Natural Language Processing (2025), **31**, pp. 986–1016
doi:10.1017/nlp.2024.7

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

A survey of context in neural machine translation and its evaluation

Sheila Castilho[1]* and Rebecca Knowles[2]*

[1]School of Applied Language & Intercultural Studies/ADAPT Center, Dublin City University, Dublin, Ireland and [2]National Research Council of Canada, Ottawa, ON, Canada
Corresponding author: Sheila Castilho; Email: sheila.castilho@adaptcentre.ie

(Received 29 February 2024; accepted 29 February 2024; first published online 17 May 2024)

Special Issue on '**The Role of Context in Neural Machine Translation Systems and its Evaluation**', guest-edited by Sheila Castilho and Rebecca Knowles

Abstract

The question of context in neural machine translation often focuses on topics related to document-level translation or intersentential context. However, there is a wide range of other aspects that can be considered under the umbrella of context. In this work, we survey ways that researchers have incorporated context into neural machine translation systems and the evaluation thereof. This includes building translation systems that operate at the paragraph level or the document level or ones that translate at the sentence level but incorporate information from other sentences. We also consider how issues like terminology consistency, anaphora, and world knowledge or external information can be considered as types of context relevant to the task of machine translation and its evaluation. Closely tied to these topics is the question of how to best evaluate machine translation output in a way that is sensitive to the contexts in which it appears. To this end, we discuss work on incorporating context into both human and automatic evaluations of machine translation quality. Furthermore, we also discuss recent experiments in the field as they relate to the use of large language models in translation and evaluation. We conclude with a view of the future of machine translation, where we expect to see issues of context continue to come to the forefront.

Keywords: machine translation; evaluation

1. Introduction

The rise of neural machine translation (NMT) in recent years, marked by the advent of sophisticated neural models—such as those introduced by Sutskever et al. (2014), Bahdanau et al. (2015) and Vaswani et al. (2017)—has significantly propelled the field of machine translation (MT), leading to substantial enhancements in translation quality (Castilho et al., 2017). Throughout this period of change, much of the research in the field has remained focused on the sentence-level translation paradigm: building MT systems that take as input a single sentence (or other short segment, such as a title or phrase) and produce as output a translation in the target language. Nonetheless, the intricate role of context in shaping translation outcomes remains a critical and evolving area of exploration.

The importance of context has been highlighted by both the claims of human parity in NMT (Hassan et al., 2018, i.a.) as well as by the analyses performed in response, which have shown

*Both authors contributed equally to this survey.

that this appearance of parity may disappear when evaluating NMT output with context (Toral et al. 2018; Läubli et al., 2018, i.a.). There is a strong and important link between context in NMT and context in NMT evaluation. When we perform evaluation of NMT without considering context, we may fail to consider features like consistency or discourse that are important to human end-users of translation. This runs the risk of undervaluing approaches to NMT that produce improvements in these context-sensitive areas. Improving our ability to evaluate the level of whole documents or incorporating other aspects of context will allow us to measure and drive improvements in NMT. A careful evaluation of the methodologies required for evaluating context-aware systems is crucial as evaluation facilitates the identification of gains and shortcomings in these systems.

But what *is* "context"? Some works refer to document-level translation and evaluation, while others discuss discourse phenomena, and yet others refer to context-aware translation. We will use all of this terminology in this survey, noting that there is often overlap in the use of each of these terms. Most commonly, studies of context in NMT have focused on intersentential context, some span of sentences that exist within some document around the sentence being translated. This will be one of the main types of context we consider in this survey, but there are other components of context as more broadly defined that we will also consider. Melby and Foster (2010) provide a typology of contexts, including just not these textual contexts like intersentential context, but also related documents or resources as well as real-world information. Throughout this survey, we will occasionally refer back to the typology of Melby and Foster (2010) as a conceptual starting point, though we may use the definitions loosely or discuss types of data that may overlap with multiple context types. These broader aspects of "context" could include the real-world context in which translation is occurring, the intended audience, the level of formality required, the incorporation of terminology, and much more. In this survey, we will examine the ways that these approaches have been explored in the literature, drawing connections between some areas that are often viewed as discrete tasks, separate from the question of context, but which can still be brought together under this broad umbrella.

This survey aims to look into the multifaceted relationship between context and MT (Section 2), describing prior surveys of document-level MT (Section 2.1), dissecting various dimensions such as intersentential context (Section 2.2), world knowledge and external information (Section 2.3), and the treatment of terminology (Section 2.4). Beyond translation, the survey also investigates the pivotal role of context in the evaluation of machine-generated translations (Section 3), both from a human perspective (Section 3.1) and through automated metrics (Section 3.2). We also discuss recent interest in how large language models can play a role in the study of context for MT and its evaluation (Section 4). We limit the scope of our discussion primarily to the topic of text-to-text translation, with very brief discussion of other types of translation. As we navigate through the complexities of context in text-to-text MT, this survey aims to provide a comprehensive overview, shedding light on current insights and paving the way for future advancements in the field. We end with a summary and discussion of future directions for study (Section 5).

2. Context and machine translation

Despite the advancements in NMT, integrating context remains a critical and underexplored avenue for enhancing the accuracy and fluency of machine-generated translations. While contemporary MT predominantly operates within a sentence-level paradigm (Wicks and Post, 2022), there is a growing emphasis on overcoming the challenge of producing coherent document-level MT translations that make use of a broader context. Consequently, recent efforts by researchers involve incorporating discourse into NMT systems (Wang, 2019; Lopes et al. 2020, i.a.).

While the term "document level" has been used somewhat loosely in reference to MT systems that handle context beyond the sentence level (i.e., intersentential context), the precise definition

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

of what constitutes a document-level context remains a subject of debate (Castilho, 2022). Much of the recent work on document-level or context-aware NMT has focused on systems that primarily rely on a context span limited to sentence pairs immediately surrounding the sentence to be translated (Tiedemann and Scherrer, 2017; Bawden et al. 2018; Müller et al., 2018), or within document substructures (Dobreva, Zhou, and Bawden, 2020), with only a few venturing beyond this span (Junczys-Dowmunt, 2019; Voita, Sennrich, and Titov, 2019b; Lopes et al., 2020). Additionally, it is still unclear whether the document-level NMT models "rely on the 'right' context that is actually sufficient to disambiguate difficult translations" (Yin et al., 2021, p.788) or whether there need to be additional model improvements to ensure that they successfully access the correct context necessary.

However, performing translation at the document level (or, simply beyond the level of the single sentence) is not the only way to view the question of context in NMT. Melby and Foster (2010) describe five different aspects of context as they relate to translation, calling them *co-text* (surrounding text within a document; e.g., intersentential context), *chron-text* (versions of the text over time), *rel-text* (related documents and resources), *bi-text* (e.g., translations of text, translation memories, etc.), and *non-text* (additional real-world information used for translation). We will examine various areas where researchers have explored one or more of these types of context in their work on MT and, in later sections, in the evaluation of MT.

In this section, we delve into various dimensions of context in MT, including prior surveys on the topic, intersentential context, world knowledge, external information, and terminology.

2.1 Prior surveys of document-level NMT

Several papers have registered the advances in the document-level field of MT, surveying, discussing, and reporting different methodologies and model architectures, with some simultaneously introducing new approaches to the problems. In 2019, Popescu-Belis (2019) published a review paper describing work from 2017 to 2018 that made use of contextual information in NMT. The author gives a comprehensive overview of the rapid evolution of the NMT models and the first attempts to add context to those systems, dating back from statistical MT systems, to newer NMT models with context information. In the same year, (Kim, Tran, and Ney, 2019, p. 24) experiment with evaluating a document-level model with test sets in order to quantify "when and why document-level context improves NMT," providing insights into how and when context can be useful in NMT. The authors considered "document-level" to be one previous source sentence, and filtered words in the context, retaining only the "context words that are likely to be useful" (Kim, Tran, and Ney, 2019, p. 26). Their findings demonstrated that, at that time, (i) most of the improvements lacked clear interpretation regarding context utilisation, (ii) minimal encoding was sufficient for the context modelling, (iii) and very long context did not offer substantial benefits for NMT.

Lopes et al. (2020) provided a comparison of existing and new (at the time) document-level NMT solutions with both automatic and manual evaluations. The approaches they compared were concatenation (previous source and/or target sentence concatenated with the current), multi-source (previous two source sentences) and cache-based (all previous source and/or target sentences). For comparison, the authors also implemented one model based on the Star Transformer architecture (Guo et al., 2019), newly proposed at that time, using the cache-based approach. They found that, at that stage, strong sentence-level baselines were still outperforming existing context-aware approaches in scenarios with larger datasets.

The survey by Maruf, Saleh, and Haffari (2021) is a comprehensive analysis of the major work in the domain of document-level MT from the introduction of NMT to the date of publication. The survey encompasses research conducted within the frameworks of both RNN and Transformer, offering a comprehensive taxonomy, with different types of context span used. They also provide a review of the evaluation strategies for document-level NMT.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Finally, Jin et al. (2023) present a review of literature pointing out some key obstacles relating to discourse phenomena, context usage, model architectures, and evaluation that hinder progress in the domain. These include: (i) document-level corpora with sparse number of discourse phenomena, (ii) context being less helpful for tense and discourse markers, (iii) concatenation-based NMT systems not outperforming context-agnostic sentence-level Transformer baselines, (iv) meaningful improvement not being achieved by advanced model architectures, and (v) current evaluation metrics not adequately measuring quality. The authors propose a novel setting for document-level translation with a dataset of aligned paragraphs.

In summary: As can be seen, extensive work has been done in the past decade to understand how context influences the translation process, since it is essential for enhancing the quality and fluency of MT systems. In the rest of this survey, we will briefly discuss earlier work to provide background and then attempt to augment these existing surveys with additional recent work.

2.2 Intersentential context

We now consider the question of context as interpreted to mean intersentential context. Performing MT at the sentence level makes an implicit assumption that sentences are independent—that is, that the translation of one sentence does not depend on additional information outside of the sentence itself. This is of course trivially untrue; even a sentence as simple as "I see the bank." contains a polysemous word (bank) whose translation into another language may depend on whether this particular sentence is intended to refer to a financial institution or the side of a river, for example. Nevertheless, as evidenced by the success of the single-sentence translation paradigm, many sentences *can* be translated well enough in isolation. For those sentences that do require additional context to translate, MT research often focuses on intersentential context (part of what Melby and Foster (2010) would call "co-text"): context that crosses sentential boundaries but that exists within some sort of document boundary.[a]

Incorporating intersentential context or document context has the potential to provide discourse information such as disambiguation of pronouns, deixis, ellipsis, as well as general cohesion. For example, there may be cases of anaphoric pronouns, where information about the person or thing mentioned earlier in the text can be used to disambiguate the translation of a pronoun in a later sentence. Other portions of the text may similarly be able to fill in the necessary blanks to resolve questions related to deixis or ellipsis. As for cohesion, this could be a matter of improving consistency of terminology (see also Section 2.4) or consistency in senses for polysemous terms, since a single discourse tends to make use of just one sense per polysemous word (Gale, Church, and Yarowsky, 1992). Carpuat (2009) found that the concept of one sense per discourse also held in the case of translation, showing for phrase-based statistical MT that improving sense consistency (e.g., by incorporating document context) had the potential to improve translation quality, so we have reason to expect the same to be true in the case of NMT.

This incorporation of intersentential context can take several forms, including translation at the sentence level that incorporates information beyond the sentence being translated as well as performing translation of units larger than the sentence. Maruf et al. (2021) provide a thorough survey of approaches to handling intersentential context in NMT, so in this section we will provide only a brief overview of common approaches.

Perhaps the most intuitive of these approaches, albeit one that comes with computer-related challenges, is to perform translation of entire paragraphs or documents at one time. That is, rather than translating one sentence at a time, having the unit of translation be the paragraph or the document (i.e., document-to-document translation). This comes with several challenges from both the data side and the implementation side. Nevertheless, Junczys-Dowmunt (2019) observed

[a]Looking at a much broader intersentential context; i.e., well beyond the document level, *k*-nearest neighbour MT systems (Khandelwal et al., 2021) could arguably be viewed as a context-aware approach to MT.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

strong performance gains over sentence-level baselines by extending the length of the input to 1000 subword tokens (i.e., using paragraphs or documents as the unit of translation rather than individual sentences; Popel et al. (2019) took a similar approach, by lengthening the input to 1,000 characters or approximately 15 sentences), adding special characters to indicate sentence and document boundaries, backtranslation, and creating synthetic documents for training from parallel data that did not contain document boundary information. This last piece highlights one of the challenges of document-level MT: the fact that many large sources of training data, having been prepared for use in sentence-level translation, do not contain document boundary information. In re-examining document-to-document translation, Sun et al. (2022) found that multi-resolution training, wherein they incorporate both full documents and shorter segments, can improve performance. Al Ghussin et al. (2023) explore the possibility of automatically extracting paragraph-level data from Paracrawl, for use in training document-level MT.

Another common approach is to perform sentence-level translation, but incorporate intersentential context. This often consists of providing the system with information about several nearby sentences, from the source side, the target side, or both sides. While it is most common to use preceding intersentential context, some work has also explored the use of future context, particularly for the resolution of cataphoric pronouns (Wong, Maruf, and Haffari, 2020). Tiedemann et al. (2017) examined both extending the unit of translation to two sentences (similar to the approaches in the previous paragraph) as well as providing the preceding source sentence concatenated with the sentence of interest as the input to the model and training the model to produce a translation of the source sentence of interest only. In this way, they sidestepped the issue of modifying the translation model by modifying the data used for training and the processing of the data for decoding. Similarly, Rippeth et al. (2023) seek to improve word sense disambiguation by prefixing input data with salient keywords from the surrounding context. Jean and Cho (2019) also avoided modifying the model architecture by proposing a novel regularisation term for the learning algorithm to encourage systems to focus on useful context. A number of architecture modifications have been proposed to incorporate intersentential context, such as additional encoders and attention (Jean et al., 2017), memory networks that incorporate both source and target intersentential context (Maruf and Haffari, 2018), multi-head hierarchical attention networks (Miculicich et al., 2018), hierarchical sparse attention (Maruf, Martins, and Haffari, 2019), query-guided capsule networks (Yang et al., 2019), summarising cross-sentence context with a hierarchy of recurrent neural networks and then using this for initialisation or incorporating it via gating (Wang et al., 2017), and incorporating context information into Transformer models using context encoders that share parameters with the source encoder (Voita et al., 2018).

Huo et al. (2020) compare four different context-aware architectures, while Gete et al. (2023) compare two concatenation-based approaches on the task of pronoun translation. These represent just a taste of the many approaches that have been employed and analysed (see also, Agrawal, Turchi, and Negri, 2018; Kim et al., 2019; Macé and Servan 2019; Li et al., 2020; Kang et al., 2020; Saunders, Stahlberg, and Byrne, 2020b; Yu et al., 2020; Mansimov, Melis, and Yu, 2021, i.a.). Returning to the issue of training data that may not always have document boundaries available, Zheng et al. (2021) seek to balance local and global context, with an architecture that is flexible enough to allow for documents of any length (including single sentences), enabling the models to train on and translate either single isolated sentences or sentences in document context.

Herold and Ney (2023) address the issues of memory usage and translation performance when using large context by proposing a constrained attention approach, which reduces memory consumption while focusing on relevant portions of the context. They also consider the issue of evaluation, as we will discuss in Section 3: document-level models often show small improvements in terms of standard automatic metrics, but may exhibit performance boosts on specific tasks or targeted test sets, particularly those closely related to intersentential context.

In summary: Much of the work on MT and MT evaluation to date has focused on the sentence level. This means that the current tools for evaluation are predominantly designed to capture

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

facets of quality that can be measured at the sentence level, and may not be able to capture improvements that are related to intersentential context. We discuss these issues of evaluation more in Sections 3.2 and 3.2. Nevertheless, as has been made clear by analyses like Toral et al. (2018) and Läubli et al. (2018), there remains room for improvement over the current sentence-level paradigm of translation, some of which will likely be found through the incorporation of intersentential context. It remains to be seen what the best ways of incorporating such context are, that is whether it is sufficient to simply perform translation on units the size of whole documents, in what ways training or model architectures should be modified to incorporate more context, or whether other approaches will prevail. It is clear, however, that translation at the level of the single sentence, without intersentential context, differs at least some of the time in notable ways from optimal human references, including when it comes to issues of discourse and cohesion. We expect the best systems in the future—especially when those systems are ranked or compared by human annotators or automatic metrics in ways that take intersentential context into consideration—to incorporate intersentential context in some manner, as human translators do in their work.

2.3 World knowledge and external information

In the realm of the type of context that might be called *non-text*, as described in Melby and Foster (2010), we can consider ways in which real-world context is relevant to translation. This can include knowledge about the world, as might be found in a knowledge base, as well as knowledge about the goals of the translation task itself. While some of this information may be captured at the document level, at other times it may need to be provided in some other way to the system.

Regarding the latter, this can include how to handle desired levels of formality. Sennrich et al. (2016a) proposed the use of "side constraints" or special tokens inserted into the source sentence in order to control the formality of MT output, in their case specifically in terms of the formal/informal "you" (T-V) distinction. Similarly, Feely et al. (2019) used source-side tags to control the generation of honorifics in English to Japanese translation. Niu et al. (2018) generalised beyond pronouns and honorifics to a more general sense of (in)formality in formality-sensitive MT, using similar side constraints as well as a multi-task approach that incorporated formality transfer. Niu and Carpuat (2020) subsequently proposed an approach involving on-the-fly synthetic example generation during training to improve formality translation.

This approach to tagging input sentences with special tokens has also been applied to controlling the output domain in multi-domain NMT systems (Kobus, Crego, and Senellart, 2017). Handling multiple domains in one MT system involves either providing the NMT system with information about the desired output domain or training it to learn to handle different input domains, such as by predicting their domain from the input sentence. Beyond the use of special tokens as side constraints, approaches to handling multiple domains and training systems to perform well in multi-domain settings include jointly learning to translate and discriminate between domains (Britz, Le, and Pryzant, 2017), treating domain as a feature for factored NMT (Tars and Fishel, 2018), instance-based on-the-fly domain adaptation (Farajian et al., 2017; Li, Zhang, and Zong, 2018; Xu, Crego, and Senellart, 2019), and domain-specific multi-head attentions (Jiang et al., 2020). Pham et al. (2021) propose a number of properties that should hold for high-quality multi-domain NMT systems, examine these various approaches, and conclude that there is still work to do on smoothly handling multi-domain NMT. Domain can be interpreted quite broadly, and the successful approaches to issues of domain and formality can also be applied to a range of other related and overlapping topics. This includes jointly translating and simplifying text (Agrawal and Carpuat, 2019) or controlling the reading level at which MT output is generated (Marchisio et al., 2019). The concepts of domain and (intersentential) context have also been directly brought together in Stojanovski and Fraser (2021), who note that providing additional intersentential context can help improve performance on new domains and can improve coherence and consistency in translation.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Most of these describe multi-domain NMT systems. However, NMT systems can also be built or adapted to specifically handle one particular domain. We can consider the concept of providing domain information to a system as a type of non-text or world knowledge that is being incorporated into the MT system. Chu et al. (2017) compared several approaches to domain adaptation and Chu and Wang (2018) surveyed approaches to domain adaptation for NMT. Even more recently, Saunders (2022) surveyed both domain adaptation as well as multi-domain approaches in NMT.

While we focus primarily on topics related to "standard" text-to-text translation, we should also note that various other settings for translation may require specialised approaches; we limit our discussion to noting a few of them. In these cases, we are considering context in the sense of the use case for translation: the setting for which the particular system is designed, which may inform choices about the model or about desired qualities of the MT output. Translation for dubbing requires producing translations of similar lengths to the source (Lakew et al., 2022), while translation for subtitling may require handling additional formatting issues (Cherry et al., 2021), and simultaneous MT requires approaches to better handle potential variations in word order (Grissom II et al., 2014) or incremental decoding approaches (Gu et al., 2017; Dalvi et al., 2018). Similarly, NMT systems for use in computer-aided translation settings such as interactive translation prediction may use modified decoding or training approaches (Knowles and Koehn, 2016; Wuebker et al., 2016; Li et al., 2021). The external information provided to the MT system need not take the form of text. Saleh et al. (2019) draw connections between natural language generation from structured data and document-level translation, and both Vincent et al. (2022) and Vincent et al. (2023) seek to use metadata to inform translation. Another such example of external information that does not take the form of text is in multimodal translation, when an image and text are used as input for translation (Elliott, Frank, and Hasler, 2015; Hitschler, Schamoni, and Riezler, 2016; Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018, i.a.), or when both text and speech or some other signal are used. Translation from text or speech into sign languages also requires consideration of how to produce a signing avatar (McDonald et al., 2021), raises ethical questions around language data that contains video of individual signers (Bragg et al., 2021), and highlights the importance of language community involvement in technology development (De Meulder, 2021), all aspects of a much broader world of context surrounding the task of translation.

In addition to the *non-text* sources of real-world context, there are other text-based sources that can incorporate world knowledge and external information into NMT systems. Linguistic knowledge, typically in the form of text, but potentially in other forms, may also be a source of knowledge to be incorporated into NMT systems. Particularly in low-resource settings, incorporating forms of linguistic knowledge or additional examples through data augmentation may be beneficial (Fadaee, Bisazza, and Monz, 2017); we discuss the specific case of external lexicons in the following section. Linguistic knowledge that has already been made machine-readable through its inclusion in rule-based MT systems can be used in NMT systems (Torregrosa et al., 2019). Incorporating monolingual data is also a way of providing additional linguistic information (i.e., improving language model capabilities) as well as a way of potentially providing the system with world knowledge (e.g., names of famous figures, news and events, etc.). A common way of incorporating monolingual data is by converting it into pseudo-parallel text through the process of backtranslation (Sennrich, Haddow, and Birch, 2016b), though other approaches have also been proposed. These include performing monolingual target language automatic post-editing at the document level to "repair" translations that were produced at the sentence level and improve their consistency (Voita, Sennrich, and Titov, 2019a), using target-side monolingual data for a particular domain to train a document-level language model and incorporating this into decoding for a sentence-level NMT model (Sugiyama and Yoshinaga, 2021), and using multilingual modelling to transfer document-level translation information to language pairs without large corpora of document-level data (Zhang et al., 2022). For more discussion of monolingual data in NMT, see Burlot and Yvon (2018).

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

In summary: The external context in which translation occurs, as well as the intended audience of translation, are important factors in determining whether a given translation is appropriate. To this end, approaches to handling formality, domain, and specific use cases have been proposed. Additionally, particularly when the amount of available parallel text is limited, monolingual data can serve as useful training input for learning both linguistic features and incorporating world knowledge into NMT systems.

2.4 Terminology

Another source of external information for MT is the use of terminology. This can be viewed through the lens of context as what Melby and Foster (2010) would call *rel-text*, a text source of information outside of the particular document being translated. There is not necessarily a clean distinction between terminology and the types of world knowledge and external information that we discussed in Section 2.3; we discuss it in a separate section here due to both the large quantity of research in this area as well as the fact that the field has often treated it as a separate specific task. While the incorporation of terminology—in the form of source-target pairs of words or phrases—was relatively straightforward in the phrase-based statistical MT paradigm, it has been a topic of interest in NMT research, even resulting in a shared task at the Conference on Machine Translation in 2021 (Alam et al., 2021). This topic and the various approaches to it all share two main connections to the topic of context and NMT. The first is that this represents a form of introducing additional context (rel-text) from the lexicon or terminology resource into the process of MT. The second is that this can contribute to within-document or even across-document consistency in translation, potentially improving document coherence. This second connection is one that ties closely to (human) evaluation of MT in context, which we discuss in Section 3.1.

There exist a number of motivations for wanting to incorporate terminological resources into MT, ranging from specific domains with highly technical vocabularies to client-specific preferred terms to low-resource settings where dictionary-like text may form a major component of the available bitext. Starting with the early days of NMT that used restricted vocabularies, one area of interest has been improving the translation of rare or out-of-vocabulary words using lexical resources (Jean et al., 2015; Luong et al., 2015, i.a.). Another reason for incorporating terminology into NMT is to improve consistency, for example of a particular term that appears repeatedly in a document or broadly across a client's data. The types of terminological or lexical resources may vary based on the scenario, including but not limited to bilingual dictionary entries, term banks, or phrase tables extracted from corpora. Some are designed specifically with MT in mind as a use case, while others may need to be adapted to this (e.g., terminological resources designed for use by human translators). They may also vary in whether they contain only citation forms of words or fully inflected forms (and, in the latter case, whether they contain the full possible range of inflections or only a subset).

Yvon and Rauf (2020) provide a thorough overview of approaches to incorporate bilingual lexicons and other terminological resources into NMT. Some approaches require modifications at training time, while others only make modifications at inference time. The appropriateness of a given approach will depend on the use case and the available terminological resources. For example, if the terminology is expected to change over time, one may wish to select an approach that does not require a fixed terminology at training time. If the terminology only contains citation forms (e.g., roots or lemmas), care may need to be taken to ensure that the approach is flexible enough to support producing or recognising inflected surface forms.

Of the approaches that occur at training time, these can roughly be broken into two categories: those that rely on the fixed terminology (known in advance) for training and those that train for a specific behaviour but can incorporate flexible or changing terminological resources. A very simple approach is to treat a bilingual lexicon as simply more parallel text to use as training data.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Kothur et al. (2018) fine-tune NMT systems using bilingual lexicons of novel vocabulary in order to improve translation consistency within individual documents for these novel terms, but warn of the risks of overfitting to such small sets of training data. Given a lexicon that is fixed in advance, another way of incorporating it is to combine translation probabilities from a lexicon with the NMT model's probabilities (Arthur, Neubig, and Nakamura, 2016) or to train a model to directly generate a target word based on a source word and combine this with the model's scores (Nguyen and Chiang, 2018). These models would require modification or retraining in order to incorporate new entries in a terminology set.

There are approaches at training time that do allow for later changes to the terminology set, without retraining. These include copying mechanisms and placeholders. Post et al. (2019) train models to incorporate special placeholders, which can then be replaced in postprocessing. Dinu et al. (2019) use factors to train models to use custom terminology, relatively flexibly. Bergmanis and Pinnis (2021b) expand on this, focusing on improving performance on morphological inflections. Song et al. (2019) perform data augmentation in which the input is converted into "code-switched" data that includes desired target terminology, with the goal of encouraging the system to copy this terminology into the output. Ailem et al. (2021) combine modification to the training data (in the form of tags to indicate when terminology should be used), token masking, and a weighted cross-entropy loss in order to incorporate terminology.

For cases where the use of terms from the lexicon is strictly required, these training-time approaches may not offer strong enough guarantees that the terms will be included in the output. Approaches like lexically constrained decoding, in which the beam search algorithm is modified to force the output to include certain tokens, offer stronger guarantees (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019) at the cost of flexibility and occasional degradation of output quality. Hasler et al. (2018) present an approach to constrained decoding using finite-state machines, with improvements to the placement of the constraints in the resulting output. Susanto et al. (2020) propose an approach to lexically constrained decoding using Levenshtein transformers (Gu, Wang, and Zhao, 2019). Inspired by the XML-style constraints in phrase-based statistical MT, Chatterjee et al. (2017) propose guided decoding to enforce specified translations of text spans.

These training-time and inference-time approaches offer various tradeoffs; Exel et al. (2020) examine some of these by comparing an approach from Dinu et al. (2019) to lexically constrained decoding. Bane et al. (2023) compare three different approaches to incorporating terminology, but find that there is no clear winner; different approaches perform differently in different settings. Bergmanis and Pinnis (2021a) highlight the importance of having high-quality terminology sources, noting that it is not sufficient that they be machine-readable term banks; the term banks used by professional translators often incorporate ambiguity that the skilled translator is expected to be able to handle, but either approaches to incorporating terminology need to improve to also handle this, term banks need to be produced specifically to work well for MT, or both of these need to be combined. Examining the same shared task as the previous work, Ballier et al. (2021) also performed analysis of the provided term banks and showed a number of problematic areas. Focusing on one specific case of parliamentary text, Knowles et al. (2023) discussed tradeoffs around the question of incorporating terminology.

In summary: While often considered as a separate topic, incorporating external terminology resources can be viewed as a problem of how to best incorporate context into NMT. There exist a number of effective approaches to incorporating lexicons and other terminological resources, though these come with various tradeoffs. Strong guarantees that exact phrases will appear in the output can lead to degraded translation quality or can lead to phrases being used in contexts where they are not actually appropriate. On the other hand, soft guarantees may still leave an undesired level of variation. Future work in this area will likely seek to balance these, as well as consider what types of terminology resources to use, or how to make better use of terminology resources that were initially designed for human translator use.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

3. Context and evaluation of machine translation

Despite the advances in context-aware MT, evaluating the efficacy of such systems remains a challenge. The best practices for human evaluation of context-aware MT are still being developed and evaluated, which in turn means that the gold standard for evaluating context-aware or document-level automatic metrics is also in flux. In the meantime, it is common to use sentence-level automatic metrics (either at the sentence level or applied to documents or multi-segment chunks) to perform evaluation. Automatic metrics designed for sentence-level translations may not precisely capture the quality of translations when assessed at the document level (Smith, 2017; Lopes et al., 2020, i.a.). In some cases, this is because they are simply applied at the sentence level and as such have no access to intersentential context and document-level information. However, even when traditionally sentence-level metrics are applied to a larger span of segments or even entire documents, they may fail to fully or informatively capture the kinds of document-level phenomena that still separate human translations from high-quality sentence-level MT. We can consider the example of BLEU, as a widely used metric. An inherent limitation of BLEU—characterised by its "complete unspecificity and uniform weight assignment to any overlap with the reference" (Hardmeier, 2012)—renders it insensitive to context-aware improvements targeted by document-level MT systems. This insensitivity is apparent in scenarios where context-aware enhancements (such as the translation of pronouns) impact only a limited subset of words within a given text and where coreference or other analyses of coherence would be necessary to properly evaluate their importance. It will not generally be clear from a change in BLEU score whether the resulting improvements or degradation are due to document-level phenomena or other changes. This is not specific to BLEU alone, but is common to metrics that do not take larger-scale context and linguistic structure into account.

The perceived inadequacy of conventional metrics has prompted the MT community to contemplate the incorporation of context into evaluation methodologies. Consequently, diverse evaluation metrics encompassing automatic, semi-automatic, and human assessments have been explored to comprehensively gauge the performance of context-aware MT systems. In this section, we will describe a few methods that have been applied to the task of adding context to MT evaluation both in human and automatic methods.

3.1 Context, human evaluation, and semi-automatic evaluation

With the improvement of translation quality in recent years, discriminating MT output from human translation has proven difficult with the current human evaluation methods (Läubli et al., 2020; Kocmi et al. 2023), and it has become clear that new forms of best practices are required. Therefore, human evaluation of MT at the document level has attracted much attention as it allows for a more thorough examination of the output quality with intersentential context (Toral et al., 2018; Läubli et al., 2018). Research has been carried out in order to find out how much context needs to be shown to translators (Castilho et al. 2020), as well as different methodologies for adding context into human evaluation of MT (Castilho, 2020, Freitag et al. 2021a; Freitag et al., 2021a).

The findings from Castilho et al. (2020) revealed that over 33% of the tested sentences needed more context than just the sentence itself for effective translation or evaluation, and from those, 23% required more than two preceding sentences to be adequately assessed.[b] This is often described as the "context span" needed for disambiguation. Challenges impeding translation included ambiguity, terminology, and gender agreement. Equipped with these findings, the author conducted a series of experiments, testing different methodologies for adding context to human evaluation of MT.

[b]The experiment was performed with 300 sentences, and the domains were literature, subtitles and user reviews.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Castilho et al. (2020) looked into methodologies and inter-annotator agreement (IAA) between single-sentence and document-level evaluation setups. In the first study, translators were tasked with evaluating the MT output in terms of fluency, adequacy (using a Likert scale), ranking, and error annotation. Two evaluation setups were employed: (i) translators assigned a single score per isolated sentence, and (ii) translators assigned a single score per entire document. The results indicated that IAA scores for the document-level setup reached negative levels, and the satisfaction level of translators with this methodology was notably low. Despite this, the document-level setup successfully mitigated cases of misevaluation that occurred in the single-sentences setup (e.g., for sentences where there is an ambiguity about translation when the sentence is shown in isolation but not when intersentential context is provided).

Building on these results, Castilho (2021) modified the document-level setup and repeated the experiment with a larger group of translators. The study compared IAA in the evaluation of: (i) random isolated single sentences, (ii) individual sentences evaluated with access to the full document, and (iii) full documents. The findings demonstrated that a methodology where translators assess individual sentences within the context of a document (ii) achieved a satisfactory level of IAA compared to the random single-sentence methodology (i). Conversely, a methodology where translators assigned a single score per document (iii) showed a significantly lower level of IAA. The author illustrated that the approach of assigning one score per sentence in context effectively avoided misevaluation cases, which are prevalent in random-sentence-based evaluation setups. Moreover, she argued that the higher IAA in the random single-sentence setup might be attributed to raters accepting translations when adequacy is ambiguous but the translation is potentially correct, particularly if it is fluent. Consequently, the author recommended avoiding the single random sentence evaluation method, emphasising the heightened risk of misevaluation, particularly when assessing the quality of NMT systems due to their improved fluency level.

Still with regard to context span, Castilho et al. (2020) found that the different issues reported in the DELA corpus (Castilho et al. 2021) required differently sized context spans to be solved. Interestingly, the domains of the data also played a major role in determining the required size of the context span. For example, the subtitle domain (TED talks) required the longest context span to solve the issue of grammatical number.

Finally, when testing shorter context spans that contain the issues from the DELA corpus, Castilho et al. (2023) found that the position of the context span (before or after the source sentence) did not seem to affect the results much, although for lexical ambiguity, gender and number the most correct translations happen when the context is positioned before the sentence being translated. Interestingly, single sentences that contained the cue for the solution of the issues were not always enough for the systems to translate them correctly.

In 2019, the Conference for Machine Translation (WMT) also ventured into document-level human assessment for the news domain (Barrault et al., 2019). Employing the direct assessment (DA) methodology (Graham et al., 2016), crowdworkers were tasked with assigning scores (ranging from 0 to 100) to individual sentences. Raters were instructed to evaluate (i) randomly selected segments, (ii) consecutive segments in their original order, and (iii) entire texts. Subsequently, in the following year, WMT20 adopted a modified approach by expanding the contextual span to encompass full documents. Raters were required to evaluate specific segments while having access to the complete document and to assess the overall translation quality of the content (Barrault et al., 2020). Since 2022, the human evaluation is performed as a source-based ("bilingual") DA + SQM (Scalar Quality Metrics) of individual segments in document context. Annotators are presented with the entire translated document snippet (typically about 10 sentences) randomly selected from competing systems with additional static contexts, and asked to rate the translation of individual segments and then the entire snippet on sliding scales between 0 and 100. However, the slider is presented to the annotator with seven labelled tick marks, thus incorporating the SQM component (Kocmi et al., 2022).

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

The Metrics Shared Task, running since 2008, has compared a number of evaluation metrics submitted by the community. Until 2020, the Metrics Shared Task used the official human scores from the WMT News Translation Task to evaluate the submitted metrics. From 2021 onward (Freitag et al. 2021b), the organisers have also collected their own human evaluation for three language pairs from professional translators via Multidimensional Quality Metrics (MQM; Lommel, Uszkoreit, and Burchardt, 2014), a framework for identifying errors in translations within an error typology, with severity rankings (Freitag et al., 2023). Annotators are given access to the full document context when performing these MQM annotations.

Focusing on the broader definition of context, Licht et al. (2022) proposed XSTS, a Crosslingual Semantic Textual Similarity scoring approach for human evaluation of machine translation. They propose it in the setting of low-resource translation, where the main goal may be on ensuring that translations are adequate, with fluency potentially a secondary concern. Additionally, it ties to other questions of the context in which translation is used, such as social media text, where it may be more appropriate to score some level of faithfulness (i.e., not penalising grammatical disfluencies or linguistic variation that appears naturally in the source) rather than standard measures of fluency.

Finally, another prevalent approach to evaluating translation quality at the document level involves the use of test suites, as they gauge the model's proficiency in translating specific discourse-level phenomena including but not limited to anaphora, deixis, ellipsis, lexical ambiguity and consistency (Rios Gonzales, Mascarell, and Sennrich, 2017; Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Voita, Sennrich, and Titov, 2019b; Cai and Xiong, 2020, i.a.). Some of these also focus specifically on issues related to bias, such as the translation of gendered pronouns, a topic at the intersection of context (due to anaphora resolution, for example) and bias (Saunders, Sallis, and Byrne, 2020a). Castilho et al. (2023) used a test suite covering a number of context-related issues, including grammatical gender, to test different NMT systems and a ChatGPT[c] model. The findings indicated that all systems encountered difficulties in accurately translating grammatical gender, even when provided with contextual information. Additionally, the authors observed a consistent misidentification of the speaker's gender by the GPT model when prompted to determine the gender in sentences without explicit speaker indications, particularly when specific adjectives were employed.

While test suites fall somewhere between the categories of human evaluation and automatic evaluation, we discuss them in this section because of the large amount of human labour typically involved in creating them, though in their use, they may either involve human annotation or automatic scoring. These test suites can incorporate both accurate and inaccurate translations corresponding to particular phenomena, thereby enabling the assessment of the model's accuracy in identifying the correct translation. Nonetheless, the availability of test suites delineated explicitly at the document level remains relatively limited (Vojtěchová et al., 2019; Rysová et al., 2019; Castilho et al., 2021). Post and Junczys-Dowmunt (2023) propose generative variants of existing contrastive test sets that are better able to discriminate different levels of quality for document-level systems. They have stated the importance of using "discourse-dense" datasets to evaluate document systems as they have shown that "the gap between translating without and with context is much larger on the discourse-dense subset" (Post and Junczys-Dowmunt, 2023). Similarly focusing on "discourse-dense" data, Wicks and Post (2023) propose an approach to building test sets that specifically focus on sentences that require additional intersentential context in order to be correctly translated. These test sets could then be used for both human and automatic evaluations that target discourse and document-level phenomena. Fernandes et al. (2023b) release benchmark data and taggers for 14 language pairs to identify phenomena that require discourse and intersentential context to disambiguate, enabling evaluation of these phenomena. They also modify Conditional Cross-Mutual Information—which they had previously used to identify the

[c]https://chat.openai.com/

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

extent to which context is used by context-aware models (Fernandes et al., 2021)—to a pointwise version for identifying particular words that are strongly context-dependent.

Another area where context and evaluation meet—when we consider a broad definition of context—is the topic of situated or task-specific evaluation of machine translation. We briefly describe a few examples of this here. Zouhar et al. (2021) evaluate NMT output by examining post-editing time; since post-editing is a real use case for translation, examining the time required to post-edit a translation provides a task-specific view of the quality of translation in the context in which it may be used (e.g., by a language service provider). Where the use case of interest is user engagement or understanding, NMT and human translation can be extrinsically compared by measuring user behaviour in A/B tests (Kovacs and DeNero, 2022). Similarly, Schmidtke (2016) used customer feedback in the format of asking whether users of help pages had found the information they needed in order to measure MT success and allow for improvements to the translations. In an in-vivo study, Mehandru et al. (2023) examined NMT quality estimation and backtranslation as tools for helping physicians using NMT to detect clinically harmful translations in discharge instructions. Roelofsen et al. (2021) provide methodological recommendations for performing evaluations of avatars for text-to-sign language translation, with a focus on two aspects of context: the community of potential users of the technology and the format in which the evaluation is performed (online).

In summary: As we explore the complexities associated with the integration of context into MT systems, the landscape of human evaluation emerges as an essential area requiring rigorous examination. Although human scores produced following current best practices have traditionally served as fundamental criteria for assessing MT performance, the exclusive dependence on these segment-level scores may not provide a sufficiently comprehensive framework to address the intricate challenges introduced by context-aware translations.

We have seen that current best practices often gravitate towards expert assessments of adequacy and fluency, and comparative analyses, with different methodologies. Moreover, the best methodologies to run context-aware human evaluations are still to be identified. The best practices of the future will need to incorporate context, in order to remain a suitable gold standard for evaluating context-aware capabilities of MT models.

The future direction of human evaluation within the context-aware MT domain is expected to present both challenges and transformative opportunities. This evolution is expected to advance improvements in translation and is also likely to have overlap and intersections with questions about bias and ethics.

3.2 Context and automatic evaluation

As mentioned previously, standard automatic evaluation metrics fail to evaluate a number of important features in the quality of longer texts since they do not seek out or measure specific discourse phenomena in the translation (Maruf et al., 2021). Therefore, developing "more robust, automatic, and interpretable document-level translation metrics" is essential (Jin et al., 2023, p.15253).

Post and Junczys-Dowmunt (2023) have identified three impediments to moving the MT field to context-aware translation, one of them being the current state of MT evaluation metrics. According to the authors, "document-level phenomena are rare enough that gains there are unlikely to be reflected with current test-set level metrics, leading to a perception, perhaps, of diminishing returns, and to the idea that the effort is not worthwhile when all costs are considered"—they argue that without an appropriate way of measuring improvements, we can't expect to see or motivate improvement. While the human evaluation improvements we discussed in the prior section are important, automatic metrics are used in the day-to-day development and rapid iteration of MT models, so it is vital to have appropriate ones available.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Efforts have been made to extend traditional automatic MT metrics to documents by merging the sentences within a document into a single expanded sentence (concatenation) and then applying the traditional metric to evaluate it (Wong and Kit, 2012; Gong, Zhang, and Zhou, 2015; Xiong et al., 2019; Liu et al., 2020; Saunders, Stahlberg, and Byrne, 2020b, i.a.). However, "this type of evaluation is still limited as it does not consider discourse aspects of the text" (Maruf et al., 2021, p. 22). In particular, this approach does not directly address important issues like consistency in terminology, anaphora and other discourse features, and overall document coherence. Some of them may be handled indirectly, for example, if the reference text being evaluated against is consistent in its terminology use and the MT output is not, the MT output may be penalised. However, any such penalty may not be directly observable or actionable as being context-related.

Several automatic metrics specifically designed for document-level MT have been introduced in the past years. Jiang et al. (2022) introduced BlonDe, a document-level automatic metric that calculates the similarity-based F1 measure of discourse-related spans across categories, taking discourse coherence into consideration. The authors compare BlonDe with 11 other metrics, including standard sentence-level metrics, document-level metrics,[d] and embedding-based metrics, evaluating them all in terms of their Pearson correlation with human assessment. The human assessment of the metric was conducted with both a sentence-level evaluation, where isolated sentences are shown, and a document-level evaluation where entire documents are shown and raters evaluate the overall quality of sequential blocks of sentences. Results showed that BlonDe is better than the other metrics at distinguishing between context-aware and context-agnostic MT systems.

The method proposed by Vernikos et al. (2022) extends pre-trained MT metrics to integrate context by constructing sentence representations from context. The approach is applied to four widely used pre-trained metrics: BERTScore (Zhang et al., 2020), Prism (Thompson and Post, 2020), COMET (Rei et al. 2020), and the reference-free metric COMET-QE (Rei et al., 2020). Evaluation involved testing the system-level correlation with human judgments using MQM judgments from the WMT21 metrics task (Freitag et al., 2021b), along with test suites. Comparisons were made with the sentence-level versions of each metric and the BlonDe metric (Jiang et al., 2022). The results demonstrated that incorporating document-level context in pre-trained metrics enhances correlation with human judgments. Additionally, the study revealed that BlonDe shows lower performance compared to both pre-trained metrics and the proposed document-level extensions. However, it is worth noting that although document context is used, this metric, now commonly called Doc-COMET, still scores single sentences one at a time.

Zhao et al. (2023) integrate discourse and BERT representations to introduce DiscoScore, a discourse metric, allowing for customisation with nouns or semantic entities. The metric was compared to 16 standard and discourse metrics using various test sets and human judgments from MQM. The study reveals that DiscoScore shows robust system-level correlation with human ratings, surpassing the current state-of-the-art BARTScore[e] in coherence, factual consistency, and other aspects, with an average improvement of over 10 correlation points. However, while it performs well at the system level over whole test sets, when it is evaluated at finer-grained levels such as the document, DiscoScore does not outperform BARTScore.

Raunak et al. (2023a) used two versions of the COMET metric, COMET20-QE (Rei et al., 2020) and COMETKiwi (Rei et al., 2022), to evaluate the performance of a simpler approach: to concatenate sentences in both the source and hypothesis. The authors present SLIDE (SLIding Document Evaluator), a document-level metric which operates by processing blocks of sentences through a sliding window across each document in the test set, with each chunk being input into an unaltered, readily available quality estimation model. The authors compared the results with Doc-COMET (Vernikos et al., 2022) on the ContraPro dataset (Müller et al., 2018). They also

[d]The document-level metric used by the authors is an intersentential linguistic feature of cohesion and coherence developed by Wong et al. (2012) which is incorporated into the sentence-based metrics.

[e]BARTScore is a metric that leverages BART, a pre-trained sequence-to-sequence model, developed by Yuan et al. (2021).

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

checked pairwise system ranking accuracy against the WMT22-MQM annotations (Freitag et al., 2022). Results showed that SLIDE achieved higher pairwise system ranking accuracy compared to its sentence-level baseline, and in certain instances, it even closed the gap with reference-based metrics. This highlights the potential of using quality estimation metrics alongside or instead of reference-based ones.

The evaluation of paragraphs was also a focus of study. Deutsch et al. (2023) examined the possibility of using neural metrics, originally trained at the sentence level, for the evaluation of paragraphs. The evaluation of the metrics is calculated via a meta-evaluation with pairwise accuracy correlations from WMT22 MQM scores. Results show that their approach trained on paragraph-level data is as effective as sentence-level metrics, but does not necessarily out-perform them. Moreover, is important to notice that their evaluation was limited to reference-based metrics.

In summary: As can be seen, the landscape of automatic evaluation for MT has witnessed notable advancements with the introduction of various metrics designed for document and paragraph-level assessments in the past few years. Efforts to extend traditional sentence-level metrics to documents have been made, but they often fall short in capturing discourse aspects of the text (Maruf et al., 2021). SLIDE, a document-level metric, leverages source context effectively, suggesting its potential to provide information akin to human references. Additionally, studies on evaluating paragraphs using neural metrics highlight comparable effectiveness but underscore the need for further exploration, particularly beyond reference-based metrics.

Recent introductions showcase promising performance at the system level. However, challenges persist, as some metrics, including BlonDe, exhibit lower efficacy compared to pre-trained metrics or document-level extensions.

Moreover, what the community still needs to understand is the limitation of each of these proposed approaches. There is also the need to refine existing metrics, address limitations in capturing nuanced linguistic phenomena, and explore more comprehensive evaluation strategies. All of these are critical for advancing the field of automatic MT evaluation. Future research should aim to develop metrics that align more closely with human judgments across diverse linguistic contexts and encompass a broader understanding of document- and paragraph-level translations. Another fruitful area of research may be in explainable metrics, which could provide more information to system developers about what areas of translation quality still need improvement.

4. Large language models

Since the initial call for papers for this special issue, we have seen a major shift in the size and performance of large language models (LLMs). This has included an interest in experimenting with LLMs for the task of MT as well as evaluating LLMs on MT tasks. We begin by discussing LLMs for translation and then examine their uses in evaluation.

4.1 LLMs for translation

The most recent edition of the Conference on Machine Translation marked the first time that LLMs were used directly as submissions in the shared task on General Machine Translation (Kocmi et al., 2023). That task found that GPT-4[f] performed in the top cluster for systems translating into English, while it dropped somewhat in the out-of-English direction.

While there has been wide interest in examining LLM performance on a wide range of tasks, there is a particular reason that LLM performance on MT tasks is closely tied to this work on context: the use of LLMs that permit very large intersentential context windows. Pawar et al.

[f]https://openai.com/gpt-4

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

(2024) provide a survey of techniques used for increasing the context span that is used by LLMs. Given our discussion of the role of intersentential or document context in MT, it is clear that the approaches from LLMs provide an opportunity to explore these questions. This can be explored both from the perspective of using LLMs for translation, as well as when or how such approaches could be incorporated into MT systems. A number of works have begun to examine how LLMs can be used as translation systems, examining their performance either at the sentence level or beyond. Other work has touched on other types of context examined in this survey.

Examining ChatGPT's performance as an MT system across a wide range of languages and comparing against other LLMs and dedicated MT systems, Robinson et al. (2023) observe that LLMs exhibit competitive behaviour on some high-resource languages, but underperform dedicated MT systems for the majority of low-resource languages. These benchmarks are intended to provide end users with an understanding of the level of quality they might expect for translation into their languages using LLMs. The paper finds that the number of existing Wikipedia pages in a given language is a strong predictor of LLM performance as a translation system for that language. Manakhimova et al. (2023) examine the performance of GPT-4 on fine-grained linguistic nuances, explored through the use of test suites. They observe performance competitive with dedicated MT systems, though they also note that there are still areas with room for improvement.

In their submission to the shared task, the Lan-Bridge team explored different natural language prompts for using GPT-3.5 for translation between Chinese and English (Wu and Hu, 2023). These included prompts that focused on sentence-by-sentence translation, as well as ones that asked the model to attend to intersentential context as well, concluding that the latter improved translation quality according to automatic metrics. Wang et al. (2023) also examine a variety of prompts for document-level translation with LLMs, comparing several MT models and testing their performance on the discourse-focused test suite from Voita et al. (2019b), finding recent performance improvements on discourse in document-level translation with GPT-4. Karpinska and Iyyer (2023) consider the task of literary translation, prompting GPT-3.5 to perform sentence-level translation, sentence-level translation with context, and paragraph-level translation. Human evaluation finds the best performance with paragraph-level translation, but still notes critical errors even in these better translations. Also focusing in the literature domain, Thai et al. (2022) explored using LLMs to perform post-editing in order to mitigate discourse errors, after finding a large gap in quality between MT and human literary translation in human evaluations. Much of the work on LLMs has focused on prompting and few-shot learning (Briakou, Cherry, and Foster, 2023; Raunak, Menezes, and Awadalla, 2023b, i.a.), but in addition to these, Zhang et al. (2023) explore finetuning of LLMs for translation, finding performance improvements even when finetuning only a very small fraction of the model's parameters. Iyer et al. (2023) consider the translation of ambiguous words with LLMs, including few-shot (or in-context) learning, where additional examples are fed to the model in order for it to use that contextual information for disambiguation.

On the topic of formality control discussed in Section 2.3, Marrese-Taylor et al. (2023) explore prompting LLMs to handle formality control for English to Japanese translation, finding it to be a viable approach. Touching on the topics discussed in Section 2.4, several papers have considered how to use LLMs to incorporate terminology into translation. Moslem et al. (2023) explore ways of using LLMs for improving terminology translation, both by using the LLM to produce synthetic bitext with the terminology, as well as by using the LLM to postedit MT output to incorporate desired terminology. On the same topic, Bogoychev and Chen (2023) prompt LLMs to revise existing translations so that they incorporate terminology constraints.

In general, the papers discussed so far have looked at evaluating or improving performance of LLMs as translation systems. Petrick et al. (2023) explore ways of fusing document-level language models with NMT systems, including LLMs. This provides an example of ways we might see knowledge from LLMs incorporated into dedicated MT systems, an avenue that could be explored in parallel to that of treating LLMs as MT systems.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

In summary: The topic of LLMs has seen growing interest in the field of MT. This initial interest has often focused on evaluating LLMs as translation systems, exploring how to modify prompts or provide additional context to improve translation, as well as ways to combine them with MT systems. We expect that this is an area that will continue to see growth and exploration in the immediate future, including improved understanding of suitable prompts and the data already incorporated in LLMs (Briakou et al., 2023), methods of finetuning, or methods of combining LLMs and MT systems. As with all work on LLMs, a challenge for the research field will be how to ensure reproducibility and fair evaluation, given a landscape of proprietary models where full information about training data may not be available.

4.2 LLMs for evaluation

As previously discussed, we have seen a growing interest in the application of LLMs to the domain of automatic translation, among many other avenues of exploration. Beyond translation, researchers have explored the potential of LLMs to assess the quality of automatic translations as well. An example is the work by Kocmi and Federmann (2023b) who introduced GEMBA, a GPT-based metric designed for translation evaluation. The authors compared the metric in two forms—with a reference translation and without—against COMET and BLEURT scores, and compared the results of the MQM-based human labels from the WMT22 Metrics Shared task. Results on the system level show that the metric achieved state-of-the-art performance even outperforming traditional metrics and human evaluations in some cases. The authors state that this is a great step towards document-level evaluation as these models have the ability to use much larger context windows.

Another metric leveraging LLMs is the one introduced by Fernandes et al. (2023a). Their work investigates the capabilities of these models to perform automatic MQM-style error annotation and score prediction. The authors introduce AutoMQM, a prompt technique for evaluating translation quality that instructs the model to identify and categorise errors rather than asking it to assign a numerical score. Their results show that prompting LLMs to predict a quality score does not improve performance over the trained standard automatic metrics. Moreover, the authors claim that without fine-tuning, AutoMQM can provide interpretable results through error spans that correlate well with human annotations. GEMBA-MQM (Kocmi and Federmann, 2023a), a follow-up work to Kocmi and Federmann (2023b), also aims to use LLMs to identify MQM-style error spans, though Lo, Larkin & Knowles (2023) note that it may struggle to distinguish between lower-quality systems (which are distinguishable by other existing metrics).

In the area of human judgements on question-answering systems, but potentially applicable to NMT evaluation, Wadhwa et al. (2023) use LLMs to rescale human judgments by analysing human explanations of the scores they assigned. The methodology involves inputting both the label and its corresponding explanation provided by annotators into the LLM, generating a score on a scale from 0 to 100 based on a rubric. The findings demonstrate a strong correlation with expert annotations.

Finally, in their recent work, Huang et al. (2024) investigate how LLMs leverage source and reference information in evaluating translations, by employing coarse-grained and fine-grained prompts. The authors instruct both open and closed LLMs to predict coarse-grained quality scores like GEMBA, but given different information such as sources and references. Furthermore, the authors adopt the AutoMQM prompt template for fine-grained error detection. The results highlight that reference information enhances system-level accuracy and segment-level correlations. Interestingly, the utilisation of source information is at times counterproductive, suggesting limitations in cross-lingual capability when employing LLMs for evaluating translated sentences.

In summary: As can be seen, the rise of LLMs in the field of MT evaluation shows substantial promise and brings intriguing challenges. These new developments with the capacity of handling more context can reshape the way we evaluate translations. However, as we dive in this emerging

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

area, it becomes clear that there are still aspects that need refinement. The ongoing pursuit of best strategies to use LLMs in MT evaluation will likely be an area of continuing interest. Moreover, the way we test these metrics should be strict and well-defined, with human evaluation playing a part in the process. These metrics should also be evaluated on a wide range of levels of system quality, as well as across diverse language pairs to highlight potential weaknesses in both the approach to evaluation (Lo et al., 2023) and the underlying model's language-specific capabilities (see, e.g., Robinson et al., 2023). The road ahead involves not only addressing the current limitations but also pushing the boundaries, guiding the field toward a more comprehensive and context-aware era in MT evaluation.

5. Conclusion and future outlook

The exploration of context in MT has brought to light the intricate dynamics between linguistic elements and external factors, shedding light on various dimensions such as intersentential context, world knowledge, external information, and the treatment of terminology.

In this survey, we have examined diverse methodologies for integrating various types of context into NMT systems, offering a comprehensive map of the evolving landscape in this domain. From leveraging intersentential context to harnessing world knowledge and external information, and addressing the nuances of terminology, these explorations have unveiled insights into handling the contextual dimensions of machine-generated translations.

This broad range of types of context is rarely viewed under this single umbrella of "context in NMT," but rather seen as a number of discrete challenges or tasks. By bringing them together in this survey, we can see ways in which there are connections between different types of context and the approaches applied to address them. For example, while handling terminology is often considered its own task, it shares with topics like polysemy, anaphora, and other areas the challenge of introducing consistency across sentences in translation. Adopting a broad view of what constitutes "context" could lead to fruitful exchanges between these different areas of study.

In the evaluation area, we have discussed efforts to adapt and extend human evaluation to account for document-level and intersentential context. These approaches face challenges related to balancing annotator fatigue and preference, inter-annotator agreement, and ability to capture discourse and other context-related features. Much like the initial approaches to incorporating intersentential context into NMT systems, some of the first attempts to incorporating context into automatic metrics have involved expanding the span considered by the metric to the paragraph or document level. We have observed that this still has weaknesses, as the metrics used may not capture important discourse features or certain types of consistency (and inconsistency) in translation. Novel approaches that focus on these questions are currently being developed and evaluated. One additional challenge to building automatic metrics that incorporate context is the fact that best practices for human evaluation that incorporates context are still in the process of being developed and refined. Automatic metrics are typically tested by comparing them against human evaluations; while best practices for context-aware human evaluation are still in flux, the goalposts for automatic metrics keep moving.

At the same time as this shift towards context-aware evaluation, we have seen continued interest in evaluating MT for biases and along ethical dimensions. While the approaches that we have described in this section offer valuable insights into the capabilities and limitations of context-aware MT systems, they may not fully address the ethical dimensions of complex problems such as racial, gender, and other biases. For some of these questions, there are already clear points of connection between context and the examination of bias, such as gender and anaphoric pronouns. For others, these connections may currently be less clear. Consequently, it is unknown whether these existing evaluation approaches can effectively gauge biases, thereby providing opportunities to mitigate them through improving the MT systems. We will likely need context-aware evaluation scoring approaches and methodologies that can adeptly navigate these complex ethical issues.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Simultaneously, we should be conscious of the ethics and best practices surrounding the practices of human annotation itself (e.g., fair pay, best practices that consider annotator preferences and ergonomics, whether the data being annotated contains graphic or disturbing content, clear instructions and guidelines, inclusion of language communities in the whole process of system design rather than only evaluation, and more). The growing recognition of ethical considerations underscores the urgency of reevaluating and potentially re-calibrating our evaluation frameworks to encompass a broader spectrum of evaluative criteria.

Of course, the tasks of translation and evaluation are not independent of one another. We can see this interplay in many areas related to context. Studying how much context is necessary in order to perform high-quality evaluation of translation output can also inform us about when and how much context is needed to perform the translation task itself. As MT continues to become more widely used, in a broader range of settings, we also expect to see the growth—both within the MT field itself and in fields using it—of extrinsic evaluations or task-specific MT evaluations. The field of MT and the fields using MT would likely both benefit from careful and considered collaborations on MT use cases and their appropriate evaluations, drawing on the MT researchers' knowledge of the systems and the users' knowledge of the setting.

As we look to the future, we anticipate that context will continue to play an even greater role. Already we see this in the much larger context windows used by LLMs. Motivated by the increasing performance of MT systems and the subsequent challenges in evaluation, the community has seen the need for improved context-aware evaluation and a shift towards document-level translation. Nevertheless, how best to perform this evaluation or these larger context translations remains an area with much room for innovation as well as careful analysis. We know that even imperfect metrics (for example BLEU) have contributed to progress in the field, but we expect that improved techniques for automatic and human evaluation of translations *in context* will also benefit researchers focusing on improving document-level or context-aware translation. We expect to see research pushing the boundaries with respect to context on both the translation and evaluation fronts in the coming years.

Acknowledgements. We thank Chi-kiu Lo, Michel Simard, and Gabriel Bernier-Colborne for their comments and feedback on this work. The first author had the financial support of Science Foundation Ireland at the ADAPT Centre, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University [13/RC/2106_P2].

References

Agrawal R., **Turchi M.** and **Negri M.** (2018). Contextual handling in neural machine translation: Look behind, ahead and on both sides. In **Pérez-Ortiz J.A.**, **Sánchez-Martínez F.**, **Esplà-Gomis M.**, **Popović M.**, **Rico C.**, **Martins A.**, **Van den Bogaert J.** and **Forcada M.L.** (eds), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alicante, Spain, pp. 31–40.

Agrawal S. and **Carpuat M.** (2019). Controlling text complexity in neural machine translation. In **Inui K.**, **Jiang J.**, **Ng V.** and **Wan X.** (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 1549–1564.

Ailem M., **Liu J.** and **Qader R.** (2021). Encouraging neural machine translation to satisfy terminology constraints. In **Zong C.**, **Xia F.**, **Li W.** and **Navigli R.** (eds), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online. Association for Computational Linguistics, pp. 1450–1455.

Al Ghussin Y., **Zhang J.** and **van Genabith J.** (2023). Exploring paracrawl for document-level neural machine translation. In **Vlachos A.** and **Augenstein I.** (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics, pp. 1304–1310.

Alam M.M.I., **Kvapilíková I.**, **Anastasopoulos A.**, **Besacier L.**, **Dinu G.**, **Federico M.**, **Gallé M.**, **Jung K.**, **Koehn P.** and **Nikoulina V.** (2021). Findings of the WMT shared task on machine translation using terminologies. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussa M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Kocmi T.**, **Martins A.**, **Morishita M.** and **Monz C.** (eds), *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 652–663.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Arthur P., **Neubig G.** and **Nakamura S.** (2016). Incorporating discrete translation lexicons into neural machine translation. In **Su J.**, **Duh K.** and **Carreras X.** (eds), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Association for Computational Linguistics, pp. 1557–1567.

Bahdanau D., **Cho K.** and **Bengio Y.** (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*, San Diego, CA.

Ballier N., **Cho D.**, **Faye B.**, **Ke Z.-Y.**, **Martikainen H.**, **Pecman M.**, **Wisniewski G.**, **Yunès J.-B.**, **Zhu L.** and **Zimina-Poirot M.** (2021). The SPECTRANS system description for the WMT21 terminology task. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussa M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Kocmi T.**, **Martins A.**, **Morishita M.** and **Monz C.** (eds), *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 813–820.

Bane F., **Zaretskaya A.**, **Miró T.B.**, **Uguet C.S.** and **Torres J.** (2023). Coming to terms with glossary enforcement: A study of three approaches to enforcing terminology in NMT. In **Nurminen M.**, **Brenner J.**, **Koponen M.**, **Latomaa S.**, **Mikhailov M.**, **Schierl F.**, **Ranasinghe T.**, **Vanmassenhove E.**, **Vidal S.A.**, **Aranberri N.**, **Nunziatini M.**, **Escartín C.P.**, **Forcada M.**, **Popovic M.**, **Scarton C.** and **Moniz H.** (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland. European Association for Machine Translation, pp. 345–353.

Barrault L., **Biesialska M.**, **Bojar O.**, **Costa-jussà M.R.**, **Federmann C.**, **Graham Y.**, **Grundkiewicz R.**, **Haddow B.**, **Huck M.**, **Joanis E.**, **Kocmi T.**, **Koehn P.**, **Lo C.-k.**, **Ljubešić N.**, **Monz C.**, **Morishita M.**, **Nagata M.**, **Nakazawa T.**, **Pal S.**, **Post M.** and **Zampieri M.** (2020). Findings of the 2020 conference on machine translation (WMT20). In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Graham Y.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.** and **Negri M.** (eds), *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 1–55.

Barrault L., **Bojar O.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Koehn P.**, **Malmasi S.**, **Monz C.**, **Müller M.**, **Pal S.**, **Post M.** and **Zampieri M.** (2019). Findings of the 2019 conference on machine translation (WMT19). In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy. Association for Computational Linguistics, pp. 1–61.

Barrault L., **Bougares F.**, **Specia L.**, **Lala C.**, **Elliott D.** and **Frank S.** (2018). Findings of the third shared task on multimodal machine translation. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Specia L.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels. Association for Computational Linguistics, pp. 304–323.

Bawden R., **Sennrich R.**, **Birch A.** and **Haddow B.** (2018). Evaluating discourse phenomena in neural machine translation. In **Walker M.**, **Ji H.** and **Stent A.** (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1304–1313.

Bergmanis T. and **Pinnis M.** (2021a). Dynamic terminology integration for COVID-19 and other emerging domains. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussa M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Kocmi T.**, **Martins A.**, **Morishita M.** and **Monz C.** (eds), *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 821–827.

Bergmanis T. and **Pinnis M.** (2021b). Facilitating terminology translation with target lemma annotations. In **Merlo P.**, **Tiedemann J.** and **Tsarfaty R.** (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online. Association for Computational Linguistics, pp. 3105–3111.

Bogoychev N. and **Chen P.** (2023). Terminology-aware translation with constrained decoding and large language model prompting. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 890–896.

Bragg D., **Caselli N.**, **Hochgesang J.A.**, **Huenerfauth M.**, **Katz-Hernandez L.**, **Koller O.**, **Kushalnagar R.**, **Vogler C.** and **Ladner R.E.** (2021). The fate landscape of sign language AI datasets: An interdisciplinary perspective. *ACM Transactions on Computer-Human Interaction*, **14**(2).

Briakou E., **Cherry C.** and **Foster G.** (2023). Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In **Rogers A.**, **Boyd-Graber J.** and **Okazaki N.** (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics, pp. 9432–9452.

Britz D., **Le Q.** and **Pryzant R.** (2017). Effective domain mixing for neural machine translation. In **Bojar O.**, **Buck C.**, **Chatterjee R.**, **Federmann C.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.** and **Kreutzer J.** (eds), *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 118–126.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Burlot F. and **Yvon F.** (2018). Using monolingual data in neural machine translation: a systematic study. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Specia L.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics, pp. 144–155.

Cai X. and **Xiong D.** (2020). A test suite for evaluating discourse phenomena in document-level neural machine translation. In **Liu Q.**, **Xiong D.**, **Ge S.** and **Zhang X.** (eds), *Proceedings of the Second International Workshop of Discourse Processing*, Suzhou, China. Association for Computational Linguistics, pp. 13–17.

Carpuat M. (2009). One translation per discourse. In **Agirre E.**, **Màrquez L.** and **Wicentowski R.** (eds), *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, Boulder, Colorado. Association for Computational Linguistics, pp. 19–27.

Castilho S. (2020). On the same page? Comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Graham Y.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.** and **Negri M.** (eds), *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 1150–1159.

Castilho S. (2021). Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In **Belz A.**, **Agarwal S.**, **Graham Y.**, **Reiter E.** and **Shimorina A.** (eds), *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, Online. Association for Computational Linguistics, pp. 34–45.

Castilho S. (2022). How much context span is enough? Examining context-related issues for document-level MT. In **Calzolari N.**, **Béchet F.**, **Blache P.**, **Choukri K.**, **Cieri C.**, **Declerck T.**, **Goggi S.**, **Isahara H.**, **Maegaard B.**, **Mariani J.**, **Mazo H.**, **Odijk J.** and **Piperidis S.** (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 3017–3025.

Castilho S., **Cavalheiro Camargo J.L.**, **Menezes M.** and **Way A.** (2021). DELA corpus - a document-level corpus annotated with context-related issues. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussa M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Kocmi T.**, **Martins A.**, **Morishita M.** and **Monz C.** (eds), *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 566–577.

Castilho S., **Mallon C.Q.**, **Meister R.** and **Yue S.** (2023). Do online machine translation systems care for context? What about a GPT model? In **Nurminen M.**, **Brenner J.**, **Koponen M.**, **Latomaa S.**, **Mikhailov M.**, **Schierl F.**, **Ranasinghe T.**, **Vanmassenhove E.**, **Vidal S.A.**, **Aranberri N.**, **Nunziatini M.**, **Escartín C.P.**, **Forcada M.**, **Popovic M.**, **Scarton C.** and **Moniz H.** (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland. European Association for Machine Translation, pp. 393–417.

Castilho S., **Moorkens J.**, **Gaspari F.**, **Calixto I.**, **Tinsley J.** and **Way A.** (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, **108**(1), 109–120.

Castilho S., **Popović M.** and **Way A.** (2020). On context span needed for machine translation evaluation. In **Calzolari N.**, **Béchet F.**, **Blache P.**, **Choukri K.**, **Cieri C.**, **Declerck T.**, **Goggi S.**, **Isahara H.**, **Maegaard B.**, **Mariani J.**, **Mazo H.**, **Moreno A.**, **Odijk J.** and **Piperidis S.** (eds), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 3735–3742.

Chatterjee R., **Negri M.**, **Turchi M.**, **Federico M.**, **Specia L.** and **Blain F.** (2017). Guiding neural machine translation decoding with external knowledge. In **Bojar O.**, **Buck C.**, **Chatterjee R.**, **Federmann C.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.** and **Kreutzer J.** (eds), *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 157–168.

Cherry C., **Arivazhagan N.**, **Padfield D.** and **Krikun M.** (2021). Subtitle translation as markup translation. In *Proceedings of Interspeech 2021*, pp. 2237–2241.

Chu C., **Dabre R.** and **Kurohashi S.** (2017). An empirical comparison of domain adaptation methods for neural machine translation. In **Barzilay R.** and **Kan M.-Y.** (eds), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 385–391.

Chu C. and **Wang R.** (2018). A survey of domain adaptation for neural machine translation. In **Bender E.M.**, **Derczynski L.** and **Isabelle P.** (eds), *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 1304–1319.

Dalvi F., **Durrani N.**, **Sajjad H.** and **Vogel S.** (2018). Incremental decoding and training methods for simultaneous translation in neural machine translation. In **Walker M.**, **Ji H.** and **Stent A.** (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

De Meulder M. (2021). Is "good enough" good enough? ethical and responsible development of sign language technologies. In **Shterionov D.** (ed.), *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, Virtual. Association for Machine Translation in the Americas, pp. 12–22.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Deutsch D., **Juraska J.**, **Finkelstein M.** and **Freitag M.** (2023). Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 996–1013.

Dinu G., **Mathur P.**, **Federico M.** and **Al-Onaizan Y.** (2019). Training neural machine translation to apply terminology constraints. In **Korhonen A.**, **Traum D.** and **Màrquez L.** (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 3063–3068.

Dobreva R., **Zhou J.** and **Bawden R.** (2020). Document sub-structure in neural machine translation. In **Calzolari N.**, **Béchet F.**, **Blache P.**, **Choukri K.**, **Cieri C.**, **Declerck T.**, **Goggi S.**, **Isahara H.**, **Maegaard B.**, **Mariani J.**, **Mazo H.**, **Moreno A.**, **Odijk J.** and **Piperidis S.** (eds), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 3657–3667.

Elliott D., **Frank S.**, **Barrault L.**, **Bougares F.** and **Specia L.** (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In **Bojar O.**, **Buck C.**, **Chatterjee R.**, **Federmann C.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.** and **Kreutzer J.** (eds), *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 215–233.

Elliott D., **Frank S.** and **Hasler E.** (2015). Multi-language image description with neural sequence models. CoRR, abs/1510.04709.

Exel M., **Buschbeck B.**, **Brandt L.** and **Doneva S.** (2020). Terminology-constrained neural machine translation at SAP. In **Martins A.**, **Moniz H.**, **Fumega S.**, **Martins B.**, **Batista F.**, **Coheur L.**, **Parra C.**, **Trancoso I.**, **Turchi M.**, **Bisazza A.**, **Moorkens J.**, **Guerberof A.**, **Nurminen M.**, **Marg L.** and **Forcada M.L.** (eds), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal. European Association for Machine Translation, pp. 271–280.

Fadaee M., **Bisazza A.** and **Monz C.** (2017). Data augmentation for low-resource neural machine translation. In **Barzilay R.** and **Kan M.-Y.** (eds), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 567–573.

Farajian M.A., **Turchi M.**, **Negri M.** and **Federico M.** (2017). Multi-domain neural machine translation through unsupervised adaptation. In **Bojar O.**, **Buck C.**, **Chatterjee R.**, **Federmann C.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.** and **Kreutzer J.** (eds), *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 127–137.

Feely W., **Hasler E.** and **de Gispert A.** (2019). Controlling Japanese honorifics in English-to-Japanese neural machine translation. In **Nakazawa T.**, **Ding C.**, **Dabre R.**, **Kunchukuttan A.**, **Doi N.**, **Oda Y.**, **Bojar O.**, **Parida S.**, **Goto I.** and **Mino H.** (eds), Proceedings of the 6th Workshop on Asian Translation, Hong Kong, China. Association for Computational Linguistics, pp. 45–53.

Fernandes P., **Deutsch D.**, **Finkelstein M.**, **Riley P.**, **Martins A.**, **Neubig G.**, **Garg A.**, **Clark J.**, **Freitag M.** and **Firat O.** (2023a). The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 1066–1083.

Fernandes P., **Yin K.**, **Liu E.**, **Martins A.** and **Neubig G.** (2023b). When does translation require context? a data-driven, multilingual exploration. In **Rogers A.**, **Boyd-Graber J.** and **Okazaki N.** (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics, pp. 606–626.

Fernandes P., **Yin K.**, **Neubig G.** and **Martins A.F.T.** (2021). Measuring and increasing context usage in context-aware machine translation. In **Zong C.**, **Xia F.**, **Li W.** and **Navigli R.** (eds), *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)*, Online. Association for Computational Linguistics, pp. 6467–6478.

Freitag M., **Foster G.**, **Grangier D.**, **Ratnakar V.**, **Tan Q.** and **Macherey W.** (2021a). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, **9**, 1460–1474.

Freitag M., **Mathur N.**, **Lo C.-k.**, **Avramidis E.**, **Rei R.**, **Thompson B.**, **Kocmi T.**, **Blain F.**, **Deutsch D.**, **Stewart C.**, **Zerva C.**, **Castilho S.**, **Lavie A.** and **Foster G.** (2023). Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 578–628.

Freitag M., **Rei R.**, **Mathur N.**, **Lo C.-k.**, **Stewart C.**, **Avramidis E.**, **Kocmi T.**, **Foster G.**, **Lavie A.** and **Martins A.F.T.** (2022). Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In **Koehn P.**, **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Jimeno Yepes A.**, **Kocmi T.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Popel M.**, **Turchi M.** and **Zampieri M.** (eds), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 46–68.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Freitag M., **Rei R.**, **Mathur N.**, **Lo C.-k.**, **Stewart C.**, **Foster G.**, **Lavie A.** and **Bojar O.** (2021b). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussa M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Kocmi T.**, **Martins A.**, **Morishita M.** and **Monz C.** (eds), *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 733–774.

Gale W.A., **Church K.W.** and **Yarowsky D.** (1992). One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman*, New York, February 23-26.

Gete H., **Etchegoyhen T.** and **Labaka G.** (2023). What works when in context-aware neural machine translation? In **Nurminen M.**, **Brenner J.**, **Koponen M.**, **Latomaa S.**, **Mikhailov M.**, **Schierl F.**, **Ranasinghe T.**, **Vanmassenhove E.**, **Vidal S.A.**, **Aranberri N.**, **Nunziatini M.**, **Escartín C.P.**, **Forcada M.**, **Popovic M.**, **Scarton C.** and **Moniz H.** (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland. European Association for Machine Translation, pp. 147–156.

Gong Z., **Zhang M.** and **Zhou G.** (2015). Document-level machine translation evaluation with gist consistency and text cohesion. In **Webber B.**, **Carpuat M.**, **Popescu-Belis A.** and **Hardmeier C.** (eds), *Proceedings of the Second Workshop on Discourse in Machine Translation*, Lisbon, Portugal. Association for Computational Linguistics, pp. 33–40.

Graham Y., **Baldwin T.**, **Dowling M.**, **Eskevich M.**, **Lynn T.** and **Tounsi L.** (2016). Is all that glitters in machine translation quality estimation really gold? In **Matsumoto Y.** and **Prasad R.** (eds), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan. The COLING 2016 Organizing Committee, pp. 3124–3134.

Grissom II **A.**, **He H.**, **Boyd-Graber J.**, **Morgan J.** and **Daumé** III **H.** (2014). Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In **Moschitti A.**, **Pang B.** and **Daelemans W.** (eds), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics, pp. 1342–1352.

Gu J., **Neubig G.**, **Cho K.** and **Li V.O.** (2017). Learning to translate in real-time with neural machine translation. In **Lapata M.**, **Blunsom P.** and **Koller A.** (eds), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain. Association for Computational Linguistics, pp. 1053–1062.

Gu J., **Wang C.** and **Zhao J.** (2019). Levenshtein transformer. In **Wallach H.**, **Larochelle H.**, **Beygelzimer A.**, **d' Alché-Buc F.**, **Fox E.** and **Garnett R.** (eds), *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., pp. 11179–11189.

Guillou L., **Hardmeier C.**, **Lapshinova-Koltunski E.** and **Loáiciga S.** (2018). A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Specia L.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels. Association for Computational Linguistics, pp. 570–577.

Guo Q., **Qiu X.**, **Liu P.**, **Shao Y.**, **Xue X.** and **Zhang Z.** (2019). Star-transformer. In **Burstein J.**, **Doran C.** and **Solorio T.** (eds), *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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 1315–1325.

Hardmeier C. (2012). Discourse in statistical machine translation. A survey and a case study. Discours. Revue de linguistique, psycholinguistique et informatique. *A Journal of Linguistics, Psycholinguistics and Computational Linguistics*, **11**.

Hasler E., **de Gispert A.**, **Iglesias G.** and **Byrne B.** (2018). Neural machine translation decoding with terminology constraints. In **Walker M.**, **Ji H.** and **Stent A.** (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 506–512.

Hassan H., **Aue A.**, **Chen C.**, **Chowdhary V.**, **Clark J.**, **Federmann C.**, **Huang X.**, **Junczys-Dowmunt M.**, **Lewis W.**, **Li M.**, **Liu S.**, **Liu T.**, **Luo R.**, **Menezes A.**, **Qin T.**, **Seide F.**, **Tan X.**, **Tian F.**, **Wu L.**, **Wu S.**, **Xia Y.**, **Zhang D.**, **Zhang Z.** and **Zhou M.** (2018). Achieving human parity on automatic Chinese to English news translation. CoRR, abs/1803.05567.

Herold C. and **Ney H.** (2023). Improving long context document-level machine translation. In **Strube M.**, **Braud C.**, **Hardmeier C.**, **Li J.J.**, **Loaiciga S.** and **Zeldes A.** (eds), *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, Toronto, Canada. Association for Computational Linguistics, pp. 112–125.

Hitschler J., **Schamoni S.** and **Riezler S.** (2016). Multimodal pivots for image caption translation. In **Erk K.** and **Smith N.A.** (eds), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 2399–2409.

Hokamp C. and **Liu Q.** (2017). Lexically constrained decoding for sequence generation using grid beam search. In **Barzilay R.** and **Kan M.-Y.** (eds), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 1535–1546.

Hu J.E., **Khayrallah H.**, **Culkin R.**, **Xia P.**, **Chen T.**, **Post M.** and **Van Durme B.** (2019). Improved lexically constrained decoding for translation and monolingual rewriting. In **Burstein J.**, **Doran C.** and **Solorio T.** (eds), *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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 839–850.

Huang X., **Zhang Z.**, **Geng X.**, **Du Y.**, **Chen J.** and **Huang S.** (2024). Lost in the source language: How large language models evaluate the quality of machine translation.

Huo J., **Herold C.**, **Gao Y.**, **Dahlmann L.**, **Khadivi S.** and **Ney H.** (2020). Diving deep into context-aware neural machine translation. In **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Graham Y.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.** and **Negri M.** (eds), *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics, pp. 604–616.

Iyer V., **Chen P.** and **Birch A.** (2023). Towards effective disambiguation for machine translation with large language models. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 482–495.

Jean S. and **Cho K.** (2019). Context-aware learning for neural machine translation. CoRR, abs/1903.04715.

Jean S., **Cho K.**, **Memisevic R.** and **Bengio Y.** (2015). On using very large target vocabulary for neural machine translation. In **Zong C.** and **Strube M.** (eds), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China. Association for Computational Linguistics, pp. 1–10.

Jean S., **Lauly S.**, **Firat O.** and **Cho K.** (2017). Does neural machine translation benefit from larger context?

Jiang H., **Liang C.**, **Wang C.** and **Zhao T.** (2020). Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. In **Jurafsky D.**, **Chai J.**, **Schluter N.** and **Tetreault J.** (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 1823–1834.

Jiang Y., **Liu T.**, **Ma S.**, **Zhang D.**, **Yang J.**, **Huang H.**, **Sennrich R.**, **Cotterell R.**, **Sachan M.** and **Zhou M.** (2022). BlonDe: An automatic evaluation metric for document-level machine translation. In **Carpuat M.**, **de Marneffe M.-C.** and **Meza Ruiz I.V.** (eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics, pp. 1550–1565.

Jin L., **He J.**, **May J.** and **Ma X.** (2023). Challenges in context-aware neural machine translation. In **Bouamor H.**, **Pino J.** and **Bali K.** (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, pp. 15246–15263.

Junczys-Dowmunt M. (2019). Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy. Association for Computational Linguistics, pp. 225–233.

Kang X., **Zhao Y.**, **Zhang J.** and **Zong C.** (2020). Dynamic context selection for document-level neural machine translation via reinforcement learning. In **Webber B.**, **Cohn T.**, **He Y.** and **Liu Y.** (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 2242–2254.

Karpinska M. and **Iyyer M.** (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 419–451.

Khandelwal U., **Fan A.**, **Jurafsky D.**, **Zettlemoyer L.** and **Lewis M.** (2021). Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Kim Y., **Tran D.T.** and **Ney H.** (2019). When and why is document-level context useful in neural machine translation? In **Popescu-Belis A.**, **Loáiciga S.**, **Hardmeier C.** and **Xiong D.** (eds), *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, Hong Kong, China. Association for Computational Linguistics, pp. 24–34.

Knowles R. and **Koehn P.** (2016). Neural interactive translation prediction. In **Green S.** and **Schwartz L.** (eds), *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, Austin, TX, USA. The Association for Machine Translation in the Americas, pp. 107–120.

Knowles R., **Larkin S.**, **Tessier M.** and **Simard M.** (2023). Terminology in neural machine translation: A case study of the Canadian Hansard. In **Nurminen M.**, **Brenner J.**, **Koponen M.**, **Latomaa S.**, **Mikhailov M.**, **Schierl F.**, **Ranasinghe T.**, **Vanmassenhove E.**, **Vidal S.A.**, **Aranberri N.**, **Nunziatini M.**, **Escartín C.P.**, **Forcada M.**, **Popovic M.**, **Scarton C.** and **Moniz H.** (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland. European Association for Machine Translation, pp. 481–488.

Kobus C., **Crego J.** and **Senellart J.** (2017). Domain control for neural machine translation. In **Mitkov R.** and **Angelova G.** (eds), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria. INCOMA Ltd., pp. 372–378

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Kocmi T., **Avramidis E.**, **Bawden R.**, **Bojar O.**, **Dvorkovich A.**, **Federmann C.**, **Fishel M.**, **Freitag M.**, **Gowda T.**, **Grundkiewicz R.**, **Haddow B.**, **Koehn P.**, **Marie B.**, **Monz C.**, **Morishita M.**, **Murray K.**, **Nagata M.**, **Nakazawa T.**, **Popel M.**, **Popović M.** and **Shmatova M.** (2023). Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 1–42.

Kocmi T. and **Federmann C.** (2023a). GEMBA-MQM: Detecting translation quality error spans with GPT-4. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 768–775.

Kocmi T. and **Federmann C.** (2023b). Large language models are state-of-the-art evaluators of translation quality. In **Nurminen M.**, **Brenner J.**, **Koponen M.**, **Latomaa S.**, **Mikhailov M.**, **Schierl F.**, **Ranasinghe T.**, **Vanmassenhove E.**, **Vidal S.A.**, **Aranberri N.**, **Nunziatini M.**, **Escartín C.P.**, **Forcada M.**, **Popovic M.**, **Scarton C.** and **Moniz H.** (eds), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland. European Association for Machine Translation, pp. 193–203.

Kocmi T., **Matsushita H.** and **Federmann C.** (2022). MS-COMET: More and better human judgements improve metric performance. In **Koehn P.**, **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Jimeno Yepes A.**, **Kocmi T.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Popel M.**, **Turchi M.** and **Zampieri M.** (eds), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 541–548.

Kothur S.S.R., **Knowles R.** and **Koehn P.** (2018). Document-level adaptation for neural machine translation. In **Birch A.**, **Finch A.**, **Luong T.**, **Neubig G.** and **Oda Y.** (eds), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia. Association for Computational Linguistics, pp. 64–73.

Kovacs G. and **DeNero J.** (2022). Measuring the effects of human and machine translation on website engagement. In **Duh K.** and **Guzmán F.** (eds), *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Orlando, USA. Association for Machine Translation in the Americas, pp. 298–308.

Lakew S.M., **Virkar Y.**, **Mathur P.** and **Federico M.** (2022). Isometric MT: Neural machine translation for automatic dubbing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6242–6246.

Läubli S., **Castilho S.**, **Neubig G.**, **Sennrich R.**, **Shen Q.** and **Toral A.** (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, **67**, 653–672.

Läubli S., **Sennrich R.** and **Volk M.** (2018). Has machine translation achieved human parity? a case for document-level evaluation. In **Riloff E.**, **Chiang D.**, **Hockenmaier J.** and **Tsujii J.** (eds), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 4791–4796.

Li B., **Liu H.**, **Wang Z.**, **Jiang Y.**, **Xiao T.**, **Zhu J.**, **Liu T.** and **Li C.** (2020). Does multi-encoder help? a case study on context-aware neural machine translation. In **Jurafsky D.**, **Chai J.**, **Schluter N.** and **Tetreault J.** (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 3512–3518.

Li H., **Liu L.**, **Huang G.** and **Shi S.** (2021). GWLAN: General word-level AutocompletioN for computer-aided translation. In **Zong C.**, **Xia F.**, **Li W.** and **Navigli R.** (eds), *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)*, Online. Association for Computational Linguistics, pp. 4792–4802.

Li X., **Zhang J.** and **Zong C.** (2018). One sentence one model for neural machine translation. In **Calzolari N.**, **Choukri K.**, **Cieri C.**, **Declerck T.**, **Goggi S.**, **Hasida K.**, **Isahara H.**, **Maegaard B.**, **Mariani J.**, **Mazo H.**, **Moreno A.**, **Odijk J.**, **Piperidis S.** and **Tokunaga T.** (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Licht D., **Gao C.**, **Lam J.**, **Guzman F.**, **Diab M.** and **Koehn P.** (2022). Consistent human evaluation of machine translation across language pairs. In **Duh K.** and **Guzmán F.** (eds), *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Orlando, USA. Association for Machine Translation in the Americas, pp. 309–321.

Liu Y., **Gu J.**, **Goyal N.**, **Li X.**, **Edunov S.**, **Ghazvininejad M.**, **Lewis M.** and **Zettlemoyer L.** (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742.

Lo C.-k., **Larkin S.** and **Knowles R.** (2023). Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 776–799.

Lommel A., **Uszkoreit H.** and **Burchardt A.** (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, **12**, 455–463.

Lopes A., **Farajian M.A.**, **Bawden R.**, **Zhang M.** and **Martins A.F.T.** (2020). Document-level neural MT: A systematic comparison. In **Martins A.**, **Moniz H.**, **Fumega S.**, **Martins B.**, **Batista F.**, **Coheur L.**, **Parra C.**, **Trancoso I.**, **Turchi M.**, **Bisazza A.**, **Moorkens J.**, **Guerberof A.**, **Nurminen M.**, **Marg L.** and **Forcada M.L.** (eds), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal. European Association for Machine Translation, pp. 225–234.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Luong T., **Sutskever I.**, **Le Q.**, **Vinyals O.** and **Zaremba W.** (2015). Addressing the rare word problem in neural machine translation. In **Zong C.** and **Strube M.** (eds), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China. Association for Computational Linguistics, pp. 11–19.

Macé V. and **Servan C.** (2019). Using whole document context in neural machine translation. In **Niehues J.**, **Cattoni R.**, **Stüker S.**, **Negri M.**, **Turchi M.**, **Ha T.-L.**, **Salesky E.**, **Sanabria R.**, **Barrault L.**, **Specia L.** and **Federico M.** (eds), *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Manakhimova S., **Avramidis E.**, **Macketanz V.**, **Lapshinova-Koltunski E.**, **Bagdasarov S.** and **Möller S.** (2023). Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 224–245.

Mansimov E., **Melis G.** and **Yu L.** (2021). Capturing document context inside sentence-level neural machine translation models with self-training. In **Braud C.**, **Hardmeier C.**, **Li J.J.**, **Louis A.**, **Strube M.** and **Zeldes A.** (eds), *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics, pp. 143–153.

Marchisio K., **Guo J.**, **Lai C.-I.** and **Koehn P.** (2019). Controlling the reading level of machine translation output. In **Forcada M.**, **Way A.**, **Haddow B.** and **Sennrich R.** (eds), *Proceedings of Machine Translation Summit XVII: Research Track*, Dublin, Ireland. European Association for Machine Translation, pp. 193–203.

Marrese-Taylor E., **Wang P.C.** and **Matsuo Y.** (2023). Towards better evaluation for formality-controlled English-Japanese machine translation. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 551–560.

Maruf S. and **Haffari G.** (2018). Document context neural machine translation with memory networks. In **Gurevych I.** and **Miyao Y.** (eds), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 1275–1284.

Maruf S., **Martins A.F.T.** and **Haffari G.** (2019). Selective attention for context-aware neural machine translation. In **Burstein J.**, **Doran C.** and **Solorio T.** (eds), *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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 3092–3102.

Maruf S., **Saleh F.** and **Haffari G.** (2021). A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys* **54**(2).

McDonald J.C., **Wolfe R.**, **Efthimiou E.**, **Fontinea E.**, **Picron F.**, **Van Landuyt D.**, **Sioen T.**, **Braffort A.**, **Filhol M.**, **Ebling S.**, **Hanke T.** and **Krausneker V.** (2021). The myth of signing avatars. In **Shterionov D.** (ed.), *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, Virtual. Association for Machine Translation in the Americas, pp. 33–42.

Mehandru N., **Agrawal S.**, **Xiao Y.**, **Gao G.**, **Khoong E.**, **Carpuat M.** and **Salehi N.** (2023). Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In **Bouamor H.**, **Pino J.** and **Bali K.** (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, pp. 11633–11647.

Melby A. and **Foster C.** (2010). Context in translation: Definition, access and teamwork. *The International Journal for Translation & Interpreting Research*, **2**.

Miculicich L., **Ram D.**, **Pappas N.** and **Henderson J.** (2018). Document-level neural machine translation with hierarchical attention networks. In **Riloff E.**, **Chiang D.**, **Hockenmaier J.** and **Tsujii J.** (eds), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 2947–2954.

Moslem Y., **Romani G.**, **Molaei M.**, **Kelleher J.D.**, **Haque R.** and **Way A.** (2023). Domain terminology integration into machine translation: Leveraging large language models. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 902–911.

Müller M., **Rios A.**, **Voita E.** and **Sennrich R.** (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Specia L.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics, pp. 61–72.

Nguyen T. and **Chiang D.** (2018). Improving lexical choice in neural machine translation. In **Walker M.**, **Ji H.** and **Stent A.** (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 334–343.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Niu X. and **Carpuat M.** (2020). Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **34**, pp. 8568–8575.

Niu X., **Rao S.** and **Carpuat M.** (2018). Multi-task neural models for translating between styles within and across languages. In **Bender E.M.**, **Derczynski L.** and **Isabelle P.** (eds), *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 1008–1021.

Pawar S., **Tonmoy S.M.T.I.**, **Zaman S.M.M.**, **Jain V.**, **Chadha A.** and **Das A.** (2024). The what, why, and how of context length extension techniques in large language models – A detailed survey.

Petrick F., **Herold C.**, **Petrushkov P.**, **Khadivi S.** and **Ney H.** (2023). Document-level language models for machine translation. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 375–391.

Pham M., **Crego J.M.** and **Yvon F.** (2021). Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, **9**, 17–35.

Popel M., **Macháček D.**, **Auersperger M.**, **Bojar O.** and **Pecina P.** (2019). English-Czech systems in WMT19: Document-level transformer. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy. Association for Computational Linguistics, pp. 342–348.

Popescu-Belis A. (2019). Context in neural machine translation: A review of models and evaluations.

Post M., **Ding S.**, **Martindale M.** and **Wu W.** (2019). An exploration of placeholding in neural machine translation. In **Forcada M.**, **Way A.**, **Haddow B.** and **Sennrich R.** (eds), *Proceedings of Machine Translation Summit XVII: Research Track*, Dublin, Ireland. European Association for Machine Translation, pp. 182–192.

Post M. and **Junczys-Dowmunt M.** (2023). Escaping the sentence-level paradigm in machine translation. CoRR, abs/2304.12959.

Post M. and **Vilar D.** (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In **Walker M.**, **Ji H.** and **Stent A.** (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1314–1324.

Raunak V., **Kocmi T.** and **Post M.** (2023a). Evaluating metrics for document-context evaluation in machine translation. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 812–814.

Raunak V., **Menezes A.** and **Awadalla H.** 2023b. Dissecting in-context learning of translations in GPT-3. In **Bouamor H.**, **Pino J.** and **Bali K.** (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics, pp. 866–872.

Rei R., **Stewart C.**, **Farinha A.C.** and **Lavie A.** (2020). COMET: A neural framework for MT evaluation. In **Webber B.**, **Cohn T.**, **He Y.** and **Liu Y.** (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 2685–2702.

Rei R., **Treviso M.**, **Guerreiro N.M.**, **Zerva C.**, **Farinha A.C.**, **Maroti C.**, **C. de Souza J. G.**, **Glushkova T.**, **Alves D.**, **Coheur L.**, **Lavie A.** and **Martins A.F.T.** (2022). CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In **Koehn P.**, **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Jimeno Yepes A.**, **Kocmi T.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Popel M.**, **Turchi M.** and **Zampieri M.** (eds), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 634–645.

Rios Gonzales A., **Mascarell L.** and **Sennrich R.** (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In **Bojar O.**, **Buck C.**, **Chatterjee R.**, **Federmann C.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.** and **Kreutzer J.** (eds), *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 11–19.

Rippeth E., **Carpuat M.**, **Duh K.** and **Post M.** (2023). Improving word sense disambiguation in neural machine translation with salient document context.

Robinson N., **Ogayo P.**, **Mortensen D.R.** and **Neubig G.** (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 392–418.

Roelofsen F., **Esselink L.**, **Mende-Gillings S.**, **de Meulder M.**, **Sijm N.** and **Smeijers A.** (2021). Online evaluation of text-to-sign translation by deaf end users: Some methodological recommendations (short paper). In **Shterionov D.** (ed.), *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, Virtual. Association for Machine Translation in the Americas, pp. 82–87.

Rysová K., **Rysová M.**, **Musil T.**, **Poláková L.** and **Bojar O.** (2019). A test suite and manual evaluation of document-level NMT at WMT19. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy. Association for Computational Linguistics, pp. 455–463.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Saleh F., **Berard A.**, **Calapodescu I.** and **Besacier L.** (2019). Naver labs Europe's systems for the document-level generation and translation task at WNGT 2019. In **Birch A.**, **Finch A.**, **Hayashi H.**, **Konstas I.**, **Luong T.**, **Neubig G.**, **Oda Y.** and **Sudoh K.** (eds), *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Hong Kong. Association for Computational Linguistics, pp. 273–279.

Saunders D. (2022). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, **75**.

Saunders D., **Sallis R.** and **Byrne B.** (2020a). Neural machine translation doesn't translate gender coreference right unless you make it. In **Costa-jussà M.R.**, **Hardmeier C.**, **Radford W.** and **Webster K.** (eds), *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, Barcelona, Spain (Online). Association for Computational Linguistics, pp. 35–43.

Saunders D., **Stahlberg F.** and **Byrne B.** (2020b). Using context in neural machine translation training objectives. In **Jurafsky D.**, **Chai J.**, **Schluter N.** and **Tetreault J.** (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 7764–7770.

Schmidtke D. (2016). Large scale machine translation publishing, with acceptable quality, for Microsoft Support content. In *AMTA 2016 Workshop on Interacting with Machine Translation (iMT 2016)*, Austin, TX, USA. The Association for Machine Translation in the Americas.

Sennrich R., **Haddow B.** and **Birch A.** (2016a). Controlling politeness in neural machine translation via side constraints. In **Knight K.**, **Nenkova A.** and **Rambow O.** (eds), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics, pp. 35–40.

Sennrich R., **Haddow B.** and **Birch A.** (2016b). Improving neural machine translation models with monolingual data. In **Erk K.** and **Smith N.A.** (eds), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 86–96.

Smith K.S. (2017). On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 110–121.

Song K., **Zhang Y.**, **Yu H.**, **Luo W.**, **Wang K.** and **Zhang M.** (2019). Code-switching for enhancing NMT with pre-specified translation. In **Burstein J.**, **Doran C.** and **Solorio T.** (eds), *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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 449–459.

Specia L., **Frank S.**, **Sima'an K.** and **Elliott D.** (2016). A shared task on multimodal machine translation and crosslingual image description. In **Bojar O.**, **Buck C.**, **Chatterjee R.**, **Federmann C.**, **Guillou L.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Névéol A.**, **Neves M.**, **Pecina P.**, **Popel M.**, **Koehn P.**, **Monz C.**, **Negri M.**, **Post M.**, **Specia L.**, **Verspoor K.**, **Tiedemann J.** and **Turchi M.** (eds), *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany. Association for Computational Linguistics, pp. 543–553.

Stojanovski D. and **Fraser A.** (2021). Addressing zero-resource domains using document-level context in neural machine translation. In **Ben-David E.**, **Cohen S.**, **McDonald R.**, **Plank B.**, **Reichart R.**, **Rotman G.** and **Ziser Y.** (eds), *Proceedings of the Second Workshop on Domain Adaptation for NLP*, Kyiv, Ukraine. Association for Computational Linguistics, pp. 80–93.

Sugiyama A. and **Yoshinaga N.** (2021). Context-aware decoder for neural machine translation using a target-side document-level language model. In **Toutanova K.**, **Rumshisky A.**, **Zettlemoyer L.**, **Hakkani-Tur D.**, **Beltagy I.**, **Bethard S.**, **Cotterell R.**, **Chakraborty T.** and **Zhou Y.** (eds), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics, pp. 5781–5791.

Sun Z., **Wang M.**, **Zhou H.**, **Zhao C.**, **Huang S.**, **Chen J.** and **Li L.** (2022). Rethinking document-level neural machine translation. In **Muresan S.**, **Nakov P.** and **Villavicencio A.** (eds), *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland. Association for Computational Linguistics, pp. 3537–3548.

Susanto R.H., **Chollampatt S.** and **Tan L.** (2020). Lexically constrained neural machine translation with Levenshtein transformer. In **Jurafsky D.**, **Chai J.**, **Schluter N.** and **Tetreault J.** (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 3536–3543.

Sutskever I., **Vinyals O.** and **Le Q.V.** (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*, Cambridge, MA, USA. MIT Press, pp. 3104–3112.

Tars S. and **Fishel M.** (2018). Multi-domain neural machine translation. In **Pérez-Ortiz J.A.**, **Sánchez-Martínez F.**, **Esplà-Gomis M.**, **Popović M.**, **Rico C.**, **Martins A.**, **Van den Bogaert J.** and **Forcada M.L.** (eds), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alicante, Spain, pp. 279–288.

Thai K., **Karpinska M.**, **Krishna K.**, **Ray B.**, **Inghilleri M.**, **Wieting J.** and **Iyyer M.** (2022). Exploring document-level literary machine translation with parallel paragraphs from world literature. In **Goldberg Y.**, **Kozareva Z.** and **Zhang Y.** (eds), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 9882–9902.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Thompson B. and **Post M.** (2020). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In **Webber B.**, **Cohn T.**, **He Y.** and **Liu Y.** (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 90–121.

Tiedemann J. and **Scherrer Y.** (2017). Neural machine translation with extended context. In **Webber B.**, **Popescu-Belis A.** and **Tiedemann J.** (eds), *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 82–92.

Toral A., **Castilho S.**, **Hu K.** and **Way A.** (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Specia L.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics, pp. 113–123.

Torregrosa D., **Pasricha N.**, **Masoud M.**, **Chakravarthi B.R.**, **Alonso J.**, **Casas N.** and **Arcan M.** (2019). Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In **Forcada M.**, **Way A.**, **Tinsley J.**, **Shterionov D.**, **Rico C.** and **Gaspari F.** (eds), *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, Dublin, Ireland. European Association for Machine Translation, pp. 125–133.

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser L.** and **Polosukhin I.** (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, pp. 5998–6008.

Vernikos G., **Thompson B.**, **Mathur P.** and **Federico M.** (2022). Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In **Koehn P.**, **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Jimeno Yepes A.**, **Kocmi T.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Popel M.**, **Turchi M.** and **Zampieri M.** (eds), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 118–128.

Vincent S., **Flynn R.** and **Scarton C.** (2023). MTCue: Learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. In **Rogers A.**, **Boyd-Graber J.** and **Okazaki N.** (eds), *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics, pp. 8210–8226.

Vincent S.T., **Barrault L.** and **Scarton C.** (2022). Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation. In **Moniz H.**, **Macken L.**, **Rufener A.**, **Barrault L.**, **Costa-jussà M.R.**, **Declercq C.**, **Koponen M.**, **Kemp E.**, **Pilos S.**, **Forcada M.L.**, **Scarton C.**, **Van den Bogaert J.**, **Daems J.**, **Tezcan A.**, **Vanroy B.** and **Fonteyne M.** (eds), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, Ghent, Belgium. European Association for Machine Translation, pp. 121–130.

Voita E., **Sennrich R.** and **Titov I.** (2019a). Context-aware monolingual repair for neural machine translation. In **Inui K.**, **Jiang J.**, **Ng V.** and **Wan X.** (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 877–886.

Voita E., **Sennrich R.** and **Titov I.** (2019b). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In **Korhonen A.**, **Traum D.** and **Màrquez L.** (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 1198–1212.

Voita E., **Serdyukov P.**, **Sennrich R.** and **Titov I.** (2018). Context-aware neural machine translation learns anaphora resolution. In **Gurevych I.** and **Miyao Y.** (eds), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 1264–1274.

Vojtěchová T., **Novák M.**, **Klouček M.** and **Bojar O.** (2019). SAO WMT19 test suite: Machine translation of audit reports. In **Bojar O.**, **Chatterjee R.**, **Federmann C.**, **Fishel M.**, **Graham Y.**, **Haddow B.**, **Huck M.**, **Yepes A.J.**, **Koehn P.**, **Martins A.**, **Monz C.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Post M.**, **Turchi M.** and **Verspoor K.** (eds), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy. Association for Computational Linguistics, pp. 481–493.

Wadhwa M., **Chen J.**, **Li J.J.** and **Durrett G.** (2023). Using natural language explanations to rescale human judgments.

Wang L. (2019). Discourse-aware neural machine translation. PhD thesis, Ph. D. thesis, Dublin City University, Dublin, Ireland.

Wang L., **Lyu C.**, **Ji T.**, **Zhang Z.**, **Yu D.**, **Shi S.** and **Tu Z.** (2023). Document-level machine translation with large language models. In **Bouamor H.**, **Pino J.** and **Bali K.** (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, pp. 16646–16661.

Wang L., **Tu Z.**, **Way A.** and **Liu Q.** (2017). Exploiting cross-sentence context for neural machine translation. In **Palmer M.**, **Hwa R.** and **Riedel S.** (eds), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 2826–2831.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Wicks R. and **Post M.** (2022). Does sentence segmentation matter for machine translation? In **Koehn P.**, **Barrault L.**, **Bojar O.**, **Bougares F.**, **Chatterjee R.**, **Costa-jussà M.R.**, **Federmann C.**, **Fishel M.**, **Fraser A.**, **Freitag M.**, **Graham Y.**, **Grundkiewicz R.**, **Guzman P.**, **Haddow B.**, **Huck M.**, **Jimeno Yepes A.**, **Kocmi T.**, **Martins A.**, **Morishita M.**, **Monz C.**, **Nagata M.**, **Nakazawa T.**, **Negri M.**, **Névéol A.**, **Neves M.**, **Popel M.**, **Turchi M.** and **Zampieri M.** (eds), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 843–854.

Wicks R. and **Post M.** (2023). Identifying context-dependent translations for evaluation set production. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 452–467.

Wong B.T.M. and **Kit C.** (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In **Tsujii J.**, **Henderson J.** and **Pașca M.** (eds), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea. Association for Computational Linguistics, pp. 1060–1068.

Wong K., **Maruf S.** and **Haffari G.** (2020). Contextual neural machine translation improves translation of cataphoric pronouns. In **Jurafsky D.**, **Chai J.**, **Schluter N.** and **Tetreault J.** (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 5971–5978.

Wu Y. and **Hu G.** (2023). Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 166–169.

Wuebker J., **Green S.**, **DeNero J.**, **Hasan S.** and **Luong M.-T.** (2016). Models and inference for prefix-constrained machine translation. In **Erk K.** and **Smith N.A.** (eds), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 66–75.

Xiong H., **He Z.**, **Wu H.** and **Wang H.** (2019). Modeling coherence for discourse neural machine translation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19*. AAAI Press.

Xu J., **Crego J.** and **Senellart J.** (2019). Lexical micro-adaptation for neural machine translation. In **Niehues J.**, **Cattoni R.**, **Stüker S.**, **Negri M.**, **Turchi M.**, **Ha T.-L.**, **Salesky E.**, **Sanabria R.**, **Barrault L.**, **Specia L.** and **Federico M.** (eds), *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Yang Z., **Zhang J.**, **Meng F.**, **Gu S.**, **Feng Y.** and **Zhou J.** (2019). Enhancing context modeling with a query-guided capsule network for document-level translation. In **Inui K.**, **Jiang J.**, **Ng V.** and **Wan X.** (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 1527–1537.

Yin K., **Fernandes P.**, **Pruthi D.**, **Chaudhary A.**, **Martins A.F.T.** and **Neubig G.** (2021). Do context-aware translation models pay the right attention? In **Zong C.**, **Xia F.**, **Li W.** and **Navigli R.** (eds), *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)*, Online. Association for Computational Linguistics, pp. 788–801.

Yu L., **Sartran L.**, **Stokowiec W.**, **Ling W.**, **Kong L.**, **Blunsom P.** and **Dyer C.** (2020). Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, **8**, 346–360.

Yuan W., **Neubig G.** and **Liu P.** (2021). Bartscore: Evaluating generated text as text generation. In **Ranzato M.**, **Beygelzimer A.**, **Dauphin Y.**, **Liang P.** and **Vaughan J.W.** (eds), *Advances in Neural Information Processing Systems*, vol. **34**. Curran Associates, Inc., pp. 27263–27277

Yvon F. and **Rauf S.A.** (2020). Utilisation de ressources lexicales et terminologiques en traduction neuronale. Technical report, LIMSI-CNRS.

Zhang B., **Bapna A.**, **Johnson M.**, **Dabirmoghaddam A.**, **Arivazhagan N.** and **Firat O.** (2022). Multilingual document-level translation enables zero-shot transfer from sentences to documents. In **Muresan S.**, **Nakov P.** and **Villavicencio A.** (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics, pp. 4176–4192.

Zhang T., **Kishore V.**, **Wu F.**, **Weinberger K.Q.** and **Artzi Y.** (2020). Bertscore: Evaluating text generation with Bert. In *International Conference on Learning Representations*.

Zhang X., **Rajabi N.**, **Duh K.** and **Koehn P.** (2023). Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In **Koehn P.**, **Haddow B.**, **Kocmi T.** and **Monz C.** (eds), *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics, pp. 468–481.

Zhao W., **Strube M.** and **Eger S.** (2023). DiscoScore: Evaluating text generation with BERT and discourse coherence. In **Vlachos A.** and **Augenstein I.** (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics, pp. 3865–3883.

https://doi.org/10.1017/nlp.2024.7 Published online by Cambridge University Press

Zheng Z., **Yue X.**, **Huang S.**, **Chen J.** and **Birch A.** (2021). Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20*.

Zouhar V., **Popel M.**, **Bojar O.** and **Tamchyna A.** (2021). Neural machine translation quality and post-editing performance. In Moens M.-F., Huang, X., Specia, L. and Yih, S. W.-t. (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 10204–10214.

Cite this article: Castilho S and Knowles R (2025). A survey of context in neural machine translation and its evaluation. *Natural Language Processing* **31**, 986–1016. https://doi.org/10.1017/nlp.2024.7