seattle, United States. Association for Computational Linguistics.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023. P-transformer: Towards better document-to-document neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:3859–3870.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Zewei Sun, Shanbo Cheng, Shujian Huang, and Mingxuan Wang. 2023. Only 5% attention is all you need: Efficient long-range document-level neural machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 733–743, Nusa Dua, Bali. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022a. 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)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022b. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 33–44, Dubrovnik, Croatia. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Zhiyi Ma, Sergey Edunov, and Michael Auli. 2021. A comparison of approaches to document-level machine translation. arXiv preprint arXiv:1910.07481.

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

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

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Lesly Miculicich, 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, Brussels, Belgium. Association for Computational Linguistics.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. *Preprint*, arXiv:1901.09115.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, Belgium, Brussels. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. 2023. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, Toronto, Canada. Association for Computational Linguistics.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes, and Adina Williams. 2022. The curious case of absolute position embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4449–4472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2023. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14590–14604, Toronto, Canada. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):1–28.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. 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 877–886, Hong Kong, China. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2022. Does sentence segmentation matter for machine translation? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level 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 1527–1537, Hong Kong, China. Association for Computational Linguistics.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F.T. Martins, and Graham Neubig. 2021. 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)*, pages 788–801, Online. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

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, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proc. International Conference on Learning Representations*.

Liang Zhao, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. 2024. Length extrapolation of transformers: A survey from the perspective of positional encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9959–9977, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556, Singapore. Association for Computational Linguistics.

A Appendix

A.1 The UNIFPE Algorithm

The UNIFPE algorithm briefly described in Section 4.2 is formalised in Algorithm 1.

Data: l_i: The input length
Data: M: The target max context length
Data: $\text{List}_{\text{p}_\text{k}}$: the distribution of p_k for each offset k in $[0, M - l_i]$
$\text{List}_{\text{p}_\text{k}} \longleftarrow$ Initialized to 0 for each element
$m \longleftarrow \lfloor M/l_i \rfloor$ nb. of possible non-zero p_k
$R_n \longleftarrow$ the remainder of M divided by l_i
$p_0 \longleftarrow \frac{1}{m}$the probability of each non-zero p_k
if $M < 2l_i$ **then**
  $\text{List}_{\text{p}_\text{k}} \longleftarrow p(k' = 0) = 1$ i.e. $p_{k=0} = 1$
else
  $k^* \longleftarrow$ a random integer in $[0, m)$
  for $k \in [0, M - l_i]$ **do**
    if $k\%l_i == 0$ and $k < k^*$ **then**
      $\text{List}_{\text{p}_\text{k}} \longleftarrow p(k' = k) = p_0$
    end
    if $(k - R_n)\%l_i == 0$ and $k^* < k \leq M - l_i$ **then**
      $\text{List}_{\text{p}_\text{k}} \longleftarrow p(k' = k) = p_0$
    end
  end
end
return $\text{List}_{\text{p}_\text{k}}$

Algorithm 1: UNIFPE: the pseudo-uniform position indices mapping algorithm.

A.2 A Call for Correctly Using BLEU Scores

As illustrated in Figure 2, d-BLEU and ds-BLEU are always larger than BLEU. When BLEU decreases due to the degradation of translation quality, d-BLEU remains stable because of the higher probability to find n-gram matches in longer sequences. In contrast, ds-BLEU consistently decreases when BLEU diminishes, as it applies a macro-average to compute the corpus-level score, which is more sensitive to the translation quality of each document than d-BLEU. Therefore, d-BLEU, ds-BLEU and BLEU are not comparable and d-BLEU is not suitable for analysing length issues in document-level evaluation of MT.

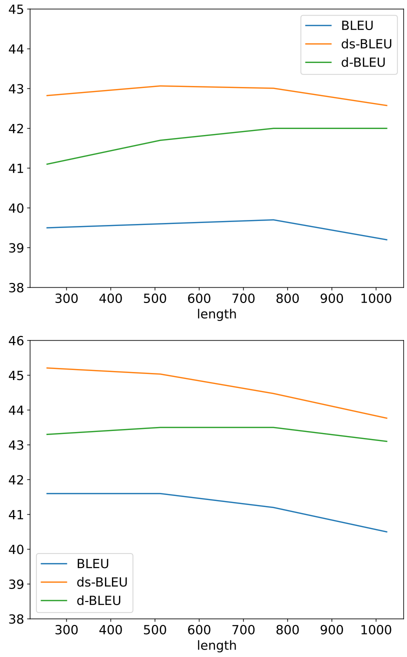

Figure 2: BLEU, ds-BLEU and d-BLEU scores for IWSLT **tst2015**, translating and evaluating *pseudo-documents* of increasing lengths $[256, 512, 768, 1204]$, using FT-NLLB-U (top) and UNIF-TOWER-U (bottom). Note that d-BLEU is computed for pseudo-documents while ds-BLEU is computed for concatenated full talks.

A.3 Data Statistics and Other Details

Full data statistics are given in Tables 8 and 9. All the full TED talks in our corpora start with the title, then the description and the talk before being split into pseudo-documents. `<description>` and `<title>` tags are removed. When preparing our training and validation sets **TED-U** and **TED-G** (see Section 5.1), if concatenating the last sentence pair (x_n, y_n) into the current pseudo-document pair exceeds l_{max}, (x_n, y_n) will yield a single parallel sequence, to respect the maximum length l_{max}. The length distribution is illustrated in Figure 3.

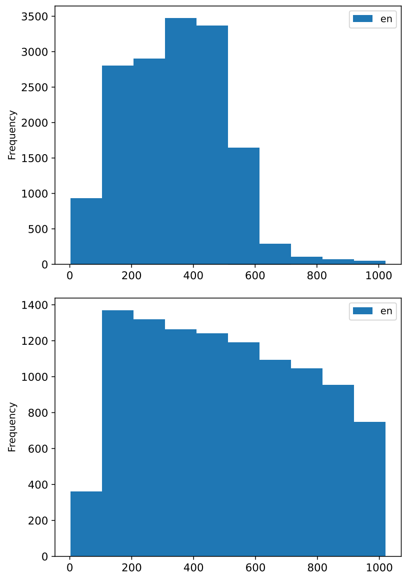

Figure 3: Source document length distribution in the training set of **TED-G** (top) and **TED-U** (bottom).

		2014	2015	2016	2017
NLLB	sent	84	85	85	84
	256	68	69	68	66
	512	49	47	47	46
	768	43	42	40	41
	1024	36	37	36	36
TOWERBASE	sent	84	85	85	85
	256	80	81	82	80
	512	79	80	82	80
	768	78	78	80	80
	1024	76	73	78	76
	1200	73	74	77	76
	1600	70	72	65	68
	2048	52	63	65	57

Table 7: 100×COMET scores for NLLB (top) and TOWERBASE (bottom).

A.4 Experimental Settings

This section presents detailed experiment settings for fine-tuning NLLB and TOWERBASE.

For NLLB, we fine-tuned the pretrained model with learning rate 5e−4, 500 warm-up steps, 4 parallel pseudo-documents per batch and 32 gradient accumulation steps. An early stopping criterion with a patience of 5 epochs is also applied, according to the d-BLEU evaluated on the validation set. For inference on test sets, the beam size is set to 5 and the batch size is set to 4.

For TOWERBASE, We performed supervised fine-tuning using QLoRA (Dettmers et al., 2023) and bfloat16.[20] The batch size is 8 with 2 gradient accumulation steps. The learning rate is 2e−5 adjusted by a *cosine* schedule, without warm-up steps nor packing. We fine-tuned the model for two epochs and saved checkpoints every 50 steps in the second epoch. We then chose the checkpoint with the best d-BLEU on the validation set. Inference is performed without additional in-context examples, with bfloat16 and greedy search.

		count	mean	min	max
TED-full	train	1831	2915 /3515	56 /62	8706 /9706
	dev	19	2861 /3460	680 /867	6076 /7590
TED-G	train	15625	341 /411	3 /2	1022 /1460
	dev	160	339 /410	12 /17	959 /1203
TED-U	train	10582	504 /608	3 /1	1020 /1527
	dev	106	512 /620	42 /41	991 /1276

Table 8: Statistics of the TED talks training and dev sets. ‘count’ denotes the number of parallel pseudo-documents. ‘mean’, ‘min’ and ‘max’ represent the average, minimum and maximum lengths of English/French pseudo-documents respectively, in NLLB tokens.

A.5 Detailed Evaluation Results

The paired comparison and the complete BLEU and COMET scores for each test set are given in Tables 11 to 15.

Document-level Fine-tuning Table 11 reports average differences of automatic scores between fine-tuned MT systems and the corresponding pre-trained models (NLLB or TOWERBASE), for varying test document lengths. ds-BLEUs are averaged over 52 complete TED talks and COMET scores are averaged over 5,103 sentences. Fine-tuning significantly improves over base conditions for all lengths, with larger increases for longer test texts, where the baseline scores were initially very poor. Both metrics yield consistent conclusions, except for the sentence-level assessment of NLLB fine-tuned on **TED-U**, which is slightly worse than the baseline according to ds-BLEU (-0.8), but for which COMET detects no difference. For TOWERBASE, DLFT is always beneficial.

Realignment Issues Since the COMET score is sentence-based, its computation requires a realignment between hypotheses and reference sentences in the Doc2Doc scenario. However, due to issues with long documents, translation hypotheses can be incomplete, resulting in empty alignment for some sentences. These untranslated sentences often occur in the final part of long documents. Table 10

[20]The prompt for fine-tuning is “Translate the following text from English into French.\nEnglish: SRC\nFrench: TGT”, and the zero-shot prompt for the pretrained model TOWERBASE is “English: SRC\nFrench:”.

	l_{max}	2014 count	2014 min	2014 max	2014 mean	2015 count	2015 min	2015 max	2015 mean	2016 count	2016 min	2016 max	2016 mean	2017 count	2017 min	2017 max	2017 mean
EN	sent	1335	2	112	23	1104	2	119	23	1185	1	151	24	1479	2	162	23
EN	256	129	65	286	234	107	71	325	234	123	61	255	232	144	65	271	234
EN	512	68	85	511	443	56	53	510	447	63	56	511	454	74	73	511	456
EN	768	48	116	767	628	40	86	766	626	45	104	767	635	51	57	767	662
EN	1024	37	83	1022	815	30	68	1023	835	35	115	1023	817	40	65	1023	844
EN	1200	32	54	1218	942	26	71	1198	963	31	73	1216	922	34	125	1203	992
EN	1600	26	135	1597	1160	24	114	1599	1043	23	191	1616	1243	27	229	1635	1250
EN	2048	20	569	2091	1507	16	176	2072	1565	21	247	2046	1361	23	65	2045	1467
EN	doc	15	995	4116	2010	12	1256	3359	2086	13	842	3366	2199	12	1909	3722	2812
FR	sent	1335	2	158	28	1104	2	145	27	1185	1	180	29	1479	2	211	27
FR	256	129	80	380	295	107	85	355	282	123	69	345	276	144	78	375	275
FR	512	68	106	717	559	56	62	679	540	63	70	672	539	74	83	737	535
FR	768	48	142	1065	792	40	102	1009	755	45	112	985	755	51	68	1083	776
FR	1024	37	100	1436	1027	30	64	1349	1007	35	125	1314	970	40	80	1431	990
FR	1200	32	61	1641	1188	26	85	1577	1162	31	93	1511	1096	34	134	1714	1164
FR	1600	26	173	2188	1462	24	156	2116	1259	23	209	2074	1477	27	279	2261	1466
FR	2048	20	657	2613	1901	16	218	2602	1889	21	280	2626	1617	23	80	2714	1721
FR	doc	15	1289	4983	2534	12	1609	4013	2518	13	1004	4179	2613	12	2473	4464	3299

Table 9: Statistics of the test sets based on talks from IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017** (see Section 5.1). 'count' refers to the number of parallel pseudo-documents. 'mean', 'min' and 'max' denote the average, minimum and maximum lengths of the source (i.e. English, top) or the reference (i.e. French, bottom) pseudo-documents. All lengths are counted in NLLB tokens.

displays the statistics of empty alignments across all the $5,103$ sentences. This issue is more severe for NLLB models than TowerBase models, which is consistent with the poor BP reported in Tables 12 and 13.

	NLLB	FT-U	Unif-U	FT-G	Unif-G
sent	0	0	0	0	0
256	557	6	6	6	6
512	1231	5	9	12	11
768	1618	6	10	250	280
1024	886	53	34	491	486
1200	1179	437	207	576	675
1600	352	465	657	789	840
2048	456	644	843	801	1089

	Tower	FT-U	Unif-U	FT-G	Unif-G
sent	0	0	0	0	0
256	79	3	8	2	3
512	58	2	2	3	2
768	65	4	3	3	5
1024	45	17	21	19	22
1200	107	19	17	13	19
1600	91	73	50	40	54
2048	151	94	84	66	98

Table 10: Number of empty alignments across all the $5,103$ sentences in our test sets for NLLB (top) and TowerBase (bottom) models.

		ds-BLEU		COMET	
		FT-U	FT-G	FT-U	FT-G
NLLB	sent	-0.8 (0.00)	-0.8 (0.01)	-	-0.3 (0.01)
	256	7.6 (0.00)	8.0 (0.00)	13.0 (0.00)	13.8 (0.00)
	512	26.3 (0.00)	27.0 (0.00)	33.4 (0.00)	33.8 (0.00)
	768	33.5 (0.00)	28.0 (0.00)	39.0 (0.00)	27.8 (0.00)
	1024	33.5 (0.00)	21.2 (0.00)	41.9 (0.00)	22.6 (0.00)
TOWERBASE	sent	2.4 (0.00)	2.6 (0.00)	0.7 (0.00)	0.7 (0.00)
	256	2.1 (0.00)	1.6 (0.00)	2.3 (0.00)	2.3 (0.00)
	512	2.1 (0.00)	2.0 (0.01)	2.4 (0.00)	2.6 (0.00)
	768	3.5 (0.00)	3.2 (0.00)	3.7 (0.00)	3.7 (0.00)
	1024	5.6 (0.00)	5.0 (0.00)	5.6 (0.00)	5.5 (0.00)
	1200	5.4 (0.00)	5.1 (0.00)	5.5 (0.00)	5.4 (0.00)
	1600	8.4 (0.00)	8.3 (0.00)	10.1 (0.00)	10.3 (0.00)
	2048	6.8 (0.00)	6.1 (0.00)	8.8 (0.00)	7.9 (0.00)

Table 11: Average difference (and p-values) in ds-BLEU or 100×COMET between fine-tuned models (FT) and the corresponding pretrained models NLLB (top) and TOWERBASE (bottom). U and G denote the corpora **TED-U** and **TED-G** respectively. - for p-values > 0.05. Positive values indicate that the fine-tuned model improves over the baseline.

		2014	2015	2016	2017
sent	FT	43.8 (0.98)	44.0 (0.99)	40.4 (1.00)	41.3 (1.00)
	Unif	42.4 (0.98)	42.5 (0.99)	39.7 (1.00)	40.2 (1.00)
	SHAPE	40.7 (0.97)	41.5 (0.98)	38.3 (1.00)	39.4 (1.00)
256	FT	43.5 (0.98)	43.1 (0.99)	40.4 (1.00)	40.8 (1.00)
	Unif	43.9 (0.98)	43.2 (0.99)	39.7 (1.00)	40.6 (1.00)
	SHAPE	43.3 (0.98)	43.2 (0.99)	39.8 (1.00)	40.2 (0.99)
512	FT	43.5 (0.98)	43.4 (0.98)	40.2 (1.00)	40.4 (1.00)
	Unif	43.8 (0.98)	43.8 (0.99)	39.8 (1.00)	41.0 (1.00)
	SHAPE	40.6 (0.91)	40.5 (0.92)	37.0 (0.9)	40.4 (0.98)
768	FT	36.6 (0.87)	36.4 (0.88)	35.3 (0.92)	35.6 (0.93)
	Unif	41.8 (0.95)	39.1 (0.88)	38.1 (0.95)	39.4 (0.97)
	SHAPE	34.0 (0.75)	34.9 (0.77)	33.8 (0.84)	34.9 (0.85)
1024	FT	28.6 (0.70)	29.1 (0.75)	28.9 (0.80)	28.7 (0.79)
	Unif	36.1 (0.81)	38.7 (0.87)	34.9 (0.87)	37.2 (0.92)
	SHAPE	32.4 (0.73)	30.9 (0.69)	30.8 (0.75)	28.0 (0.69)
1200	FT	25.2 (0.64)	25.8 (0.74)	24.1 (0.71)	24.3 (0.73)
	Unif	34.6 (0.80)	36.4 (0.82)	30.0 (0.74)	32.4 (0.80)
	SHAPE	27.2 (0.61)	30.3 (0.68)	24.6 (0.64)	23.9 (0.60)
1600	FT	18.2 (0.53)	19.3 (0.62)	19.2 (0.62)	19.6 (0.59)
	Unif	25.5 (0.59)	30.1 (0.70)	26.9 (0.68)	29.4 (0.72)
	SHAPE	22.0 (0.50)	21.7 (0.50)	22.1 (0.56)	20.6 (0.57)
2048	FT	15.4 (0.49)	12.4 (0.52)	16.7 (0.58)	14.8 (0.61)
	Unif	22.0 (0.51)	21.6 (0.50)	24.3 (0.60)	20.6 (0.53)
	SHAPE	18.7 (0.43)	15.1 (0.44)	14.6 (0.38)	13.4 (0.48)

		2014	2015	2016	2017
sent	FT	84	85	85	84
	Unif	82	84	84	83
	SHAPE	82	83	83	83
256	FT	81	82	82	81
	Unif	81	82	82	81
	SHAPE	80	82	81	80
512	FT	81	82	81	80
	Unif	81	82	81	81
	SHAPE	77	78	76	77
768	FT	69	69	70	69
	Unif	74	75	74	73
	SHAPE	68	69	68	68
1024	FT	58	59	60	58
	Unif	67	65	67	67
	SHAPE	60	62	61	59
1200	FT	56	56	55	53
	Unif	61	65	62	60
	SHAPE	55	59	55	54
1600	FT	50	49	49	49
	Unif	55	58	54	56
	SHAPE	51	52	50	48
2048	FT	47	42	48	45
	Unif	51	49	51	48
	SHAPE	46	43	47	44

Table 12: ds-BLEU (and brevity penalty) (left) and 100×COMET (right) scores for FT-NLLB-G (FT), UNIF-NLLB-G (Unif), and SHAPE-NLLB-U (SHAPE) trained on **TED-G** with target max source document length $M = 2048$.

		2014	2015	2016	2017
sent	FT	44.2 (0.99)	43.5 (0.99)	40.4 (1.00)	41.4 (1.00)
	Unif	40.1 (0.95)	40.4 (0.97)	38.0 (1.00)	38.1 (0.99)
	SHAPE	39.4 (0.97)	40.0 (0.98)	36.3 (1.00)	37.8 (0.99)
256	FT	43.2 (0.98)	42.8 (0.99)	39.7 (1.00)	40.5 (1.00)
	Unif	42.8 (0.98)	42.4 (0.99)	39.5 (1.00)	40.3 (1.00)
	SHAPE	42.0 (0.97)	42.5 (0.99)	39.6 (1.00)	39.4 (1.00)
512	FT	42.9 (0.98)	43.1 (0.99)	39.2 (1.00)	39.4 (1.00)
	Unif	43.4 (0.98)	43.0 (0.99)	39.8 (1.00)	40.5 (1.00)
	SHAPE	39.7 (0.89)	41.1 (0.94)	38.2 (0.95)	39.0 (0.98)
768	FT	43.5 (0.98)	43.0 (0.99)	39.4 (1.00)	39.8 (1.00)
	Unif	44.0 (0.98)	43.7 (0.99)	40.4 (1.00)	40.9 (1.00)
	SHAPE	39.6 (0.88)	41.4 (0.93)	37.4 (0.93)	36.8 (0.91)
1024	FT	42.6 (0.96)	42.6 (0.97)	39.2 (1.00)	39.6 (1.00)
	Unif	42.6 (0.96)	44.1 (0.98)	39.6 (1.00)	40.6 (0.99)
	SHAPE	40.3 (0.91)	42.8 (0.96)	36.8 (0.92)	37.8 (0.94)
1200	FT	38.3 (0.88)	39.3 (0.91)	36.3 (0.92)	36.4 (0.92)
	Unif	39.5 (0.89)	43.0 (0.97)	38.0 (0.95)	39.2 (0.98)
	SHAPE	36.9 (0.84)	37.2 (0.87)	32.6 (0.82)	34.3 (0.86)
1600	FT	31.5 (0.77)	30.4 (0.73)	30.9 (0.83)	30.3 (0.80)
	Unif	31.5 (0.72)	34.4 (0.82)	34.2 (0.88)	33.8 (0.84)
	SHAPE	28.0 (0.68)	29.4 (0.68)	31.1 (0.79)	31.7 (0.82)
2048	FT	27.4 (0.69)	24.0 (0.63)	26.7 (0.75)	23.6 (0.71)
	Unif	30.2 (0.68)	25.3 (0.60)	31.5 (0.79)	28.0 (0.68)
	SHAPE	26.1 (0.63)	27.9 (0.66)	26.8 (0.69)	26.9 (0.74)

		2014	2015	2016	2017
sent	FT	84	85	85	84
	Unif	82	83	83	83
	SHAPE	81	82	82	82
256	FT	80	81	81	81
	Unif	80	81	81	81
	SHAPE	79	81	81	80
512	FT	80	81	81	81
	Unif	81	82	81	81
	SHAPE	75	79	79	78
768	FT	80	81	81	81
	Unif	81	82	82	81
	SHAPE	75	78	75	75
1024	FT	78	79	78	78
	Unif	80	81	81	80
	SHAPE	74	78	75	73
1200	FT	71	73	70	69
	Unif	76	79	76	75
	SHAPE	68	70	69	66
1600	FT	61	60	61	60
	Unif	62	61	61	64
	SHAPE	57	59	62	59
2048	FT	57	53	57	54
	Unif	59	57	57	54
	SHAPE	51	52	57	53

Table 13: ds-BLEU (and brevity penalty) (left) and $100\times$COMET (right) scores for FT-NLLB-U (FT) UNIF-NLLB-U (Unif), and SHAPE-NLLB-U (SHAPE) trained on **TED-U** with target max source document length $M = 2048$.

		2014	2015	2016	2017
sent	FT	46.5 (0.98)	45.1 (0.99)	42.3 (1.00)	41.0 (1.00)
	Unif	46.5 (0.98)	45.0 (0.99)	42.3 (1.00)	41.1 (1.00)
	SHAPE	46.4 (0.98)	45.2 (0.99)	42.4 (1.00)	41.2 (1.00)
256	FT	44.6 (0.98)	45.1 (0.99)	42.3 (1.00)	41.9 (1.00)
	Unif	44.5 (0.99)	45.1 (0.99)	42.2 (1.00)	41.8 (1.00)
	SHAPE	46.2 (0.98)	45.2 (0.99)	42.4 (1.00)	42.1 (1.00)
512	FT	43.7 (0.98)	45.0 (1.00)	41.4 (1.00)	41.6 (1.00)
	Unif	43.7 (0.98)	44.8 (1.00)	41.4 (1.00)	41.4 (1.00)
	SHAPE	45.5 (0.98)	45.1 (0.99)	41.6 (1.00)	41.6 (1.00)
768	FT	44.0 (0.98)	44.2 (0.99)	40.3 (1.00)	40.7 (1.00)
	Unif	44.0 (0.98)	44.2 (0.99)	40.2 (1.00)	40.7 (1.00)
	SHAPE	45.6 (0.98)	44.4 (0.99)	41.3 (1.00)	41.5 (1.00)
1024	FT	42.7 (0.98)	40.6 (0.99)	38.9 (1.00)	40.4 (1.00)
	Unif	42.2 (0.97)	42.6 (0.99)	39.4 (1.00)	40.1 (1.00)
	SHAPE	44.5 (0.98)	40.9 (0.99)	39.2 (1.00)	39.7 (1.00)
1200	FT	42.6 (0.98)	43.0 (0.99)	38.9 (1.00)	40.8 (1.00)
	Unif	42.5 (0.98)	42.7 (0.99)	39.3 (1.00)	40.6 (1.00)
	SHAPE	44.0 (0.98)	42.8 (0.99)	39.3 (1.00)	40.6 (1.00)
1600	FT	42.0 (0.97)	41.1 (0.98)	37.0 (1.00)	38.7 (1.00)
	Unif	42.0 (0.97)	40.0 (0.98)	37.5 (1.00)	37.3 (1.00)
	SHAPE	42.2 (0.97)	40.6 (0.98)	38.5 (1.00)	39.2 (0.99)
2048	FT	33.0 (0.99)	32.5 (0.99)	31.6 (1.00)	29.2 (0.97)
	Unif	33.1 (0.99)	33.7 (0.99)	32.2 (0.98)	26.5 (0.97)
	SHAPE	34.1 (0.97)	35.1 (0.98)	32.1 (1.00)	30.2 (0.97)

		2014	2015	2016	2017
sent	FT	85	86	85	85
	Unif	85	86	85	85
	SHAPE	85	86	85	85
256	FT	83	84	83	83
	Unif	83	84	83	83
	SHAPE	83	84	83	83
512	FT	82	84	83	82
	Unif	82	84	83	82
	SHAPE	82	84	83	82
768	FT	82	83	83	82
	Unif	82	83	83	82
	SHAPE	82	84	83	82
1024	FT	81	81	82	81
	Unif	81	82	83	81
	SHAPE	81	81	83	81
1200	FT	78	82	81	81
	Unif	78	82	81	81
	SHAPE	80	82	82	81
1600	FT	80	79	79	80
	Unif	79	78	79	78
	SHAPE	79	79	80	79
2048	FT	68	69	68	64
	Unif	69	70	70	62
	SHAPE	65	73	70	68

Table 14: ds-BLEU (and brevity penalty) (left) and $100\times$COMET (right) scores for FT-TOWER-G (FT), UNIF-TOWER-G (Unif) and SHAPE-TOWER-G (SHAPE) trained on **TED-G** with target max source document length 2048 ($M = 4096$).

		2014	2015	2016	2017
sent	FT	46.2 (0.99)	45.1 (0.99)	42.1 (1.00)	40.8 (1.00)
	Unif	46.4 (0.98)	45.1 (0.99)	42.2 (1.00)	40.9 (1.00)
	SHAPE	46.3 (0.98)	45.2 (0.99)	42.4 (1.00)	41.0 (1.00)
256	FT	46.3 (0.98)	45.1 (0.99)	42.3 (1.00)	41.8 (1.00)
	Unif	44.5 (0.98)	45.2 (0.99)	42.3 (1.00)	41.9 (1.00)
	SHAPE	46.0 (0.98)	45.3 (0.99)	42.4 (1.00)	42.0 (1.00)
512	FT	44.4 (0.98)	44.9 (0.99)	41.2 (1.00)	41.6 (1.00)
	Unif	43.0 (0.98)	45.0 (0.99)	41.4 (1.00)	41.6 (1.00)
	SHAPE	44.5 (0.98)	45.0 (0.99)	41.3 (1.00)	41.6 (1.00)
768	FT	44.1 (0.98)	44.6 (0.99)	41.1 (1.00)	40.8 (1.00)
	Unif	43.2 (0.99)	44.5 (0.99)	41.4 (1.00)	41.0 (1.00)
	SHAPE	44.2 (0.99)	44.5 (0.99)	41.2 (1.00)	41.7 (1.00)
1024	FT	43.9 (0.97)	41.6 (0.99)	39.6 (1.00)	39.7 (1.00)
	Unif	42.7 (0.98)	43.8 (0.99)	39.7 (1.00)	39.6 (1.00)
	SHAPE	44.1 (0.98)	42.9 (0.99)	39.8 (1.00)	39.5 (1.00)
1200	FT	43.2 (0.97)	42.9 (0.99)	39.5 (1.00)	40.8 (1.00)
	Unif	43.2 (0.98)	43.1 (0.99)	38.2 (1.00)	40.9 (1.00)
	SHAPE	43.8 (0.98)	43.2 (0.99)	39.2 (1.00)	39.7 (1.00)
1600	FT	41.6 (0.95)	40.5 (0.97)	37.6 (1.00)	39.8 (0.99)
	Unif	41.1 (0.97)	41.6 (0.98)	36.5 (1.00)	38.0 (1.00)
	SHAPE	42.6 (0.96)	42.1 (0.98)	37.3 (1.00)	40.5 (1.00)
2048	FT	34.4 (0.96)	35.3 (0.97)	31.2 (0.99)	28.2 (0.96)
	Unif	34.6 (0.97)	35.5 (0.98)	30.2 (1.00)	30.9 (0.97)
	SHAPE	35.2 (0.97)	34.6 (0.95)	31.9 (0.99)	31.5 (0.96)

		2014	2015	2016	2017
sent	FT	85	86	85	85
	Unif	85	86	85	85
	SHAPE	85	86	85	86
256	FT	83	84	83	83
	Unif	83	84	83	82
	SHAPE	83	84	83	82
512	FT	82	84	83	82
	Unif	82	84	83	82
	SHAPE	82	84	83	82
768	FT	82	83	83	82
	Unif	82	84	83	82
	SHAPE	82	83	83	82
1024	FT	81	81	83	81
	Unif	81	83	83	81
	SHAPE	81	82	83	81
1200	FT	80	82	81	81
	Unif	80	82	81	81
	SHAPE	80	82	81	80
1600	FT	79	79	80	80
	Unif	80	80	78	79
	SHAPE	79	81	78	80
2048	FT	68	72	69	64
	Unif	70	72	69	67
	SHAPE	70	73	70	68

Table 15: ds-BLEU (and brevity penalty) (left) and $100\times$COMET (right) for FT-TOWER-U (FT), UNIF-TOWER-U (Unif) and SHAPE-TOWER-U (SHAPE) trained on **TED-U** with target max source document length 2048 ($M = 4096$).

	TED-U		**TED-G**		FT	Unif	SHAPE
	FT vs Unif	FT vs SHAPE	FT vs Unif	FT vs SHAPE	U vs G	U vs G	U vs G
sent	-0.1 (0.20)	**-0.2 (0.01)**	-0.0 (0.60)	-0.1 (0.18)	**-0.1 (0.05)**	-0.1 (0.19)	-0.1 (0.36)
256	0.5 (0.32)	-0.0 (0.87)	0.1 (0.22)	-0.5 (0.24)	0.5 (0.32)	0.1 (0.21)	-0.1 (0.48)
512	0.3 (0.55)	-0.1 (0.10)	0.1 (0.08)	-0.6 (0.23)	0.1 (0.87)	-0.1 (0.82)	-0.4 (0.28)
768	0.2 (0.46)	-0.2 (0.67)	0.0 (0.84)	**-0.9 (0.05)**	0.3 (0.11)	0.2 (0.65)	-0.4 (0.19)
1024	-0.2 (0.82)	-0.3 (0.49)	-0.4 (0.44)	-0.5 (0.35)	0.6 (0.29)	0.4 (0.34)	0.5 (0.38)
1200	0.2 (0.44)	0.1 (0.82)	0.0 (0.70)	-0.4 (0.06)	0.3 (0.24)	0.1 (0.71)	-0.2 (0.50)
1600	0.6 (0.45)	**-0.7 (0.01)**	0.4 (0.58)	-0.5 (0.40)	0.1 (0.84)	0.0 (1.00)	0.4 (0.49)
2048	-0.5 (0.61)	-1.0 (0.22)	0.2 (0.84)	-1.3 (0.19)	0.7 (0.46)	1.3 (0.13)	0.4 (0.44)
sent	-0.0 (0.52)	-0.0 (0.93)	0.0 (0.38)	0.0 (0.60)	-0.0 (0.95)	0.0 (0.24)	0.0 (0.56)
256	0.0 (0.60)	0.0 (0.85)	0.0 (0.73)	-0.1 (0.07)	-0.0 (0.63)	-0.1 (0.46)	**-0.2 (0.05)**
512	-0.1 (0.15)	**-0.2 (0.04)**	0.1 (0.16)	-0.1 (0.27)	-0.2 (0.09)	0.0 (0.76)	-0.0 (0.62)
768	-0.1 (0.29)	0.1 (0.28)	-0.1 (0.34)	**-0.3 (0.00)**	0.0 (0.87)	0.0 (0.69)	**-0.4 (0.00)**
1024	-0.6 (0.22)	-0.5 (0.32)	-0.3 (0.46)	-0.2 (0.09)	0.1 (0.45)	0.3 (0.10)	0.4 (0.42)
1200	0.1 (0.69)	0.2 (0.51)	0.1 (0.50)	-0.4 (0.16)	0.1 (0.61)	0.0 (0.93)	-0.4 (0.12)
1600	0.2 (0.69)	-0.4 (0.27)	0.6 (0.37)	0.1 (0.70)	-0.2 (0.62)	0.2 (0.73)	0.4 (0.49)
2048	-0.9 (0.40)	-1.3 (0.16)	-0.2 (0.84)	-1.7 (0.13)	0.9 (0.50)	1.5 (0.10)	0.4 (0.47)

Table 16: Average difference (and p-values) in **ds-BLEU** (top) evaluated on full TED talks and $100\times$**COMET** (bottom) evaluated on realigned sentences for TOWERBASE. Left and middle: paired comparison between the original fine-tuning (FT), UNIFPE and SHAPE on **TED-U** and **TED-G** respectively. Right: differences between fine-tuning on TED-U (U) and TED-G (G). A positive value indicates that in the comparison pair, the translation of the first item achieves higher scores than that of the second. Significant differences with p-values < 0.05 are in bold.