Handling Very Long Contexts in Neural Machine Translation: a Survey

Ziqian Peng, Rachel Bawden, François Yvon

;

Handling Very Long Contexts in Neural Machine Translation: a Survey

Ziqian Peng, Rachel Bawden and François Yvon

June 2024

MaTOS — Livrable D3-2.1

Machine Translation for Open Science - ANR-22-CE23-0033

Handling Very Long Contexts in Neural Machine Translation: a Survey

Ziqian Peng, Rachel Bawden and François Yvon

June 2024

Abstract

This report examines methods for integrating an extended discourse context in machine translation, focusing on neural translation methods. Machine translation systems generally translate each sentence independently of its neighbors, which yields systematic errors resulting from a limited discourse context. Therefore, various approaches have been proposed to incorporate cross-sentential context, mostly based on the predominant Transformer architecture. Recently, the introduction of large language models (LLMs) also created novel opportunities to process long-range dependencies, inspiring several context-aware machine translation approaches. We present the challenges of translating long inputs, then investigate encoder-decoder architectures and LLM-based approaches, with a brief overview of efficient transformer implementations as a common background. Furthermore, we also discuss strategies to extend other NLP tasks to a longer context, and list recently available open-source document-level parallel corpus for future exploration. We conclude with a summary of current work and the main research directions.

Résumé

Ce rapport étudie les méthodes visant à intégrer un contexte discursif étendu en traduction automatique (TA), en se focalisant sur les méthodes de traduction neuronales. Les systèmes de traduction automatique traduisent en général chaque phrase indépendamment de ses voisines, ce qui entraîne des erreurs systématiques qui résultent d'un contexte discursif trop étroit. Diverses approches ont été proposées pour intégrer le contexte au-delà de la phrase courante, en s'appuyant sur l'architecture transformeur, qui est l'architecture prédominante en TA. Récemment, l'introduction de grands modèles de langue (LLM) a également créé de nouvelles opportunités pour traiter les dépendances à longue portée, donnant lieu à la formulation d'approches holistiques de la traduction, qui prennent en compte un contexte étendu. Nous discutons des défis que pose la traduction de longs documents, avant de présenter

les méthodes proposées pour les architectures encodeurs-décodeurs et les approches à base de LLM, avec un bref aperçu des implémentations efficaces pour les transformeurs, qui subsubmment ces deux types de modèles. En complément, nous considérons également des stratégies d'extension de la fenêtre du contexte pour d'autres tâches de TAL; nous avons également listé des corpus de documents parallèles récemment disponibles en source ouverte, pour une exploration future. Nous concluons par un résumé des travaux actuels et des principales directions de recherche.

Contents

1 Introduction

Neural machine translation (NMT) has experienced remarkable progress at the sentence level with the adoption of the Transformer architecture [Vaswani et al., 2017]. Yet, human evaluations continue to demonstrate the superiority of human translations, as long as they can access the translation of a longer context, or even that of the entire document [Läubli et al., 2018]. This is because sentence-level MT (SLMT) is, by design, unable to handle certain linguistic issues.

Discourse-level issues in Machine Translation With only intra-sentence information, sentence-level neural machine translation models cannot correctly handle certain discourse phenomena, such as the resolution of anaphoric references, formality, consistency, and coherence issues, which all require a long-term context [Bawden et al., 2018, Voita et al., 2019b, Maruf et al., 2019].

We list below the most discussed context-aware discourse issues to provide the reader with the necessary linguistic background information. More informative and comprehensive presentations are given in, e.g. [Joty et al., 2017, Popescu-Belis, 2019, Zhang, 2020, Abdul Rauf and Yvon, 2020].

1. **anaphora and coreference** refers to the process of establishing connections between references of the same entity. For example, in the case of anaphoric pronouns:

   Mary will join us for dinner, if *she* is in town.

2. **deixis** are referential expressions, whose interpretation in an utterance depends on extra-linguistic factors, such as a specific time, place, or person in context.

3. **ellipsis** is the omission of one or more words from a clause that are nevertheless understood in the context of the sentence. For example, "*You might do it, but I won't (do it)*".

4. **lexical consistency** describes the logical alignment and uniformity of terms and entities within a text. It ensures that the information presented does not contradict itself and maintains a stable tone and style throughout. Other important cohesion-building devices are repetitions, collocations, tense or pronoun use, etc.

5. **word sense disambiguation (WSD)** deals with the determination of a correct meaning or sense of an expression in a given context [Agirre and Stevenson, 2022]. It is usually associated with discourse phenomena, as lexical ambiguities within a document need to be resolved in a mutually consistent manner, participating also in the building of a cohesive text.

6. **discourse connectives** are the cohesive markers that join clauses in texts and indicate discourse relations between adjacent spans. These include words such as: "*although*", "*while*", "*however*", "*since*", "*for example*", etc. [Meyer and Webber, 2013]. Correctly translating connectives is key to making the target text fluid and logically coherent.

Extended Contexts: Local and Global Definitions The issues listed above are diverse, both in their frequency of occurrence and in the severity of associated translation errors. Another key distinction is between their spread: some can usually be resolved by enlarging the sentential context with a couple of preceding or following sentences: this is, for instance, the case of co-references. Others will require a more global view, as, for instance, lexical consistency issues. This distinction is however not always made in the literature, where the "document" in "document-level MT" sometimes refers to a handful of consecutive sentences, whose size is chosen on linguistics, or computational ground[1] Abdul Rauf and Yvon, 2020, tab. 7; Deutsch et al., 2023, Castilho and Knowles, 2024). To clarify this ambiguity, we will make sure to distinguish between *Context-Aware MT*[2] (CAMT) to refer to methods that handle a local context and genuine *Document-Level MT* (DLMT) when a global document context is considered.[3]

Contextualizing machine translation with inter-sentence context thus is necessary to boost machine translation quality, but tricky to deal with, because of the noise in extended contexts and the complexity of handling long input and output sequences. Different discourse phenomena can have diverse distributions across different languages, and they are usually sparse in data [Lupo et al., 2022a, Jin et al., 2023]. This means that including a larger context may also introduce noise and distract the attention mechanism of sentence-level NMT models. This is because enlarging the context has the effect of spreading the attention weights of the current target token throughout the full context rather than within the current sentence.

Another source of inefficiency stems from the self-attention mechanism itself, the computation of which has quadratic complexity, making it very costly for long sequences. Even though a wide range of efficient transformers have been proposed to tackle this problem (see the review of Tay et al. [2023b]), none of them can significantly go beyond the original transformer in terms of both quality and speed [Tay et al., 2021]. This is not the sole consequence of processing larger contexts: it also implies smaller batch sizes and a reduced number of gradient updates. When generating the target texts, the beam search procedure also has to consider longer branches [Herold and Ney, 2023b]. Finally, if the past and future context of current sentences are not efficiently incorporated, this may lead to problems related to search error and label bias [Stahlberg and Byrne, 2019].

[1] E.g. limiting the context window to some arbitrary maximal size to reduce the memory footprint or the processing complexity.

[2] We assume here that all context is textual, which is a gross approximation, as in principle the context could include many other sources of (useful) information, such as the domain and style of the source text, the identity or intent of the writer, etc. We also reckon that DLMT could be seen as a special case of CAMT, with 'context' encompassing both local and global information. We use *Extended Context* for this, and implicitly assume that CAMT uses only local context.

[3] A convenient way to formalize this is as follows: assuming $\mathbf{x} = x_1 \dots x_I$ and $\mathbf{y} = y_1 \dots y_J$ are the source and target sentences, the conditional distribution can be factorized as $P(\mathbf{y}|\mathbf{x}) = \prod_j P(y_j|y_{<j}, \mathbf{x})$. CAMT will typically make additional hypotheses to simplify each term in this product, assuming local dependencies, while DLMT tackles the generic problem. Note that the former approach typically assumes one-to-one correspondences between source and target sentences (and also $I = J$), making it possible to simplify as $P(y_j|y_{<j}, \mathbf{x})$ as e.g. $P(y_i|y_i, s_{x-1}, t_{x-1})$. DLMT makes no such assumption and can handle arbitrary cases of many-to-one correspondences.

Organisation of the report A wide range of research is actively conducted in this subfield to explore efficient approaches that improve NMT with very long contexts. Recently, the introduction of Large Language Models (LLMs) also opens new opportunities to better translate extended context, with the prospect of even handling. In this survey, we aim to investigate findings and approaches regarding machine translation with such longer contexts. In Section 2, we analyze and discuss the challenges associated with long-context MT. We then summarize current approaches proposed in the literature, encompassing traditional encoder-decoder methods (Section 3) as well as LLM-based techniques (Section 4). Additionally, we present other related and insightful explorations to deal with long-range dependencies in Section 5, even though they have not been extensively tested on translation tasks.

Evaluating DLMT Our focus in this report is on translation techniques, yet, before beginning this overview, it is worth emphasizing the importance of document-level / context-aware metrics for DLMT and their advancement. Adequate context-aware metrics are crucial to assess the effectiveness of novel methods in DLMT, and to indicate the correct direction for follow-up research. However, traditional automatic metrics such as BLEU [Papineni et al., 2002] and COMET [Rei et al., 2020] are not sensitive to discourse phenomena [Bawden et al., 2018, He et al., 2023]. The most common alternatives, which use contrastive test suites ([Guillou and Hardmeier, 2016, Müller et al., 2018, Popescu-Belis, 2019], inter alia) usually evaluate one specific phenomenon in a fixed language pair. The development of such test suites is also an expensive process.

Several new context-aware automatic metrics have been introduced [Jin et al., 2023, Castilho and Knowles, 2024], such as CXMI [Fernandes et al., 2021], BlonDe [Jiang et al., 2022], MuDA [Fernandes et al., 2023], X-COMET [Guerreiro et al., 2023]. In addition, Vernikos et al. [2022] proposed a method based on cross-sentential contextual embeddings to convert pre-trained metrics, such as COMET and BERTScore [Zhang et al., 2020] into document-Level metrics. Nevertheless, it seems that the community is still lacking well-recognized benchmarks and metrics to use in their evaluation. Another direction is to examine different factors that influence the translation quality of DLMT. For instance, the relative impact of the past and future context [Agrawal et al., 2018], the degradation of BLEU scores associated with the translation of longer inputs [Neishi and Yoshinaga, 2019, Bao et al., 2021, Li et al., 2022], the actual utility of correctly integrating context information in DLMT [Yin et al., 2021, Mohammed and Niculae, 2024], etc. A more complete review of evaluation issues for DLMT is in a companion report [Dahan et al., 2024].

2 Challenges of Extended Contexts in NMT

Machine translation models compute the probability of appropriate target translations given the source text. Let x_i and y_j respectively denote the i^{th} source sentence (resp. the j^{th} target sentence. The translation of a document $\mathbf{x} = x_1 \cdots x_T$ of T sentences into

$\mathbf{y}$ is modeled through:

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

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

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

denotes the index of the current token in the document.

Sentence-level NMT systems usually assume that sentences are conditionally independent and that the source and target sentences stand in one-to-one correspondences, simplifying the problem as:

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

These assumptions are unrealistic, as sentences in a document depend on each other to convey a coherent message. Furthermore, multiple discourse phenomena require contextual information from surrounding sentences or even from the full document.

Following Sun et al. [2022], we call *Doc2Doc* the holistic translation of an input document as expressed in equation (1), where the model output corresponds to the entire input document.[4] This contrasts with *Sent2Sent*, where the model outputs the translation of isolated input sentences. An intermediate design is *Doc2Sent*, where the model inputs a complete block (typically consisting of a fixed-sized window of past sentences) and only retains the translation of one single sentence (typically the last one) – referred to as *focal sentence* – in the extended output. Doc2Sent remains sentence-based and needs to keep track of sentence boundaries to prepare and post-process translation blocks. Many implementations or variants of Doc2Sent, depending on the content of the block (from a pair of sentences to the full document), as well as the number and position of focal sentences. When the same sentence is translated in several blocks, a post-processing step can be used to reconcile the corresponding outputs. Finally note that Doc2Sent is highly inefficient, as the encoding (and decoding) of large blocks is repeatedly performed to translate only a much small part. Formally, Doc2Sent relies on the specification of Block$(\mathbf{x}, \mathbf{y}, t, l,)$ which computes the context block for token (t, l):[5]

[4] We choose the term Doc2Doc to remain consistent with previous studies. Holistic translation may apply to smaller portions of an input text, such as a chapter, a section, a paragraph, or even a fixed-size block. When necessary, we will update the terminology accordingly, using terms such as *Sec2Sec*, *Par2Par*, or *Block2Block*.

[5] It is custom, but not necessary, to keep Block(t, l) constant with a sentence. Likewise, it is customary, to restrict the target context to past sentences - making the generation process autoregressive.

$$P(y_1 \cdots y_T|\boldsymbol{x}) = \prod_{t=1}^{T} \prod_{l=1}^{L_t} P(y_{t,l}|\,\text{Block}(\mathbf{x}, \mathbf{y}, t, l,)). \tag{4}$$

Doc2Doc is conceptually simple, yet it introduces multiple changes compared to the reference situation where each sentence is encoded and decoded separately from the others, as in equation (3). We recap below the main differences between Doc2Doc and the alternatives for encoder-decoder architecture, bearing in mind that the same observations apply to approaches based on LLMs, when used for translation purposes [Wang et al., 2023a, Karpinska and Iyyer, 2023]. In particular, translating documents holistically means:

- the encoder views the entire source document $\mathbf{x}$, composed of T sentences $\mathbf{x} = (x_1 \dots x_T)$, as one long sequence, with or without prior identification of sentence boundaries;
- to generate the l^{th} target sentence, the decoder has access to $\mathbf{x}$ as well as to all previously translated target sentences $y_{<l} = t_1 \dots y_{l-1}$. Note that for Doc2Doc, y_l does not necessarily translate x_l.

These changes have a number of consequences, some positive (1), others negative (2-6):

1. using more complete source and target contexts gives access to more information and longer dependencies, helping lexical disambiguation, consistency and pronominal references.
2. the sequences to be processed are longer, resulting in a computational overhead, as the attention in the encoder and decoder is quadratic with respect to the length of the sequence attended to [Tay et al., 2023b].
3. attention weights are "diluted" because of a greater number of tokens in longer inputs [Herold and Ney, 2023a]. Yet, at each time step, the inference procedure needs to keep enough attention on the input part that is being translated, as information in this "local" context is much denser than in the global context. The same holds for Doc2Sent models, even though explicitly locating sentence boundaries in the input and output can help this process.
4. when decoding the tokens corresponding to source sentence x_l, the decoder can no longer rely on an explicit alignment between sentences and therefore need to only rely on the cross-attention. This amounts to computing a word alignment over the entire source document $\mathbf{x}$, a difficult process, given that word alignments are usually less intuitive than sentence alignments. This effect is analyzed in particular by Bao et al. [2021]. Again, this issue also exists, albeit in a less acute form, for Doc2Sent approaches.

5. decoding longer sequences increases the impact of search errors and of exposure bias. Recall that the latter is due to the fact that model training only considers correct target contexts ($t_{<l}$), while during inference this context may be incorrect [Ranzato et al., 2016, Mihaylova and Martins, 2019]. Decoding longer sequences also reduces the diversity of hypotheses represented in the beam search.

6. beam search is more difficult. The risk is also to exacerbate problems linked to the length of texts to be translated [Koehn and Knowles, 2017, Stahlberg and Byrne, 2019].

7. the generated sentences are no longer necessarily in one-to-one correspondence with the source sentences, which complicates, or even obtrudes, the computation of conventional metrics designed for parallel sentences.

In the next section, we discuss how these challenges are handled in the framework of standard sequence-to-sequence architectures. We start with attempts to improve the computation performed by self-attention layers (issue #2).

3 Long Contexts in Encoder-Decoder Models

The Transformer architecture of Vaswani et al. [2017] has become predominant in the machine translation field. However, the attention mechanism of the vanilla Transformer suffers from a quadratic complexity, which limits the model's capacity to process long-range dependencies embedded in long, concatenated input. Diverse approaches have been proposed to construct more efficient transformers, and have inspired work in DMT. To provide a comprehensive overview, we begin by providing some background by briefly review efficient transformers in Section 3.1, as these improvements can in fact benefit all approaches having to handle long contexts.

An essential direction involves adapting the model's architecture to facilitate more effective context use (Section 3.2)..

In addition to architecture adaptation, meaningful novel methods relative to data augmentation and training strategies have also been proposed. We summarize them respectively in Sections 3.3 and 3.4.

3.1 Efficient transformers

Vanilla Transformer The Transformer [Vaswani et al., 2017] model is designed to optimize the transformation of context-free lexical embeddings into contextual representations: each computation layer recombines all the input representations in a succession of two main operations, consisting of the **self-attention mechanism** and the transformation through a position-wise fully connected feed-forward network, to yield the input for the next layer.

Given a sequence of n tokens represented as a list of vectors $\mathbf{x} = (x_1, \ldots, x_n) \in R^{n \times d_m}$, with d_m the embedding dimension, the attention mechanism computes a similarity between each input x_i and all the other inputs x_j in X in a weight matrix $\boldsymbol{\alpha} \in R^{n \times n}$, such

that higher weight values represent stronger dependencies. Once normalized in $\boldsymbol{\alpha}$, these coefficients are used to compute the updated representation $x_i^* = \sum_j \alpha_i x_j$, which will then be added to x_i, normalized, and passed through the feed-forward layer with ReLU activation before being propagated to the upper layers.

Multi-head attention can also be applied to jointly attend to information from different representation subspaces. This strategy trains H attention heads in parallel, controlled by $\boldsymbol{\alpha}_h$ with $h = 1 \dots H$, and concatenates them. In other words, $\boldsymbol{\alpha}_h$ is computed during back-propagation at each epoch to update weights simultaneously for each position and each head. It involves three parameter matrices: the query (Q), key (K), and value (V) matrices, as described in equations (5) and (6):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\alpha_1, \dots, \alpha_H)W^O \\ \text{with } \alpha_h &= \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \end{aligned} \tag{6}$$

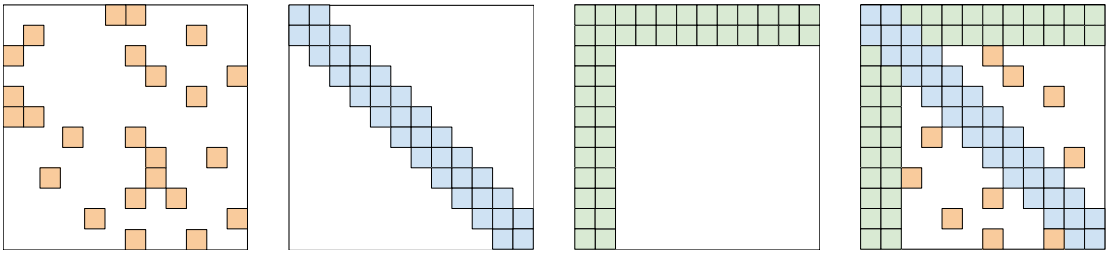
where $Q, K \in R^{n \times d_k}$ and $V \in R^{n \times d_v}$, usually $d_k = d_v$ is taken. For the rest of this section, we use d to denote both d_k and d_v. $QK^T \in R^{n \times n}$ is a dot product that measures the similarity between queries and keys. This computation has **quadratic complexity** for both time and space with respect to n, limiting the processing of input of long sequences. This is the main obstacle to enlarging the transformer context to enable the process of long-range dependencies.

Therefore, a series of *efficient Transformers* have been developed to reduce the space and time complexity. They roughly fall into three categories: linear approximation, sparse attention and other adaptation in model components, along with general efficiency techniques such as parameter sharing. Please refer to [Tay et al., 2023a] for a more comprehensive and detailed overview of this topic.

Linear approximations Approaches employing *linear approximation* aim to approximate Equation (5) with linear or logarithmic complexity. For example, Linformer [Wang et al., 2020] is based on the observation that the computation of $\boldsymbol{\alpha}_h$ can be approximated by the product of two low-rank matrices. These low-rank matrices can be obtained by introducing two random matrices $E_h, F_h \in \mathbb{R}^{n \times S}$ for each head h, used to project KW_h^K, VW_h^V from dimension $n \times d$ to $n \times S$ (cf. Equation (7)). As a result, the term in the softmax outputs an $n \times S$ matrix (instead of $n \times d$).

$$\begin{aligned} \boldsymbol{\alpha}_h &= \text{Attention}(QW_h^Q, E_h KW_h^K, F_h VW_h^V) \\ &= \text{softmax}(\frac{QW_h^Q (E_h KW_h^K)^T}{\sqrt{d}}) \times F_h VW_h^V \end{aligned}$$

By choosing $n \gg S$, we get the expected complexity reduction, at almost zero cost in terms of performance. The authors also show that some parameter sharing, such as sharing the projection matrices across layers, can also speed up the computation without harming performance too much.



(a) random attention (b) window attention (c) global attention (d) BigBird

Figure 1: The sparse attention mechanism of Bigbird. White color indicates absence of attention, Figure extracted from [Zaheer et al., 2020]

.

Sparse attention The main idea of *sparse attention* approaches is to reduce the number of effective (non-zero) coefficients in $\boldsymbol{\alpha}_h$ that need to be taken into consideration. One specific way to proceed is to reduce the number of neighbors of each token to a fixed size. Liu* et al. [2018] first restricts the attention computation to blocks of a limited size. This means that the representation of a token only recombines the representation of tokens within the same block. This reduces the contextualization of token representations but also creates boundary effects between blocks. An alternative (Memory compressed attention) that is explored in the same work uses strided convolutions to reduce the number of neighbors while preserving access to the global context. Boundary effects can also be avoided by considering neighbors in a sliding window of S tokens, as shown in Figure 1b, which means that only the near-diagonal terms of the attention matrix will be computed. Although the context is localized in the lower layers, it still remains global at the upper layers as the influence of more remote tokens propagates in the network. A further trick is to "dilate" these local contexts to speed up the diffusion in the network. To preserve the overall performance, a critical aspect is to make sure that a restricted number of positions still keep a global view over the full input, meaning that they attend to (and are attended to) by all positions (cf. Figure 1c). These positions can be described as performing local summaries that are propagated through the entire network.

A typical example that generalizes these ideas is BigBird [Zaheer et al., 2020]. As shown in Figure 1, BigBird combines not only the sliding window and global token attention but also a random attention pattern such that each query attends to r random keys. The authors show that these random neighbors help speed up the "diffusion" of information amongst tokens. Other models in this category are the Sparse Transformer [Child et al., 2019], Longformer [Beltagy et al., 2020], ETC [Ainslie et al., 2020] and Unlimiformer [Bertsch et al., 2023]. These methods are also employed in several LLMs, such as GPT-3 [Brown et al., 2020] and Mistral [Jiang et al., 2023a]. In addition, Correia et al. [2019] replaced the costly softmax operator by the less costly sparsemax [Niculae and Blondel, 2017], to enforce attention sparsity and speed up the computation without having to set a meta-parameter.

In Reformer [Kitaev et al., 2020], locally-sensitive hashing (LSH) is used to identify the most significant terms in the summation implied by the dot product QK^T (corre-

sponding to the most similar neighbors), thereby yielding sparse attention matrices and computational gains.

Others Regarding other adaptations to the network design, some representative architectures are FLASH [Hua et al., 2022] and MEGA [Ma et al., 2023].

FLASH, the abbreviation for *Fast Linear Attention with a Single Head*, proposes the *Gated Attention Unit* (GAU) structure to replace the multi-head self-attention layer. The GAU contains a single attention head, replacing the softmax in equation (5) with the more efficient ReLU^2 and applying a learnable relative position bias. It therefore contains fewer parameters than multi-head attention, while keeping a comparable level of performance. Since the complexity of GAU is still quadratic ($O(n^2)$), *mixed chunk attention* is used to approximate GAU with a $O(n)$ complexity. This approach first groups tokens into chunks then uses precise quadratic attention in GAU within a chunk to compute local attention and fast linear attention $Q(K^TV)$ across chunks to capture global long-range interactions.

Ma et al. [2023] introduced the *Moving average Equipped Gated Attention* (MEGA) mechanism. Inspired by the Exponential Moving Average (EMA) approach, which controls the influence of past time steps by a weighting factor that decreases exponentially, MEGA encodes contextual information through EMA, thereby prioritizing the local context (i.e. recent time steps, see Ma et al. [2023, fig. 1]). It also applies a reset gate and an update gate dedicated to the context, to control the impact of contextual information when computing the final output of each MEGA layer. In addition, the authors compute the attention mechanism using GAUs, and they add to the scaled dot product a relative positional bias. They further approximate ReLU^2 by a Laplace function because the unbounded values of ReLU^2 and its gradient lead to unstable model training. MEGA shows competitive performance on Long Range Arena benchmark [Tay et al., 2021]. The same authors further proposed MEGA-chunk, which applies attention to each local chunk of a fixed length, yielding a linear complexity instead of the quadratic complexity in the original setting. In theory, the effective context can go beyond the chunk boundary, as the EMA sub-layer captures local contextual information near each token and propagates this information to the next layers.

3.2 Architecture adaptation

Attempts in architecture adaptation for context-aware MT can be broadly classified based on how they incorporate context [Abdul Rauf and Yvon, 2020]: *Single-encoder methods* incorporate context and the current sentence in the same encoder (Section 3.2.1), in contrast to *multi-encoder methods* (Section 3.2.2). Other approaches rely on a dedicated memory structure (Section 3.2.3) or a multi-pass decoding scenario (Section 3.2.4). We only detail recent findings complementary to previous surveys [Abdul Rauf and Yvon, 2020, Maruf et al., 2021, Castilho and Knowles, 2024].

3.2.1 Single-encoder architectures

The methods described in this section use the same encoder to process all their inputs, irrespective of their length. They apply both to Doc2Doc or Doc2Sent approaches. They have been introduced in NMT by [Tiedemann and Scherrer, 2017], who propose to concatenate the past sentences with the current one before feeding them to NMT models - in our terms, this implements Doc2Sent, with $\text{Block}(t, l, \mathbf{x}, \mathbf{y}) = (x_{t-1}, x_t)$ or (x_{t-1}, x_t, y_{t-1}).

This simple approach tends to be competitive with its more sophisticated counterparts, given that sufficient training data is available [Lopes et al., 2020]. The study by Fernandes et al. [2021] also confirms the overall merits of this single-encoder approach. These authors also show that the usefulness of past source sentences quickly vanishes after one or two sentences. They suggest, following Bawden et al. [2018], that the target-side context may be more useful than the source-side context. They finally propose to apply word dropout to the current sentence to improve context usage and translation quality.

Recently, several single-encoder architectures have been proposed to improve this naive approach, e.g. by adopting efficient transformer techniques, such as sliding window attention or separately modeling the local and global context.

FLAT-Transformer [Ma et al., 2020] was designed to encode the current sentence and its surrounding context in a unified flat encoder, with attention blocks spanning the entire input at the bottom layers then restricted to the focal sentence at the top layers, before cross-attention is applied. This somehow addresses issue # 3. G-transformer [Bao et al., 2021] considers each document as a group of sentences and assigns each token a group tag as a sentential index. It then uses group attention to enhance attention to the local context and further combines group attention with global attention using a gate-sum module at the top layers to enable cross-sentence interaction. G-transformer is trained on parallel sequences of up to 512 tokens, and it shows stable d-BLEU [Liu et al., 2020a] scores when translating inputs containing 512 and 1024 tokens.

Zheng et al. [2021] build a local context for each sentence. They reset token positions and introduce segment embeddings when computing the sentence-level attention, then retrieve global context encoding via segment-level relative attention, and perform a gated context fusion to integrate information from any sentence in the context. DOCFLAT [Wu et al., 2023] incorporates the global contextual information using a gated flat-batch attention to optimize document translation at the batch level. Before computing the self-attention of the vanilla Transformer, DOCFLAT flattens pseudo-documents with the original order to compute attention weights at the document level, which are subsequently reshaped back to the batch dimension. A neural context information gate is applied to control the influence of the global contextual information when updating hidden representations. The inference stage involves refining sentences that were translated independently in the first pass.

Herold and Ney [2023a] try to fix the alignment issues (#3 and #4) discussed above on page 7. For this, they apply a sliding window attention by dynamically aligning each translated token with the source sentences to circumscribe the span of potentially relevant source positions. Lasformer [Liu et al., 2023b] selects the top-K important tokens with respect to a lightweight attention score and masks unimportant tokens. This method

reduces the context length, hence the time complexity, while maintaining performance comparable to the evaluation of a complete context.

3.2.2 Multi-encoder methods

These approaches encode context information separately from the focal sentence, and combine them together either inside or outside the decoder [Li et al., 2020, Abdul Rauf and Yvon, 2020]. As they handle differently the context and the focal sentence, they only apply to Doc2Sent architectures.

For integration outside the decoder, the focal sentence and its context are first encoded using a source side network, for instance a specific encoder [Voita et al., 2018, Zhang et al., 2018], or hierarchical attention networks (HANs) [Miculicich et al., 2018, Maruf et al., 2019, Yin et al., 2021]. These representations are then fused by a gated sum before being fed to the decoder. The gating mechanism enables the model to learn which additional contextual information should be included.

Regarding methods integrating context inside the decoder, the target word generation process can separately attend to the source and the context representations, in addition to the available target-side prefix. Depending on the specific architecture, this combination of source and context attention can be performed *sequentially*, as in [Tu et al., 2018, Zhang et al., 2018] or *in parallel* [Jean et al., 2017, Bawden et al., 2018, Stojanovski and Fraser, 2018]. Compared to single-encoder approaches, such strategies also enable the use simpler processing modules for the context, which is arguably less informative for the translation than the focus sentence.

3.2.3 Cache-based Approaches

Cache-based methods store a short-term memory of the recent context to boost the probabilities of target words that have recently generated [Maruf and Haffari, 2018, Tu et al., 2018, Kuang et al., 2018, Yang et al., 2019, Dobreva et al., 2020]. This can be done, for instance, using a *continuous representation* of the cache to store recent context history for use in future decoding steps [Tu et al., 2018].

Cache slots are pairs of key-value vectors, with the keys being attention context vectors, and values corresponding to the decoder states collected from previous translations. The cache can also be added to a pre-trained NMT model with fine-tuning, by updating only the new parameters related to the cache. Kuang et al. [2018] use both a dynamic cache of past words and a "topical cache" of semantically relevant keywords and compute the final word prediction probability via a gating mechanism by combining the probability estimated from the cache with the probability computed by the decoder. [Dobreva et al., 2020] use a context tag to provide the encoder with information about the document structure and a fixed-size topic cache and dynamic cache similar to the proposal of Kuang et al. [2018]. These are concatenated and passed to the output layer, thereby helping to improve the estimation of the probability distribution over words. Similar techniques have been introduced in the form of cache LMs or related mechanisms (see the discussion in [Yvon and Abdul Rauf, 2020, Section 3.1.3]).

3.2.4 Multi-pass approaches

Multi-pass systems usually introduce an additional computational component that helps to refine translations produced by context-free first-pass systems and make them more globally coherent, with the help of document-level monolingual data [Xiong et al., 2019, Voita et al., 2019a, Yu et al., 2020, Kang et al., 2020]. Such approaches are easy to implement as they only rest on the availability of sufficiently large monolingual documents in the target domain and do not require changing the first pass system. However, the two-stage generation introduces some undesirable computational complexity, because of the need to train multiple systems, and to repeatedly encode the same text during inference. This process may result in cascading errors [Xiong et al., 2019].

3.3 Data augmentation

Another research direction is data augmentation. As aforementioned, long-context MT aims to improve the translation of long sequences and the translation of diverse discourse phenomena. However, document-level parallel corpora are scarce, and it is expensive to develop new datasets. Several studies also report suboptimal results for models directly trained on parallel documents (Doc2Doc) [Zhang et al., 2018, Liu et al., 2020b, Tang et al., 2021]. Consequently, pretraining a sentence-level MT system, then fine-tuning it on document-level data has become a well-recognised training procedure for document-level MT, for instance in [Liu et al., 2020b, Lopes et al., 2020].

In the fine-tuning step, MT models are required to learn with very long source and target contexts, with potentially large position indices and long inter-sentence contexts, which are unseen during the pretraining stage. To improve training efficiency, *multi-resolution training* [Sun et al., 2022] was proposed to balance the sequence length distribution of the training data. It consists in splitting each input parallel document pair into k parallel pseudo-documents, with $K \in \{1, 2, 4, 8, \ldots\}$. For instance, an 8-sentence document is divided into 15 pseudo-documents, with one 8-sentence document, two 4-sentence sub-documents, four 2-sentence sub-documents, and eight sequences of one sentence. This method is helpful to mitigate the sparsity problem and helps Transformer models to perform document-level MT [Sun et al., 2022, Li et al., 2022].

The sparsity of discourse phenomena depending on inter-sentence context is another challenge to improve translation quality. Several approaches are proposed to increase the need for context information for translation so that models learn not to ignore contextual information when it is required. A representative of this trend is discussed by Lupo et al. [2022a], who propose to split sentences as if they were made of two sentences, thereby creating two supplementary examples of cross-sentence coreferences.

More recently, the *Importance Aware Data Augmentation* (IADA) algorithm was introduced by Wu et al. [2024b]. It augments the training data based on token importance information estimated using the norm of hidden states or training gradients. The main principle is to automatically detect and mask the important tokens in the focal sentence while perturbing less important tokens in the context so that models are encouraged to recover masked tokens using the useful contextual information (see Figure 2). In con-

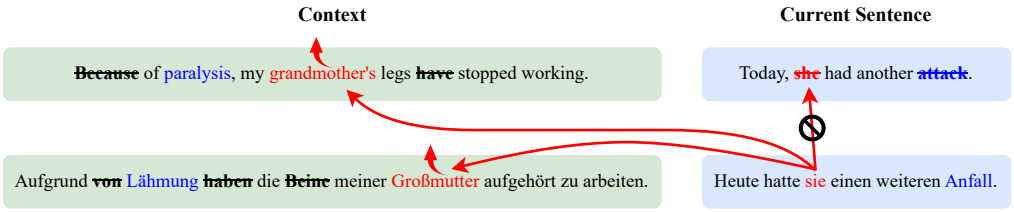


Figure 2: An example illustrating the action of IADA to increase the need for cross-sentential context, borrowed from [Wu et al., 2024b]. ~~Strikethrough~~ indicates perturbation.

sideration of the increased training difficulty from the corpora corrupted by the IADA algorithm, the loss function is also adapted to ensure training efficiency.

3.4 Training strategies

In addition to data augmentation, training strategies, such as applying multilingual denoising pre-training, refining positional information and adjusting the loss function, represent another active avenue for more effectively integrating contextual information.

Inspired by mBART25 [Liu et al., 2020a], DOCmT5 [Lee et al., 2022] was developed by continuing the pretraining of mT5 [Xue et al., 2021] using a reconstruction objective on synthetic multilingual parallel documents of up to 512 tokens. The training procedure involves reordering randomly shuffled sentences, recovering spans randomly masked in the same way as mT5, then translating into target languages. Experiments show improved d-BLEU scores, although systematic context-aware analysis are not reported.

Since information embedded in local attention is dense, whereas information in global attention is sparse and further diluted with longer context, introducing constraints to increase attention weight on the local context is a meaningful approach to assist in long context translation. For example, Lupo et al. [2022b] propose an improved concatenation method, which uses *context discount* (CD) and *segment-shifted position* to enhance attention on the current sentence. When training with an extended context, CD is used to discount the part of the loss corresponding to the generation of the target context. The authors focus on the sliding context window scenario to translate the current sentence x_k in the context of the $K-1$ past sentences. More precisely, they apply a discount $\lambda \in [0,1]$ to the context tokens' loss function as follows:

$$\mathcal{L}_{\text{CD}}(x_K^j, y_K^j) = \lambda \mathcal{L}_{context} + \mathcal{L}_{current} \tag{7}$$

where x_K^j and y_K^j denote the current sentence, preceded by $K-1$ past context sentences. The segment-shifted position technique further increases position distance between tokens from different sentences, where a constant *shift* is applied to every sentence belonging to the input sequences $X = x_1, \ldots, x_K$, to transform the token position from t to $t + k \times shift$ for all tokens of x_k, with $k \in \{1, \ldots, K\}$.

P-transformer [Li et al., 2022] adds absolute positional embedding P_{abs} to queries and keys before the dot-product of all attention operations. It also injects relative positional

embedding P_{rel} into self-attention, computing the attention vector as:

$$\text{softmax}(\frac{(Q + P_{abs})(K + P_{abs})^T}{\sqrt{d_k}} + P_{rel})V \tag{8}$$

In their experiments, long documents are split into sub-documents of up to 512 tokens. The translation quality of full documents evaluated using d-BLEU and ds-BLEU [Peng et al., 2020] was improved, and the P-transformer generalizes much better than the baseline Doc2Doc system when translating text with unseen lengths of 1024 or 2048 tokens [Li et al., 2022, Fig. 7].

Lupo et al. [2023] also study positional embeddings. They compare several segment embedding methods, including one-hot, sinusoidal, and learned segment embeddings, to explicitly encode sentence positions using the segment-shifted position scheme. The segment embedding (SE) is then concatenated with the positional embedding (PE) to form *position-segment embeddings* (PSEs), to avoid information damage which could be seen if the addition operation is used between SE and PE. The PSE, the dimension of which is $d_{PSE} = d_{SE} + d_{PE} = d_m$, is subsequently added to token representations at the input of every Transformer block. In general, their impact on BLEU scores is marginal. Results evaluated on contrastive sets of language pairs EN–DE and EN–RU confirm the effectiveness of using a context discount (see above), while PSEs appear to hurt the translation quality for EN–DE.

MT systems trained on long texts such as parallel documents are also reported to overfit the length distribution of the training set, especially when the training corpora are constrained by a predetermined maximum length [Zhuocheng et al., 2023]. To mitigate this effect, a dynamic length sampling method was proposed to progressively increase the input sequence's length during training and ensure a more uniform length distribution [Zhuocheng et al., 2023]. This idea is reminiscent of the data augmentation strategies discussed in Section 3.3 and also aims to manipulate the training length distribution. In the same work, the authors introduced *length-aware attention* (LAA), which amounts to multiplying the QK^T product with an additional factor $\frac{logL}{\log \overline{L}}$, with L the document length and $\overline{L}$ the average document length in the current batch. This has the effect of flattening the attention weights for short sentences, and contrarily of making them more concentrated for long sentences. This is again related to issue #3 in Section 2. These methods are combined with a sliding window decoding strategy that incrementally truncates the past source and target contexts to enable the translation of very long sequences.

3.5 Decoding strategies

Researchers have also explored the impact of decoding search strategies on translation performance. Stahlberg and Byrne [2019] investigate the search errors in beam search with beam size values from one to 100. The authors combine beam search and depth-first search to find the global best score, while discovering that beam search failed to find these scores but preferred shorter and even empty translations. This phenomenon affected more long input sentences.

Herold and Ney [2023b] compare several inference strategies (both Block2Block and Block2Sent) for document-level translation, such as decoding non-overlapping full segments of length k ($k = 3$), decoding the focal sentence with either the $k - 1$ preceding sentences or the $k - 1$ following sentences within the block (with $k = 3$) , two-pass decoding to refine sentence-level translations, generating full documents, with or without sentence-level beam search. A comparison with sentence-level decoding demonstrates the utility of contextual information, while no significant difference was observed between the various context-aware decoding methods. However, given the small size of the blocks used in this work, it resorts more to Context-Aware MT than to Document-level MT.

4 LLM-based methods

The introduction of large language models (LLMs) such as chatGPT[6] have received world-wide attention and revolutionized the whole field of NLP studies, improving state-of-the-art results for multiple tasks. Their multilingual nature and their capacity to handle long-range dependencies have provided new possibilities for context-aware machine translation task [Vilar et al., 2023, Briakou et al., 2023, Moslem et al., 2023, Zhang et al., 2023, Bawden and Yvon, 2023, Pang et al., 2024, Qin et al., 2024, Castilho and Knowles, 2024]. LLMs can be used both in a Doc2Sent or, increasingly, in a Doc2Doc mode, presenting the complete input text in the LLM instruction window.

Several works have explored the context-aware translation ability of LLMs through different strategies, including parameter-frozen methods such as zero-shot prompting, in-context learning (Section 4.1), and parameter-tuning strategies (Section 4.2). In Section 4.3, we finally discuss various ongoing issues regarding LLM-based translation, such as sentence alignment difficulties, positional bias, etc.

4.1 In-context learning

LLMs are known for their multi-task abilities, which can be manipulated through the use of relevant prompts [Brown et al., 2020]. Primary studies in context-aware MT using LLMs have explored in-context learning methodologies using various prompting paradigms, keeping the parameters frozen. Related works include the integration of few-shot examples with similar ambiguous contexts as input sequence for better disambiguation [Iyer et al., 2023] – where context can typically include segments preceding the focus sentence; the use of special tags to mark sentence boundaries in documents [Zhang et al., 2023]; the introduction of chain-of-dictionary prompting [Lu et al., 2023]; the selection of prompt pattern candidates generated with an LLM [Wang et al., 2023a]; and the examination of prompts of different context structures, with or without natural language instruction [Wu et al., 2024a].

Moslem et al. [2023] propose to select fuzzy matches of the current source sentence as examples in few-shot learning to improve translation quality for domain adaptation. More

[6] https://chatgpt.com/

precisely, the authors extract parallel segments from a domain-specific dataset, based on sentence-level embedding similarities, and compared prompts with one- to ten-shot fuzzy matches along with zero-shot, and a random two-shot prompt pattern. Experiments on GPT3.5 show that few-shot fuzzy match examples lead to significant improvement of translation quality and outperform conventional systems such as DeepL, NLLB-3.3B for high-resource language pairs like EN-ES, EN-FR, and EN-ZH according to spBLEU, chrf++, and COMET. This method is further augmented by appending the most similar two fuzzy matches with in-domain terminologies to assist in terminology translations in a specific domain. Moreover, this work also shows the possibility of refining low-resource language translation of encoder-decoder MT systems using very deep LLMs like GPT3.5. Even though the context is augmented, all these scenarios still produce sentence-based translations.

Hendy et al. [2023] investigated the capacity of the GPT *text-davinci-003* model to translate small blocks containing more than one sentence at a time (a Block2Block approach). . They report competitive zero-shot performance for documents of 2 to 32 sentences, and comparable 5-shot performance for documents of 10 sentences in the domain of news, evaluated using a range of automatic metrics including d-BLEU and the average of sliding window COMET adapted for document-level MT. They also reported alignment issues when translations using LLMs, due to either the merge of two sentences into one or due to ellipsis.

Wang et al. [2023a] conducted a more comprehensive examination, engaging not only metrics like d-BLEU, contrastive test suites [Voita et al., 2019b] and human evaluation but also a comparison of diverse translation systems (DeepL, G-transformer, multi-resolution Doc2Doc model, DocRepair, etc.). At the time of this publication, GPT4 showed impressive performance and ranked among the best systems across all experiments based on human evaluation, despite lagging behind DocRepair for deixis, lexical consistency, and inflection ellipsis.

In addition, Karpinska and Iyyer [2023] thoroughly compared the translation quality of sentence-level, in-context sentence-level, and paragraph-level literary translation across 18 language pairs by few-short prompting GPT-3.5, assessing the output with a human evaluation inspired by MQM [Lommel et al., 2014, Freitag et al., 2021]. Paragraph-to-paragraph[7] translations are preferred compared to the two other scenarios for various aspects, such as fewer mistranslations, grammatical errors, and inconsistencies, demonstrating the utility of longer context in translation tasks, while critical errors such as omission and mistranslation still persist.

To conclude, LLMs seem capable of extracting contextual information from prompts. A well-chosen prompt is however required for LLMs to achieve optimal performance. For the context-aware translation task, applying in-context learning is, for most LLMs,[8] not sufficient to surpass traditional encoder-decoder translation systems, especially for low-resource tasks, because of their general-purpose and English-centric properties [Zhang

[7] Most paragraphs here consist of four to nine sentences, with a minimum of two sentences and a maximum of 28 sentences.

[8] With the exception, perhaps, of closed-source models such as GPT3.5 and GPT4.

et al., 2023, Vilar et al., 2023, Hendy et al., 2023].

4.2 LLM-based Training strategies

To train LLMs as better translators, the NLP community also examined the efficiency of the standard pretraining+finetuning strategy, usually compared with translation using in-context learning and encoder-decoder systems [Iyer et al., 2023, Zhang et al., 2023].

For instance, Zhang et al. [2023] evaluate 15 open-source LLMs on sentence-level and document-level MT tasks, comparing zero-shot prompting, few-shot learning, and fine-tuning strategies, where the pseudo-documents used during fine-tuning or in-context learning contained 5, 10, or 15 sentences, and were translated as a whole. Results and analysis based on BLEU and COMET scores illustrate that only prompting is not sufficient for document-level MT, and moderate-size LLMs can outperform larger counterparts. Moreover, QLoRA fine-tuning is more efficient than full fine-tuning, which is also recognized in other publications [Alves et al., 2023, 2024]. In addition, Alves et al. [2023] observe the degradation of in-context learning capabilities and propose fine-tuning LLMs using data mixed with zero-shot and few-shot instructions to alleviate this problem.

To exploit the potential of decoder-only LLMs for the translation task, Xu et al. [2024a] introduce an efficient two-stage fine-tuning method for moderate-size LLMs of 7B or 13B parameters and release ALMA. ALMA is a fine-tuned version of LLaMA2, with comparable performance to GPT-3.5 in terms of average BLEU and COMET score of translations for 5 English-centric language pairs. This training recipe begins with continued pretraining on target-side monolingual data to learn general linguistic knowledge, followed by supervised fine-tuning using a small amount of high-quality parallel data.[9]

Based on this, Alves et al. [2024] release TowerBase and TowerInstruct. TowerBase is the result of pretraining LLaMA2 (7B or 13B) on a high-quality corpus, comprising one-third parallel data along with two-thirds monolingual data from all target languages. TowerInstruct derives from fine-tuning TowerBase on TowerBlock, a well-curated multi-task corpus, which comprises instances of multiple translation-related tasks, including sentence-level MT, context-aware MT, error-span detection, conversation, etc.

Wu et al. [2024a] also apply a two-stage training recipe when adapting moderately-sized LLMs (LlaMA2-7B, BLOOM-7B, and VICUNA-7B) for document-level MT. The authors explore different prompting strategies to integrate contextual information from the 3 preceding consecutive sentences, and compare full-parameter fine-tuning (FFT) and fine-tuning with LORA [Hu et al., 2022]. They discovered that fine-tuning moderate-size LLMs results in models that outperform GPT-4 in some translation tasks, while instruction-tuning may be hurtful for subsequent supervised fine-tuning performance. In their experiments, Parameter Efficient Fine-Tuning (PEFT) approaches such as LORA outperform Full Fine-Tuning (FFT), while FFT requires less training data. Their empirical results also show better generalization ability in translating out-of-domain text of LLM-based translation systems than the traditional encoder-decoder ones.

[9] They also reported that large amounts of parallel data may lead to catastrophic forgetting of pretrained knowledge.

Furthermore, Xu et al. [2024b] propose Contrastive Preference Optimization (CPO) to mitigate the generation of adequate but imperfect translations. Their triplet preference training data is automatically constructed according to the average quality estimation score of reference and automatic translations. The CPO objective first removes the component relevant to pretrained models in Direct Preference Optimization loss function via an approximation that reduces the memory and time complexity. Moreover, it is augmented with a behavior cloning (BC) regularizer deviated from the expectation of Kullback–Leibler (KL) divergence to encourage the model to learn the characteristics of preferred data. Empirical experiments are built on ALMA-13B-LoRA, resulting in ALMA-13B-R, which outperforms GPT-4 on average for 10 translation directions.

Very recently, the work of Wu et al. [2024c] explores multi-agent collaboration for ultra-long web novel translation. Human evaluation and LLM evaluation prefer their translations over references, despite suboptimal d-BLEU scores.

4.3 Discussion

Despite the reported exciting performance of LLMs in translation, it is still unclear how and to what extent long-term context contributes to the translation quality, especially for the translation of discourse phenomena, due to the lack of adequate recognized context-aware automatic metric for this purpose.

Publications in other NLP tasks have reported that LLMs show bias towards certain token positions. In particular, this has been observed for question-answering and key-value retrieval tasks: locating the key information at different positions of the input sequences has a clear impact on the overall performance [Liu et al., 2024, Saito et al., 2024, An et al., 2024, Levy et al., 2024]. In particular, it seems that LLMs struggle to recognize useful information in the middle of input prompts [Liu et al., 2024]. Another well-known phenomenon is the degradation of performance with the increase of the input length, which has been thoroughly examined in [Levy et al., 2024].

Saito et al. [2024] report that LLMs also struggle to extract context at the end of sequences after fine-tuning, and they suggest using denoising auto-regressive fine-tuning strategies to mitigate the problem. Alternatively, with FILM-7B, An et al. [2024] apply a novel information-intensive data-driven training strategy, which randomly places short segments containing crucial information in the whole long context of documents with balanced length distribution, thereby significantly alleviating the position bias. However, similar experiments for machine translation have not yet been explored and therefore would be a useful direction for future work.

In addition, translation using LLMs suffers from several serious shortcomings, including hallucination [Ji et al., 2023], wrong language prediction, and over-generation [Bawden and Yvon, 2023]. Generation through closed-source LLMs also faces problems of data leakage, copyright issues, and lack of reproducibility [Balloccu et al., 2024].

5 Extrapolating beyond Training Lengths

5.1 Position Encoding for Long Sequences

Transformer models are typically trained with a maximal length limit, meaning that any longer input will have to be chopped into several parts. The lack of ability of Transformer-based models to extrapolate to longer document lengths, unseen in training, is a well-known difficulty that has attracted more and more attention in the LLM era, leading notably to the emergence of several novel positional encoding strategies [Zhao et al., 2024]. In this section, we aim to summarize representative findings along these lines. Although their effects on MT tasks are not always reported, they are noteworthy directions to improve MT with long context in future work.

Since the default attention mechanism is position-agnostic, it is necessary for Transformer-based models to explicitly incorporate positional information through positional encodings. Positional encoding is an imprecise term denoting a collection of methods for the injection of positional signals, including a fixed mapping function (e.g. the absolute sinusoidal position signal in the vanilla Transformer of Vaswani et al. [2017]), and also positional encodings, which depend on learnable parameters (e.g. the fixed-size matrix for trained positional embeddings used in BERT [Devlin et al., 2019]) or positional biases integrated into attention computation (e.g. as used in ALIBI [Press et al., 2022]).

However, most models, including LLMs, are trained with a pre-defined context window size. Several publications have reported performance degradation when processing input sequences longer than this predefined max length [Chen et al., 2023, Liu et al., 2024].

Such techniques have been superseded in most recent LLMs by the use of RoPE (Rotary Positional Embedding) [Su et al., 2024]. In a nutshell, RoPE parameterizes the absolute positions with a rotation matrix, which then allows to integrate relative positions in the attention weights and uniformly boost the attention over neighboring positions. RoPE has also been found to work well with various position interpolation methods. This is reflected in Table 1.

Models	Positional Encoding	Year
T5	T5 relative bias	2020-2023
BLOOM	ALiBi	2023
LLama	RoPE + Dynamique-NTK	2023
Qwen	RoPE + Dynamique-NTK	2023
Falcon	RoPE + PI	2023
PaLM	RoPE	2022

Table 1: Positional encoding methods used in recent language models.

Position interpolation [Chen et al., 2023] was proposed to linearly down-scale position indices $0 \dots L_{\text{input}}$ to $0 \dots L_{\text{max}}$ using a factor $\frac{L_{\text{input}}}{L_{\text{max}}} \in [0, 1]$ before applying RoPE to match the original context window size, thereby enabling to process documents of

arbitrary length. This method however requires fine-tuning the model with these non-integer position indices, before it can properly handle documents whose length exceeds the $L_{\max}$ limit. Its introduction has led to active discussion in the NLP community. This is because linear interpolation results in an information loss in the local context, which corresponds to the high-frequency field of RoPE's Fourier space: this means that the interpolation mostly distorts the distance between neighbor positions. As a possible mitigation NTK-aware[10] scaled RoPE introduced a non-linear interpolation via the base change of position indices.[11]

Dynamic-NTK was introduced as an enhancement of NTK-aware scaling, involving dynamically choosing the correct scale parameter according to each input sequence length to interpolate high frequencies less and low frequencies more. This method can take effect without fine-tuning the pretrained models, although it is difficult to determine the optimal base [Peng et al., 2023b].[12] Alternatively, the NTK-by-part approach involves scaling position indices for the number of rotations of the given RoPE dimension, to preserve more information about the relative position. YaRN [Peng et al., 2023b] generalizes the ideas of Dynamic-NTK and NTK-by-part. It additionally introduces a temperature t that scales attention weight before the softmax to further reduce its perplexity.

Dynamic NTK-aware approaches are integrated into the implementation of Llama2 [Touvron et al., 2023] and Qwen [Bai et al., 2023]. These methods have been mainly tested on RoPE-based models, while its application to ALIBI [13] also improved the performance of BLOOM [Workshop et al., 2023] for long input text.

An alternative is given FIRE [Li et al., 2024], which applies a learnable continuous function such as MLP to map input position to position biases $b(i, j)$. It applies progressive interpolation using $b(i, j) = f_\theta(\frac{i-j}{max(L,i)})$ to always stay in the training domain, with L a learnable threshold to preserve the model performance on short input. A log transformation is also used to amplify local attention.

Another way to extend the context length via position encoding is *position extrapolation*. Ruoss et al. [2023] randomly map position indices to a much larger interval while preserving the original word order, thus exposing the model to large position indices derived from short training sequences. More recently, Zhu et al. [2024] proposed *Positional Skip-wisE* (PoSE) fine-tuning, facilitating context window extension using documents that are shorter than the predefined max length. It is also compatible with RoPE-based models and position interpolation strategies. PoSE 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. The number of chunks N is regarded as a trade-off for efficiency due to its negative impact on the perplexity of a large N.

[10] NTK stands for *Neural Tangent Kernel*

[11] https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have

[12] https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases

[13] https://github.com/keezen/ntk_alibi/blob/main/readme_en.md

5.2 Other Techniques for Very Long Contexts

A considerable number of techniques have been proposed to process ultra-long contexts outside the scope of MT. To begin, Pawar et al. [2024], Wang et al. [2024] and Naveed et al. [2024] provide comprehensive surveys that summarize approaches for context-length extension in LLMs, including adapted position encodings (Section 5.1), specialized attention mechanisms [Sun et al., 2023, Ding et al., 2023, Yang, 2023, Han et al., 2024, Chen et al., 2024], window-based approaches [Jin et al., 2024], prompt compression [Jiang et al., 2023b] and memory/retrieval augmented techniques [Rubin and Berant, 2023a, Mohtashami and Jaggi, 2023, Wang et al., 2023b, Liu et al., 2023a, Tworkowski et al., 2023, Packer et al., 2024].

Publications in automatic summarization [Koh et al., 2022, Phang et al., 2023] and simultaneous translation [Iranzo-Sánchez et al., 2022, Xiao et al., 2024] also explored diverse approaches to the processing of long input.

Another avenue is to construct architectures that would be more efficient than the Transformer for long sequences. A series of linear RNNs have thus been proposed, such as the variant of S4 [Gu et al., 2022b,a], H3 [Fu et al., 2023], RWKV [Peng et al., 2023a] and Mamba [Gu and Dao, 2024]. Performance on the Long Range Arena benchmark [Tay et al., 2021] and downstream tasks [Amos et al., 2024, Le Bronnec et al., 2024] shows the great potential of these architectures to deal with long-range dependencies. Vardasbi et al. [2023] test S4 for sentence-level translation, and reported that combining the transformer encoder and S4 decoder leads to optimal performance, improving the BLEU score of longer sentences.

Document layout also serves as an informative context for better text understanding. A number of recent works consider the two-dimensional layout location of tokens on a page instead of their linear positions in a long raw text [Xu et al., 2020, Nguyen et al., 2021, 2023]. It is also an interesting direction when translating structured documents like scientific articles.

6 Document-level parallel corpus

This conclusive section aims to provide an overview of parallel document corpora in addition to resources mentioned in Abdul Rauf and Yvon [2020, Section 5.1]:

- SciPar [Roussis et al., 2022]: a multilingual collection of parallel abstracts from openly published bachelor, master and doctoral theses across various fields.
- A collection of biomedical abstracts Cochrane, EDP, Medline and Sielo, used by [Abdul Rauf and Yvon, 2024]
- SCAT [Yin et al., 2021]: 14K EN–FR translations containing supporting context for ambiguous translations.
- Karpinska and Iyyer [2023] released 20 parallel literary paragraphs for 18 language pairs, extracted from recently published translations of novels.

- Al Ghussin et al. [2023] extracted parallel paragraphs from ParaCrawl [Bañón et al., 2020] using automatic sentence alignments, and released an EN–DE corpus.
- TANDO [Gete et al., 2022]: a Basque-Spanish document-level parallel corpus, containing literary documents, news and subtitles, and contrastive sets for the contextual phenomena of gender and register.
- JaParaPat [Nagata et al., 2024] is a Japanese–English parallel patent application corpus including parallel documents.
- Parallel document corpus (PDC) [Sun et al., 2022] is a 60k web-crawled set of parallel news documents in diverse domains, including politics, finance, health, and culture.

7 Conclusion and outlook

In this survey, we investigated the advancement of techniques for incorporating long contexts into machine translation. We first analyzed the benefits and challenges of introducing full documents in translation using the Doc2Doc scheme, clarifying the motivation and interest of this research field. Subsequently, we gave an overview of relevant approaches designed for traditional encoder-decoder architectures and for LLMs, with a brief review of efficient transformers to provide additional background.

Traditional methods attempt to (1) adapt model architectures, in particular incorporating contextual information in the encoder or cache memories, or refining context-agnostic translation through a multi-pass system; (2) carry out data augmentation to create datasets with more balanced length distributions or richer context-dependent phenomena; and (3) improve training strategies using multilingual denoising pretraining, adapted loss functions, and augmented positional encodings. Reported results indicate that the simple concatenation approach is a robust baseline when sufficient document-level data is available to fine tune a sentence-base model,[14] that data distribution plays an important role in MT training, that pretraining enhances translation quality and that improved positional encodings can mitigate performance degradation for the translation of long inputs.

In recent years, multiple LLM-based methods have emerged, primarily focusing on the potential of LLMs in context-aware machine translation through in-context learning and performance enhancement via fine-tuning. These studies underscore the critical importance of the quality of demonstrations in few-shot learning and of the parallel corpus. Research in this area also favors parameter-efficient fine-tuning and continued monolingual pretraining of all target languages before fine-tuning with high-quality parallel documents. Additionally, we briefly discussed specific issues related to LLM-based translation, such as sentence alignment difficulties, and potential positional biases.

[14] Note that this method has only been tested for relatively short documents, comprising a dozen of sentences.

Before listing some recently released document-level parallel corpora, we briefly reviewed other approaches for context window extensions applied to different tasks, hoping to inspire future work in context-aware MT. Although our focus is on MT techniques, it is essential to highlight that the study of context-aware metrics is a crucial research sub-field of context-aware MT. These metrics are vital for measuring the improvements of proposed approaches and guiding future explorations.

For our future work, it first seems that one of the essential factors to achieve better Doc2Doc translation is high-quality data. "High quality" means that the parallel corpus is well aligned [Xu et al., 2024a], the corpus is sufficiently informative [Lupo et al., 2022a, Wu et al., 2024b], and sufficiently balanced (for example, in sequence length [Sun et al., 2022]). Some novel techniques such as *Retrieval-Augmented Generation* [Guu et al., 2020, Lewis et al., 2020], which assists text generation by retrieving relevant information from an external knowledge base or from the past generation history [Rubin and Berant, 2023b], can effectively augment available context when translating target documents. The training strategies ought to facilitate the access of relevant information in data (e.g. using contrastive learning like CPO or adapted loss function) and to avoid introducing bias (e.g. length bias [Zhuocheng et al., 2023]). Regarding the architecture, it is also beneficial to assess the ability of recently proposed linear RNNs such as S4 and Mamba [Gu and Dao, 2024] in Doc2Doc translation tasks to guide future NMT research [Pitorro et al., 2024].

Bibliography

Sadaf Abdul Rauf and François Yvon. Document level contexts for neural machine translation. Research Report 2020-003, LIMSI-CNRS, December 2020. URL https://hal.science/hal-03687190. [Cited on pages 5, 6, 13, 15, and 25.]

Sadaf Abdul Rauf and François Yvon. Translating scientific abstracts in the bio-medical domain with structure-aware models. *Computer Speech and Language*, 87:101623, 2024. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2024.101623. URL https://www.sciencedirect.com/science/article/pii/S0885230824000068. [Cited on page 25.]

Eneko Agirre and Mark Stevenson. Word Sense Disambiguation. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 06 2022. ISBN 9780199573691. doi: 10.1093/oxfordhb/9780199573691.013.28. URL https://doi.org/10.1093/oxfordhb/9780199573691.013.28. [Cited on page 5.]

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 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*, pages 31–40, Alicante, Spain, May 2018. URL https://aclanthology.org/2018.eamt-main.1. [Cited on page 7.]

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. ETC: Encoding

long and structured inputs in transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.19. URL https://aclanthology.org/2020.emnlp-main.19. [Cited on page 12.]

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. Exploring paracrawl for document-level neural machine translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1304–1310, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.94. URL https://aclanthology.org/2023.eacl-main.94. [Cited on page 26.]

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. Steering large language models for machine translation with finetuning and in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.744. URL https://aclanthology.org/2023.findings-emnlp.744. [Cited on page 21.]

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024. [Cited on page 21.]

Ido Amos, Jonathan Berant, and Ankit Gupta. Never train from scratch: Fair comparison of long-sequence models requires data-driven priors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PdaPky8MUn. [Cited on page 25.]

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. Make your llm fully utilize the context, 2024. URL https://arxiv.org/abs/2404.16811. [Cited on page 22.]

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. [Cited on page 24.]

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In

Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.5. [Cited on page 22.]

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417. [Cited on page 26.]

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.267. URL https://aclanthology.org/2021.acl-long.267. [Cited on pages 7, 9, and 14.]

Rachel Bawden and François Yvon. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.16. [Cited on pages 19 and 22.]

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 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, June 2018. doi: 10.18653/v1/N18-1118. URL https://www.aclweb.org/anthology/N18-1118. [Cited on pages 5, 7, 14, and 15.]

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150. [Cited on page 12.]

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. Unlimiformer: Long-range transformers with unlimited length input. In *Thirty-seventh Conference*

on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=lJWUJWLCJo. [Cited on page 12.]

Eleftheria Briakou, Colin Cherry, and George Foster. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.524. URL https://aclanthology.org/2023.acl-long.524. [Cited on page 19.]

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. [Cited on pages 12 and 19.]

Sheila Castilho and Rebecca Knowles. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, page 1–31, 2024. doi: 10.1017/nlp.2024.7. [Cited on pages 6, 7, 13, and 19.]

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595. [Cited on page 23.]

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6PmJoRfdaK. [Cited on page 25.]

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. [Cited on page 12.]

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 2174–2184, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1223. URL https://aclanthology.org/D19-1223. [Cited on page 12.]

Nicolas Dahan, Rachel Bawden, and François Yvon. Survey of existing document-level metrics for machine translation. Technical report, ISIR-CNRS,Inria, 2024. [Cited on page 7.]

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.96. URL https://aclanthology.org/2023.wmt-1.96. [Cited on page 6.]

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423. [Cited on page 23.]

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens, 2023. URL https://arxiv.org/abs/2307.02486. [Cited on page 25.]

Radina Dobreva, Jie Zhou, and Rachel Bawden. Document sub-structure in neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3657–3667, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.451. [Cited on page 15.]

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 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)*, pages 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. [Cited on pages 7 and 14.]

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 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, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.36. [Cited on page 7.]

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human

evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87. [Cited on page 20.]

Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=COZDy0WYGg. [Cited on page 25.]

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. TANDO: A corpus for document-level machine translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.324. [Cited on page 26.]

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://openreview.net/forum?id=AL1fq05o7H. [Cited on pages 25 and 27.]

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=yJE7iQSAep. [Cited on page 25.]

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022b. URL https://openreview.net/forum?id=uYLFoz1vlAC. [Cited on page 25.]

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023. [Cited on page 7.]

Liane Guillou and Christian Hardmeier. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology/L16-1100. [Cited on page 7.]

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training, 2020. [Cited on page 27.]

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models, 2024. [Cited on page 25.]

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. On the blind spots of model-based evaluation metrics for text generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.674. URL https://aclanthology.org/2023.acl-long.674. [Cited on page 7.]

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023. [Cited on pages 20 and 21.]

Christian Herold and Hermann Ney. 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, July 2023a. Association for Computational Linguistics. URL https://aclanthology.org/2023.codi-1.15. [Cited on pages 9 and 14.]

Christian Herold and Hermann Ney. On search strategies for document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12827–12836, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.811. URL https://aclanthology.org/2023.findings-acl.811. [Cited on pages 6 and 18.]

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. [Cited on page 21.]

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9099–9117. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hua22a.html. [Cited on page 13.]

Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. From simultaneous to streaming machine translation by leveraging streaming history. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6972–6985, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.

18653/v1/2022.acl-long.480. URL https://aclanthology.org/2022.acl-long.480. [Cited on page 25.]

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. Towards effective disambiguation for machine translation with large language models. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.44. URL https://aclanthology.org/2023.wmt-1.44. [Cited on pages 19 and 21.]

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. Does neural machine translation benefit from larger context?, 2017. [Cited on page 15.]

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730. [Cited on page 22.]

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023a. [Cited on page 12.]

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression, 2023b. [Cited on page 25.]

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 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*, pages 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. [Cited on page 7.]

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Chia-Yuan Chang, and Xia Hu. Growlength: Accelerating LLMs pretraining by progressively growing training length, 2024. URL https://openreview.net/forum?id=vmlwllg7DJ. [Cited on page 25.]

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. Challenges in context-aware neural machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.943. URL https://aclanthology.org/2023.emnlp-main.943. [Cited on pages 6 and 7.]

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722, December 2017. doi: 10.1162/COLI_a_00298. URL https://aclanthology.org/J17-4001. [Cited on page 5.]

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.175. URL https://aclanthology.org/2020.emnlp-main.175. [Cited on page 16.]

Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.41. URL https://aclanthology.org/2023.wmt-1.41. [Cited on pages 9, 20, and 25.]

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB. [Cited on page 12.]

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204. [Cited on page 10.]

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Computing Surveys*, 55(8):1–35, December 2022. ISSN 1557-7341. doi: 10.1145/3545176. URL http://dx.doi.org/10.1145/3545176. [Cited on page 25.]

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. Modeling coherence for neural machine translation with dynamic and topic caches. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1050. [Cited on page 15.]

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels,

Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://aclanthology.org/D18-1512. [Cited on page 5.]

Florian Le Bronnec, Song Duong, Mathieu Ravaut, Alexandre Allauzen, Nancy Chen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. LOCOST: State-space models for long document abstractive summarization. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1144–1159, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.69. [Cited on page 25.]

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. DOCmT5: Document-level pretraining of multilingual language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 425–437, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.32. URL https://aclanthology.org/2022.findings-naacl.32. [Cited on page 17.]

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024. URL http://arxiv.org/abs/2402.14848. [Cited on page 22.]

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf. [Cited on page 27.]

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 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*, pages 3512–3518, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.322. URL https://www.aclweb.org/anthology/2020.acl-main.322. [Cited on page 15.]

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rR03qFesqk. [Cited on page 24.]

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. P-transformer: Towards better document-to-document neural machine translation, 2022. [Cited on pages 7, 16, 17, and 18.]

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory, 2023a. [Cited on page 25.]

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 02 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00638. URL https://doi.org/10.1162/tacl_a_00638. [Cited on pages 22 and 23.]

Peter J. Liu*, Mohammad Saleh*, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-. [Cited on page 12.]

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 11 2020a. ISSN 2307-387X. doi: 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343. [Cited on pages 14 and 17.]

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020b. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47. [Cited on page 16.]

Zihan Liu, Zewei Sun, Shanbo Cheng, Shujian Huang, and Mingxuan Wang. Only 5% attention is all you need: Efficient long-range document-level neural machine translation. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *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, November 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.47. URL https://aclanthology.org/2023.ijcnlp-main.47. [Cited on page 14.]

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12 2014. doi: 10.5565/rev/tradumatica.77. [Cited on page 20.]

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level neural MT: A systematic comparison. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the*

22nd Annual Conference of the European Association for Machine Translation, pages 225–234, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.24. [Cited on pages 14 and 16.]

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. Chain-of-dictionary prompting elicits translation in large language models, 2023. [Cited on page 19.]

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 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, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.312. URL https://aclanthology.org/2022.acl-long.312. [Cited on pages 6, 16, and 27.]

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 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), December 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.77. [Cited on page 17.]

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 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, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.insights-1.4. [Cited on page 18.]

Shuming Ma, Dongdong Zhang, and Ming Zhou. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.321. URL https://aclanthology.org/2020.acl-main.321. [Cited on page 14.]

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qNLe3iq2El. [Cited on page 13.]

Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1118. URL https://aclanthology.org/P18-1118. [Cited on page 15.]

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective attention for context-aware neural machine translation. In Jill Burstein, Christy Doran, and Thamar

Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1313. URL https://aclanthology.org/N19-1313. [Cited on pages 5 and 15.]

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2), March 2021. ISSN 0360-0300. doi: 10.1145/3441691. URL https://doi.org/10.1145/3441691. [Cited on page 13.]

Thomas Meyer and Bonnie Webber. Implicitation of discourse connectives in (machine) translation. In Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors, *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-3303. [Cited on page 5.]

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://www.aclweb.org/anthology/D18-1325. [Cited on page 15.]

Tsvetomila Mihaylova and André F. T. Martins. Scheduled sampling for transformers. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 351–356, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2049. URL https://aclanthology.org/P19-2049. [Cited on page 10.]

Wafaa Mohammed and Vlad Niculae. On measuring context utilization in document-level MT systems. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.113. [Cited on page 7.]

Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023. URL https://openreview.net/forum?id=PkoGERXS1B. [Cited on page 25.]

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. Adaptive machine translation with large language models. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini,

Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.22. [Cited on page 19.]

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium, October 2018. doi: 10.18653/v1/W18-6307. URL https://www.aclweb.org/anthology/W18-6307. [Cited on page 7.]

Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9452–9462, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.826. [Cited on page 26.]

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. [Cited on page 25.]

Masato Neishi and Naoki Yoshinaga. On the relation between position information and sentence length in neural machine translation. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 328–338, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1031. URL https://aclanthology.org/K19-1031. [Cited on page 7.]

Laura Nguyen, Thomas Scialom, Jacopo Staiano, and Benjamin Piwowarski. Skimattention: Learning to focus via document layout. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2413–2427, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.207. URL https://aclanthology.org/2021.findings-emnlp.207. [Cited on page 25.]

Laura Nguyen, Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. LoRaLay: A multilingual and multimodal dataset for long range and layout-aware summarization. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 636–651, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.46. URL https://aclanthology.org/2023.eacl-main.46. [Cited on page 25.]

Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/2d1b2a5ff364606ff041650887723470-Paper.pdf. [Cited on page 12.]

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560. [Cited on page 25.]

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. Salute the classic: Revisiting challenges of machine translation in the age of large language models, 2024. [Cited on page 19.]

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040. [Cited on page 7.]

Saurav Pawar, S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Aman Chadha, and Amitava Das. The what, why, and how of context length extension techniques in large language models – a detailed survey, 2024. URL https://arxiv.org/abs/2401.07872. [Cited on page 25.]

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936. [Cited on page 25.]

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023b. URL https://arxiv.org/abs/2309.00071. [Cited on page 24.]

Ziqian Peng, Rachel Bawden, and François Yvon. À propos des difficultés de traduire automatiquement de longs documents (about the difficulty of automatically translating long documents). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 35e édition), Traitement Automatique des Langues Naturelles (TALN,*

31e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 26e édition)., pages 122–135, Toulouse, France, 6 2020. ATALA et AFCP. URL https://aclanthology.org/2024.jeptalnrecital-taln.XX. [Cited on page 18.]

Jason Phang, Yao Zhao, and Peter Liu. Investigating efficiently extending transformers for long input summarization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3946–3961, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.240. URL https://aclanthology.org/2023.emnlp-main.240. [Cited on page 25.]

Hugo Pitorro, Pavlo Vasylenko, Marcos Treviso, and André F. T. Martins. How effective are state space models for machine translation?, 2024. URL https://arxiv.org/abs/2407.05489. [Cited on page 27.]

Andrei Popescu-Belis. Context in neural machine translation: A review of models and evaluations, 2019. [Cited on pages 5 and 7.]

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL https://arxiv.org/abs/2108.12409. [Cited on page 23.]

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. Large language models meet nlp: A survey, 2024. [Cited on page 19.]

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico, 2016. URL https://arxiv.org/pdf/1511.06732.pdf. [Cited on page 10.]

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213. [Cited on page 7.]

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.284. [Cited on page 25.]

Ohad Rubin and Jonathan Berant. Long-range language modeling with self-retrieval, 2023a. [Cited on page 25.]

Ohad Rubin and Jonathan Berant. Long-range language modeling with self-retrieval, 2023b. URL https://arxiv.org/abs/2306.13421. [Cited on page 27.]

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.161. URL https://aclanthology.org/2023.acl-short.161. [Cited on page 24.]

Kuniaki Saito, Kihyuk Sohn, Chen-Yu Lee, and Yoshitaka Ushiku. Where is the answer? investigating positional bias in language model knowledge extraction, 2024. URL https://arxiv.org/abs/2402.12170. [Cited on page 22.]

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331. [Cited on pages 6, 10, and 18.]

Dario Stojanovski and Alexander Fraser. Coreference and coherence in neural machine translation: A study using oracle experiments. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6306. URL https://aclanthology.org/W18-6306. [Cited on page 15.]

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864. [Cited on page 23.]

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14590–14604, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.816. URL https://aclanthology.org/2023.acl-long.816. [Cited on page 25.]

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 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. [Cited on pages 8, 16, 26, and 27.]

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.304. URL https://aclanthology.org/2021.findings-acl.304. [Cited on page 16.]

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k. [Cited on pages 6, 13, and 25.]

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):1–28, July 2023a. ISSN 0360-0300, 1557-7341. doi: 10.1145/3530811. URL https://dl.acm.org/doi/10.1145/3530811. [Cited on page 11.]

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):1–28, July 2023b. ISSN 0360-0300, 1557-7341. doi: 10.1145/3530811. URL https://dl.acm.org/doi/10.1145/3530811. [Cited on pages 6 and 9.]

Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors, *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4811. URL https://aclanthology.org/W17-4811. [Cited on page 14.]

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. Llama 2: Open foundation and fine-tuned chat models, 2023. [Cited on page 24.]

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 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. [Cited on page 15.]

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=s1FjXzJ0jy. [Cited on page 25.]

Ali Vardasbi, Telmo Pessoa Pires, Robin Schmidt, and Stephan Peitz. State spaces aren't enough: Machine translation needs attention. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 205–216, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.20. [Cited on page 25.]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. [Cited on pages 5, 10, and 23.]

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.6. [Cited on page 7.]

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.859. URL https://aclanthology.org/2023.acl-long.859. [Cited on pages 19 and 21.]

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1117. URL https://www.aclweb.org/anthology/P18-1117. [Cited on page 15.]

Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair for neural machine translation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1081. URL https://aclanthology.org/D19-1081. [Cited on page 16.]

Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116. URL https://aclanthology.org/P19-1116. [Cited on pages 5 and 20.]

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1036. URL https://aclanthology.org/2023.emnlp-main.1036. [Cited on pages 9, 19, and 20.]

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. URL https://arxiv.org/abs/2006.04768. [Cited on page 11.]

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=BryMFPQ4L6. [Cited on page 25.]

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models, 2024. [Cited on page 25.]

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, and Colin Raffel et al. Bloom: A 176b-parameter open-access multilingual language model, 2023. [Cited on page 24.]

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. Document flattening: Beyond concatenating context for document-level neural machine translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.33. URL https://aclanthology.org/2023.eacl-main.33. [Cited on page 14.]

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. Adapting large language models for document-level machine translation, 2024a. [Cited on pages 19 and 21.]

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. Importance-aware data augmentation for document-level neural machine translation. In Yvette Graham and Matthew Purver, editors, *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, March 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.44. [Cited on pages 16, 17, and 27.]

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts, 2024c. [Cited on page 22.]

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF. [Cited on page 25.]

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345, 2019. [Cited on page 16.]

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=farT6XXntP. [Cited on pages 21 and 27.]

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation, 2024b. [Cited on page 22.]

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20. ACM, August 2020. doi: 10.1145/3394486.3403172. URL http://dx.doi.org/10.1145/3394486.3403172. [Cited on page 25.]

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41. [Cited on page 17.]

Jianxin Yang. Longqlora: Efficient and effective method to extend context length of large language models, 2023. [Cited on page 25.]

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. Enhancing context modeling with a query-guided capsule network for document-level translation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1164. URL https://aclanthology.org/D19-1164. [Cited on page 15.]

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F.T. Martins, and Graham Neubig. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.65. URL https://aclanthology.org/2021.acl-long.65. [Cited on pages 7, 15, and 25.]

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 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. [Cited on page 16.]

François Yvon and Sadaf Abdul Rauf. Utilisation de ressources lexicales et terminologiques en traduction neuronale. Research Report 2020-001, LIMSI-CNRS, July 2020. URL https://hal.science/hal-02895535. [Cited on page 15.]

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for longer sequences, 2020. URL http://arxiv.org/pdf/2007.14062. [Cited on page 12.]

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the transformer translation model with document-level context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1049. URL https://aclanthology.org/D18-1049. [Cited on pages 15 and 16.]

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *Proc. International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr. [Cited on page 7.]

Xiaojun Zhang. A review of discourse-level machine translation. In Qun Liu, Deyi Xiong, Shili Ge, and Xiaojun Zhang, editors, *Proceedings of the Second International Workshop of Discourse Processing*, pages 4–12, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.iwdp-1.2. [Cited on page 5.]

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.43. URL https://aclanthology.org/2023.wmt-1.43. [Cited on pages 19, 20, and 21.]

Liang Zhao, Xiaocheng Feng, Xiachong Feng, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. Length extrapolation of transformers: A survey from the perspective of positional encoding, 2024. [Cited on page 23.]

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165. [Cited on page 14.]

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3Z1gxuAQrA. [Cited on page 24.]

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. Addressing the length bias challenge in document-level neural machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.773. URL https://aclanthology.org/2023.findings-emnlp.773. [Cited on pages 18 and 27.]