Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level

Nicolas Dahan, Rachel Bawden, François Yvon

;

Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level

Nicolas Dahan[1,2], **Rachel Bawden**[1] **and François Yvon**[2]

[1] **Inria and** [2] **Sorbonne Université and CNRS**

October 2024

MaTOS — Livrable D4-4.1

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

Survey of Automatic Metrics for Evaluating Machine Translation at the Document-Level

Nicolas Dahan, Rachel Bawden and François Yvon

Abstract

This report presents a survey of document-level automatic metrics for machine translation (MT), addressing the need for sophisticated evaluation methods that extend beyond sentence-level assessments. Traditional metrics, which evaluate translations on a sentence-by-sentence basis, often fail to capture the complexities of discourse phenomena, leading to gaps in assessing coherence, cohesion, and cross-sentence dependencies. The report starts by introducing the terminology and notation relevant to document-level MT evaluation. It then describes the linguistic phenomena that are crucial at the document level, related for example to lexical and grammatical cohesion, and overall text coherence, which pose significant challenges for MT systems. Following this, we explore human evaluation protocols targeting document-level translation, discussing the methodologies used to judge translation quality in a more holistic manner. Studying human judgments is necessary, as automatic metrics often aim at reproducing them. We also examine the various test sets that have been developed to support the evaluation of document-level MT. The core of the survey focuses on automatic evaluation metrics designed for document-level translation. These metrics aim to provide a more accurate representation of translation quality by considering the broader context and long-range dependencies within a text, offering a more comprehensive assessment than sentence-level metrics. The report concludes with an overview of the current trends in document-level MT evaluation, summarizing key challenges and identifying areas for future research. It emphasizes the need for the development of context-aware metrics and the importance of creating standardized, document-level test sets to advance MT evaluation.

Résumé

Ce rapport présente un état de l'art des métriques automatiques pour évaluer la traduction automatique (TA) au niveau du document, répondant au besoin de méthodes d'évaluation sophistiquées qui dépassent les évaluations au niveau de la phrase. Les métriques traditionnelles, qui évaluent les traductions phrase par phrase, échouent le plus souvent à saisir les complexités des phénomènes discursifs, ce qui entraîne des lacunes dans l'évaluation de la cohérence, de la cohésion et des dépendances interphrastiques. Dans ce rapport, nous commençons par introduire la terminologie et les notation pertinentes pour l'évaluation de la TA au niveau du document. Nous étudions ensuite sur les principaux phénomènes linguistiques qui impliquant des dépendance niveau du document, liés par exemple à la cohésion lexicale et grammaticale, ainsi qu'à la nécessité de maintenir la cohérence globale du texte, deux problèmes qui posent des défis importants pour les systèmes de TA. Par la suite, nous présentons les méthodologies d'évaluation humaine ciblant la traduction au niveau du document en discutant des protocoles utilisés pour juger la qualité de la traduction de manière holistique. Ce détour par les évaluations humaines est nécessaire, dans la mesure où les évaluations automatiques cherchent souvent à reproduire ces jugements. Nous examinons également les différents jeux de tests qui ont été développés pour soutenir l'évaluation de la TA au niveau du document. Au cœur de ce rapport, nous nous concentrons sur les métriques d'évaluation automatique conçues pour la traduction au niveau du document. Ces métriques visent à fournir une représentation plus précise de la qualité de la traduction en tenant compte du contexte plus large et des dépendances à long terme au sein d'un texte, offrant une évaluation plus complète que les métriques évaluant des phrases. Nous terminons par un aperçu des tendances actuelles dans l'évaluation de la TA au niveau du document, résumant les principaux défis et identifiant des thématiques de recherche émergentes. Nous soulignons la nécessité de développer des métriques sensibles au contexte et l'importance de créer des jeux de tests standardisés au niveau du document pour faire progresser l'évaluation de la TA.

Contents

1 Introduction **4**

2 Terminology **4**

3 Document-level linguistic phenomena **6**
3.1 Coherence . . . 6
3.2 Cohesion . . . 8

4 Human Evaluation of Document-Level MT **10**
4.1 Context-aware and document-aware human evaluation . . . 10
4.2 Document-aware meta-evaluation . . . 12

5 Global Automatic Evaluation **13**
5.1 Document-level test sets . . . 14
5.1.1 Parallel documents . . . 14
5.1.2 Test sets with error annotations . . . 15
5.2 Global automatic evaluation metrics . . . 17
5.2.1 Surface metrics . . . 18
5.2.2 Embedding-based metrics . . . 19
5.2.3 Fine-tuned metrics . . . 22
5.2.4 Zero-shot prompting metrics . . . 25

6 Targeted Evaluation of Specific Phenomena **26**
6.1 Test suites for document-level phenomena . . . 26
6.1.1 Manually-designed test suites . . . 27
6.1.2 Constrastive test sets . . . 28
6.2 Automatic metrics for evaluating specific linguistic phenomena . . . 30

7 Meta-evaluation **33**
7.1 Computing correlations with human judgments . . . 33
7.2 Evaluating metrics using challenge sets . . . 34

8 Conclusion **35**

Bibliography **36**

1 Introduction

As Machine Translation (MT) research advances toward document-level systems capable of translating text beyond individual sentences, evaluating the effectiveness of these systems remains a significant challenge. Despite these advancements, it is still common practice to use sentence-level automatic metrics, often applied to entire documents or multi-segment chunks, to assess translation quality. However, such metrics, originally designed for sentence-level translation, may fail to accurately capture the progress of document-level translation. This inadequacy has led the MT community to recognize the need to incorporate context, and document-level phenomena, and to shift beyond sentence-level evaluation methodologies [Sim Smith, 2017, Maruf et al., 2021].

Several surveys have addressed aspects of document-level MT but did not concentrate exclusively on evaluation. For instance, [Maruf et al., 2021, Abdul Rauf and Yvon, 2020, Popescu-Belis, 2019, Peng et al., 2024a] focus primarily on translation methodologies and the role of context, allocating limited attention to evaluation techniques and all fields related to automatic evaluations, such as the dichotomy between different automatic metrics or the meta-evaluation. Other works, such as [Castilho and Knowles, 2024], offer a survey of context in neural MT and its evaluation but do not focus specifically on document-level evaluation, which encompasses context-aware evaluation. Surveys by Castilho et al. [2018] and Chatzikoumi [2020] explore how to evaluate MT and approaches to human and MT quality assessment, respectively, but they address evaluation at a general level rather than concentrating on document-level specifics. Additionally, Gehrmann et al. [2023] survey obstacles in evaluation practices for text generation, highlighting issues that are relevant, but not exclusive, to document-level MT.

In contrast, this report exclusively focuses on the evaluation of translation at the document level, aiming to fill a notable gap in the literature. By concentrating on document-level automatic metrics for MT, we delve deeper into methods that extend beyond sentence-level assessments. Our survey addresses the need for sophisticated evaluation methods that can accurately capture the complexities of discourse phenomena inherent in longer texts. Our angle is particularly timely due to recent developments in the field. For example, the introduction of test sets with paragraphs as segments in the WMT Metrics Shared Task only dates back to the 2023 edition [Freitag et al., 2023], reflecting a recent and growing recognition within the community of the limitations of traditional, sentence-level metrics when applied to document-level translations. There has been a surge in enthusiasm for developing new evaluation methods that are context-aware and capable of assessing coherence and cohesion across sentences.

This report is organized as follows: in Section 2, we introduce the terms that are relevant to document-level MT evaluation. We then describe the linguistic phenomena (Section 3) that are crucial at the document level, related for example to lexical and grammatical cohesion, and overall text coherence, which pose significant challenges for MT systems. We continue with an exploration of human evaluation techniques targeting document-level translation, discussing the methodologies used to judge translation quality more holistically (Section 4). Next, we also examine the various test sets that have been developed to support the evaluation of document-level MT. The core of the survey is in Section 5, which focuses on automatic evaluation metrics designed for document-level translation. These metrics aim to provide a more accurate representation of translation quality by considering the broader context and long-range dependencies within a text, offering a more comprehensive assessment than sentence-level metrics. In Section 6, complementing global metrics, we look at methods for evaluating various specific document-level phenomena. The report concludes with an overview of the current trends in document-level MT evaluation, summarizing key challenges and identifying areas for future research. It emphasizes the need for the development of document-level metrics and the importance of creating standardized, document-level test sets to advance MT evaluation.

2 Terminology

In this section, we define the terms *segment*, *sentence*, *context*, *paragraph*, *document*, and *corpus*. Although there is a general understanding of these terms within the translation community, we find it useful to provide precise definitions that will be used throughout this report. A *segment* is defined as the individual input fed into an MT system or an scoring procedure (e.g., a benchmark usually comprises a set of segments to be translated/scored). The computation of quality assessment scores typically involve two or three input segments, corresponding to the output translation hypothesis, associated to either the target gold translation (for reference-based evaluations) or the original source segment (for referenceless evaluation), sometimes with both. The nature of segments depends

on how the task is defined; traditionally, a segment corresponds to a sentence, in which case we speak of sentence-level MT/evaluation. However, a segment may also comprise multiple sentences in the case of contextual or document-level MT/evaluation.

A *sentence* in MT is typically defined as the smallest unit of translation that contains a complete thought, typically bounded by punctuation marks such as periods, question marks, or exclamation points [Newmark, 1988]. Sentence-level translation has been the traditional focus in MT for several reasons, both theoretical and practical. In many cases, the sentence constitutes a reasonable basic unit for translation, which can be correctly translated without the need for additional context), and limiting the scope of a segment to a sentence is computationally beneficial for most MT systems [Koehn, 2010, Karpinska and Iyyer, 2023]. It also allows for straightforward alignment (between source, target, and reference sentences), useful both during model training and evaluation. However, processing isolated sentences makes it impossible to access the contextual information that may be necessary for both translation and evaluation to make optimal judgments. This happens notably when handling certain discourse phenomena such as the resolution of anaphoric references, formality, consistency, cohesion, and coherence issues [Bawden et al., 2018b, Voita et al., 2019, Stojanovski and Fraser, 2019]. Such context-dependent phenomena are reviewed in Section 3 of this report.

A *paragraph* consists of a series of sentences that are closely related and contribute to a single theme or idea [Scott et al., 1909]. A typical paragraph can be as short as a couple of sentences or contain several dozen sentences. In recent efforts towards document-level MT, paragraphs have been a common choice, as they represent a reasonable trade-off between going beyond the sentence level while ensuring that the length of the input that is translated is not too long. Recent advancements in paragraph-level MT include the work of Miculicich et al. [2018], who focus on using hierarchical attention for paragraph translation, and Karpinska and Iyyer [2023], Wang et al. [2023], who explore the use of Large Language Models (LLMs) for the same purpose. Additionally, Liu and Zhang [2020] have contributed by collecting a large number of existing document-level corpora and tools, further facilitating research in this area.

The definition of what constitutes a *document* is more difficult to determine precisely, as it is highly dependent on the type of text and on the formulation of the task in the context of document-level MT. For example, a document could typically contain multiple paragraphs but can range from short articles to entire books (with a more complex internal structure). However, the length of a document is determined in practice by those who formulate the task, with the idea that a document constitutes a set of sentences that forms a coherent whole, optionally structured into sections/paragraphs. For example, in 2023 WMT shared tasks [Kocmi et al., 2023], documents did not always constitute complete articles or texts, but samples of texts that were considered to constitute a coherent set of sentences that could be considered together. Whatever the length or structure of a document, working at the document level involves considering inter-sentential and inter-paragraph dependencies to preserve the overall structure, tone, and context of the entire text [Thai et al., 2022].

In this report, the term *context* refers to the span of sentences or words surrounding a particular segment. This concept is essential for what is known as "context-aware translation". In this scenario, the segment of focus is usually an individual sentence, augmented by the surrounding linguistic context. This context can vary in size, encompassing linguistic elements directly related to the sentence (such as the sentences that precede or follow). Additionally, in the broader scope of translation studies, the context can extend to related documents and resources, as well as real-world extra-linguistic information such as the author's background, the historical period, the general topic, the intended readership, or the overarching theme of the text, all of which can influence both translation and evaluation [Melby and Foster, 2010, Hardmeier, 2012].

It is worth noting that in the MT literature, what is referred to as document-level MT is often in reality context-aware MT [Vernikos et al., 2022]. Yet, context-aware MT differs from paragraph-level MT and document-level MT in fundamental ways: while paragraph- and document-level approaches aim to translate the text holistically (a whole paragraph/document), context-aware MT often considers the sentence as the translation unit, with an associated context providing additional disambiguation information [Deutsch et al., 2023a].

Finally, the term *corpus* refers to a set of segments, which is composed of an arbitrary conjunction of sentences, paragraphs, or documents (depending on how the corpus is designed), used for training, evaluation, or analysis. A corpus can be seen as a random set of representative samples, over which metric scores will be averaged.

3 Document-level linguistic phenomena

Traditionally, sentence-level MT operates under the assumption that sentences are independent units [Hardmeier, 2012, Tan et al., 2020]. However, this assumption often fails to hold, as the meaning and interpretation of a sentence can depend on its context within a paragraph or the entire document. *Local context*, such as the immediate surrounding sentences, can have a disambiguating effect by clarifying ambiguous terms and by understanding sentences that rely heavily on nearby information. On the other hand, *Global context* encompasses the entire document, which is crucial for maintaining overall thematic consistency, understanding the author's intent, and ensuring that the translation aligns with the document's purpose and tone. *Global context* affects decisions like making a consistent usage of terminology throughout a document as well as preserving overarching narratives or arguments.

Unlike sentence-level approaches, document-level MT explicitly rejects the premise of independence across sentences. Document-level MT is both a methodological approach—translating documents as a whole—and an objective, aiming to produce translations that consider inter-sentential dependencies [Miculicich et al., 2018, Zhang et al., 2018]. By integrating both local and global contextual information, document-level MT models strive to maintain *coherence* and *cohesion*, two interconnected concepts essential for effective communication. *Coherence* ensures that the document is understandable and logically structured, presenting a unified argument or narrative. *Cohesion*, on the other hand, supports coherence by providing the grammatical, lexical, and rhetorical links that bind sentences and paragraphs together. *Coherence* and *cohesion* are not merely stylistic choices but essential components of effective communication. *Coherence* allows the reader to grasp the overall meaning and purpose of the text, while *cohesion* makes the reading experience fluent and structured.

In addition to coherence and cohesion, document-level evaluation should also consider the stylistic, rhetorical, and pragmatic goals of the source text. For example, if the source text employs a colloquial style, this style should ideally be preserved in the translation to maintain the original tone and communicative intent. Similarly, rhetorical devices and pragmatic nuances such as politeness markers, irony, or register shifts should be appropriately conveyed in the target language to reflect the source text's intent and impact. However, since these criteria have not been the subject of extensive research, we have chosen to focus primarily on coherence and cohesion and their resolution in document-level MT.

In the context of MT, we assume that the source text is already coherent and cohesive; MT should therefore ensure that these qualities are preserved in the target text [Smith et al., 2016]. This, however, is a challenging task, as structural and linguistic differences between languages can introduce ambiguities and complexities that make maintaining coherence and cohesion difficult. Recognizing and replicating these interlinked aspects are crucial for producing translations that are not only accurate on a per-sentence basis but also contextually appropriate and interconnected across the entire document.

3.1 Coherence

Coherence in texts is essential for ensuring they are semantically meaningful and easy to follow. In part, coherence is a matter of personal perspective, and individual readers may perceive varying levels of coherence based on their external knowledge of the world [Lonsdale, 1996]. This concept is crucial in MT where the challenge lies not only in translating words and phrases accurately but also in preserving the logical flow and thematic unity of the source input throughout the target document. Importantly, coherence is not always overtly apparent; often, it is more noticeable in its absence, manifesting as a disjointed or confusing narrative [Vasconcellos, 1989].

In MT, maintaining coherence across a document involves addressing several specific types of potential incoherence that may be introduced during translation:

- Logical Progression: Ensuring the translated text follows the same logical sequence as the source text is crucial. For instance, a discussion on climate change should naturally progress from the introduction to evidence, impacts, and solutions as illustrated in Table 3. Any disruption in this order would make the text confusing and difficult to follow.

- Thematic Unity and contextual relevance: The translated text must maintain a consistent theme throughout. With MT, thematic disunity can arise due to the mistranslation of ambiguous source words. Improperly translating an ambiguous word could prompt a topical change and yield a divergence from the main argument, leading to a disjointed narrative.

- Semantic Consistency: Key terms and concepts should be consistently translated to maintain coherence. Inconsistencies in the automatic translation of terms by an MT system can disrupt the reader's understanding and the overall coherence of the text.

Many aspects and examples of coherence are more subjective and challenging to measure systematically. While important, these aspects fall outside the scope of our systematic analysis in this survey. Here, we concentrate on a small number of concrete coherence phenomena that can be systematically studied.

- The **resolution of ambiguity** is one of the most fundamental and salient challenges in MT related to coherence. It can arise because of the ambiguous nature of the source sentence (e.g. English *bank* 'side of a river' (French *rive*) or 'financial institute (French *banque*)) and/or because of semantic mismatches between source and target languages.The examples in Table 1 contrast erroneous context-free (sentence-level) translations with context-aware translations informed by a disambiguating context. Providing additional document context and/or including explicit word sense disambiguation modules (WSD) can be used to ensure that the appropriate word is used in context [Carpuat and Wu, 2007].

- Another significant challenge in MT coherence arises from issues related to **terminology**. Proper term consistency is vital in domains like scientific or technical translation, where textual data are often documents and where key terms must be translated consistently to maintain clarity and coherence. When an MT system fails to consistently translate these terms, it can lead to confusion for the reader and disrupt the overall flow of the text [Bawden et al., 2019]. Terminology issues can emerge due to the system's inability to correctly identify domain-specific terms. Domain-specific termbases or glossaries can help mitigate this issue, ensuring that the appropriate terminology is used consistently throughout the text [Hanoulle, 2017].

- **Deixis** refers to the use of words or phrases that point to a particular time ('then'), place ('here'), or person ('you') relative to the context of an utterance, with their meaning being context-dependent. In the example in Table 2, "Je vais là demain.", the spatial reference "là" can only be understood and correctly translated in light of the surrounding discourse (the previous sentence is enough with this example). MT systems must correctly handle such contexts to maintain coherence across documents. However, because different languages may use different deictic structures, errors often arise when translating pronouns, space or time references across languages. In addition, deixis elements are frequently be based on non-textual elements which can complicate their resolution.

English source	
EN	We're not very well set up for fishing and there aren't many fish.
EN	We should move to another **bank**.
Sentence-level translation	
FR	Nous ne sommes pas très bien équipés pour la pêche et il n'y a pas beaucoup de poissons.
FR	Nous devrions changer de **banque**. ✗
Context-aware translation	
FR	Nous ne sommes pas très bien installés pour pêcher et il n'y a pas beaucoup de poissons.
FR	Nous devrions changer de **rive**. ✓

Table 1: Comparison of sentence-level versus context-aware translation using DeepL (public version, consulted on the 1st August 2024).

English source	
FR	Nous sommes devant la bibliothèque.
FR	Je vais **là** demain.
Sentence-level translation	
EN	We are in front of the library.
EN	I am going **there** tomorrow. ✗
Context-aware translation	
EN	We are in front of the library.
EN	I am going **here** tomorrow. ✓

Table 2: Comparison between sentence-level and context-aware translations of an ambiguous French space deictic.

The various aspects that help to maintain coherence have not received the same level of attention when studied through the lens of MT. Ambiguity and its non-resolution, which can lead to incoherence, have been addressed by multiple studies [Weaver, 1952, Chen et al., 1999, Carpuat and Wu, 2007, Rios et al., 2018, Bawden et al., 2018b, Stahlberg and Kumar, 2022], and the same can be said, to a lesser extend, to issues related to terminology [Maia, 2005, Michon et al., 2020, ibn Alam et al., 2021, Semenov and Bojar, 2022]. By contrast, some other coherence phenomena have not been studied as extensively in MT research; this is for instance the case of the accurate translation of deictic elements [Voita et al., 2019].

Sentence	**Category**
Climate change represents one of the most significant challenges facing the planet today.	Introduction of the topic
Satellite images have documented a dramatic reduction in Arctic sea ice, indicating a warming planet.	Evidence of climate change
The agricultural sector is particularly vulnerable, experiencing reduced crop yields due to altered rainfall patterns and increased temperatures.	Impacts
To address these challenges, renewable energy sources such as wind and solar power must replace fossil fuels to reduce carbon emissions significantly.	Solutions

Table 3: An example of the coherent progression of a discussion on climate change.

3.2 Cohesion

Cohesion is a crucial surface property of a text, referring to the way textual units are grammatically or lexically linked together [Halliday and Hasan, 1976] and is a crucial aspect in document-level translation [Lei et al., 2022]. Compared to coherence, cohesion is easier to describe and recognize. Halliday and Hasan [1976] identify several types of cohesive devices, including reference, substitution/ellipsis, lexical cohesion, and conjunction. Additionally, other linguistic devices have been suggested to further bind a text together [Thornbury, 2005, Rimmer, 2006]: not all of these devices have been studied specifically for MT as they are either rare or difficult to model/evaluate.

The following set of devices play a key role in making a text cohesive:

- **Lexical cohesion**: This involves the use of vocabulary and word choice to create connections within the text. Examples include the repetition of words, the use of synonyms or near-synonyms, lexical chains, the inclusion of related words to build a semantic field and substitution. Table 4 compares lexically non-cohesive and cohesive translations of two consecutive Japanese sentences. The non-cohesive translation, carried out at the sentence level, translates the Japanese word “時計” (watch) as “clock” in the first sentence and “watch” in the second sentence, whereas the same translation should be used in both instances. In contrast, the cohesive translation, which goes beyond the sentence level, consistently translates the word as “watch” [Mishra et al., 2020].

- **Grammatical cohesion**: This encompasses the use of grammatical structures to link different parts of the text. Pronouns, conjunctions, tenses, substitution of clause elements and ellipses are common grammatical cohesive devices that help to maintain the flow of information and clarify relationships between ideas. In cases where grammatical cohesion is realized by

anaphoric pronouns, they refer back to their antecedents/referents (i.e. concepts previously introduced into the text) [Thornbury, 2005]. In Table 5, the correct French translation of the plural personal pronoun *they* "ils (m) or elles (f)" must agree with the gender of its coreferent *les filles* "the girls" in the previous sentence, a choice that can only be made correct if there is access to the previous sentence.

- **Rhetorical cohesion**: This refers to the use of rhetorical strategies to create a coherent and persuasive argument [Thornbury, 2005]. Techniques such as parallelism, contrast, and rhetorical questions can effectively link sections of the text and enhance its overall cohesiveness. However, there is as yet no research into the problems of translating rhetorical cohesion phenomena.

Japanese source	
JP	田中さん、よい時計をお持ちですね。
JP	ありがとう、この時計は祖父の形見なんです。
Sentence-level translation (incohesive translation)	
EN	You have a good clock, Mr. Tanaka.
EN	Thank you, this watch is a memento of my grandfather. ✗
Context-aware translation (cohesive translation)	
EN	You have a good watch, Mr. Tanaka.
EN	Thank you, this watch is a memento of my grandfather. ✓

Table 4: Comparison of sentence-level versus context-aware JP-EN translation of lexical cohesion device (direct repetition and ambiguity). Example from [Mishra et al., 2020].

English source	
EN	The town would burn down to the ground if we took the **girls** within a mile of that guy.
EN	Do you think **they**'ll go?
Sentence-level translation (non-cohesive translation)	
FR	La ville serait réduite en cendres si on amenait les **filles** à proximité de lui.
FR	Tu crois qu'**ils** voudront y aller? ✗
Context-aware translation (cohesive translation)	
FR	La ville serait réduite en cendres si on amenait les **filles** à proximité de lui.
FR	Tu crois qu'**elles** voudront y aller ? ✓

Table 5: Comparison of sentence versus context-aware EN–FR translation of pronouns. Example from [Lopes et al., 2020].[1]

Many cohesive devices, crucial for maintaining the fluency and readability of translated documents, have received attention in MT research. Lexical cohesion, including word repetition, synonymy, and lexical chains, has been extensively studied, with numerous works focusing on its role in ensuring consistency across translations [Wong and Kit, 2012, Xiong et al., 2013, Voita et al., 2019]. Grammatical cohesion, encompassing devices like conjunctions, ellipses, and reference resolution, has also garnered substantial attention, as evidenced by studies on the proper handling of conjunctions [Krein-Kühle, 2002, Ketabi and Jamalvand, 2012, Pan, 2014, Popović, 2019, Popović and Castilho, 2019], ellipses [Oommen and Reichstein, 1986, Voita et al., 2019, Mutal et al., 2020, Khullar, 2021], and anaphoric reference [Krein-Kühle, 2002, Voigt and Jurafsky, 2012]. However, rhetorical cohesion, which involves the use of rhetorical strategies such as parallelism and contrast to create a cohesive argument and plays a key role in linking larger textual units and enhancing the persuasiveness of a text, remains underexplored in MT studies as far as we know. The limited attention to rhetorical cohesion poses a challenge for ensuring that translations capture the full rhetorical effect of the source text, especially at the document level. These cohesive devices are particularly prominent when dealing with paragraph- or document-level translations, where resolution must extend beyond the sentence level. While traditional sentence-based approaches often miss these links, automatic techniques, whether applied locally or globally, are expected to increasingly incorporate such features to enhance the overall quality of document-level translations. This is especially crucial for capturing the linguistic subtleties needed to maintain consistency and coherence across larger textual units.

4 Human Evaluation of Document-Level MT

Before reviewing automatic evaluation metrics, it is first useful to discuss what constitutes the gold standard for evaluation, as this determines what automatic metrics are developed to imitate. The gold standard in terms of translation quality evaluation is human evaluation (i.e. collecting human judgments of the translation quality of MT outputs). We keep this discussion lightweight since a more in-depth survey of human evaluation strategies is available in a companion survey [Bénard et al., 2024].

Human protocols for evaluating MT are typically based on a comparison of system outputs with source texts, system outputs with reference translations, or a combination thereof [Bojar et al., 2016, Läubli et al., 2018]. In addition to these different comparison configurations (reference-based, source-based, source and reference-based), these protocols vary according to several key aspects:

1. the length of the segment that receives an individual human judgment (e.g., quality judgments can be expressed at the sentence, the paragraph, or the document level);

2. the amount and type of context provided to the evaluator when assessing a given segment (e.g., preceding source/target sentence(s), whole paragraph context, etc.)

3. the type of judgment provided, e.g., a single quality score for the segment, a single score for identified aspects of translation quality such as fidelity and fluency [White et al., 1994], a reranking compared to other system outputs corresponding to the same source sentences [White et al., 1994], or a fine-grained error analysis based, for instance, using Multidimensional Quality Metrics (MQM) [Lommel et al., 2013].

Whichever the protocol selected (conventions have changed over time), it is typical for individual judgments to be collected from experts or from crowdworkers [Iskender et al., 2020] and then to average these judgements over a corpus or a subsample of a corpus. In the remainder, we focus in human evaluations that can be summarized by a single numerical score; these represent, by large, the most common approach.

4.1 Context-aware and document-aware human evaluation

Traditionally, the evaluation of MT system outputs has been performed at the sentence level. This was, for instance, still the case for the WMT shared tasks[2] before 2019 [Barrault et al., 2018], with individual test set sentences presented in random order. Document-level and context-aware human evaluation of MT is increasingly recognized as essential for faithfully assessing translation quality, given that extra-sentential context can influence judgment choices [Toral et al., 2018, Barrault et al., 2019, Toral, 2020]. Evaluations that rely on single, isolated sentences can lead to misevaluations as the necessary context to evaluate context-dependent elements is not present. Providing evaluators with full document contexts has been shown to improve accuracy by enabling a better understanding of such ambiguities and agreements [Scarton et al., 2015, Toral et al., 2018, Castilho et al., 2020].

Although not all sentences require context to be correctly evaluated, this can greatly vary depending on the domains. A study by Castilho et al. [2020] showed that approximately 30% of sentences across different domains (reviews, subtitles, and literature)[3] had over 40% of annotator agreement on the necessity of inter-sentential context for accurate translation and evaluation.[4] From those sentences judged to require additional context, 23% needed more than two preceding sentences for proper evaluation [Castilho, 2021]. The sentences concerned typically presented challenges such as ambiguity, specific terminology, and gender agreement, and the study highlighted that certain error types related to cohesion and coherence are not always identifiable at the sentence level.

Manual evaluation strategies have evolved over time [Bojar et al., 2016], along with the realization that traditional (sentence-level) practices were not optimal for distinguishing the best systems [Graham et al., 2017, Mathur et al., 2020] and could skew the evaluation results [Läubli et al., 2018, Toral et al., 2018, Akhbardeh et al., 2021]. As mentioned before, WMT shared tasks have accordingly progressed from isolated sentence-level judgments to context-level and document-level ones.

[2] https://github.com/rbawden/Large-contrastive-pronoun-testset-EN-FR

[2] WMT is the Conference on Machine Translation, a specialized conference that organizes yearly translation shared tasks.

[3] In this study, 300 sentences were evaluated.

[4] However, some of these sentences were later discarded due to unrelated issues, such as problems with the English source or their position as the first sentence in a corpus. This was conducted to consider 95 sentences and not 107 (33%).

- The **Relative Ranking** (RR) method was used to evaluate sentence translations. Here, competing translations of the same sentence were ranked relative to each other. While this method enabled a direct comparison of multiple systems, it proved ill-suited to evaluate long segments (with or without their context).

- The **Direct Assessment** (DA) [Graham et al., 2016] method replaced the earlier relative ranking strategy. In this method, human evaluators score the quality of translated sentences on a continuous scale from 0 to 100, where 0 is the worst and 100 represents perfect translation quality. A major advantage of DA is its adaptability to the evaluation of long segments or the inclusion of additional context. DA configurations have evolved owing to the need to go beyond the sentence-level and to respond to the issues raised by Läubli et al. [2018], Toral et al. [2018]:

  - Segment-based DA with document context (+DC/-DC): Starting in 2019, evaluators rated individual segments[5] while considering surrounding sentences as context. However, they could not revisit previous segments during evaluation [Barrault et al., 2019].
  - Document-based DA (DR+DC): This method attempted to provide a single score for an entire document but faced statistical power issues and inconclusive ties [Graham et al., 2019, 2020b], leading to its exclusion in future evaluations.
  - Contextual Adaptations in DA [Barrault et al., 2020], DA evolved into Segment Rating with Full Document (SR+FD), allowing evaluators to review segment-level scores with full document access. This shift improved the reliability of context-based evaluations by giving annotators broader access to related content across languages with available context [Castilho et al., 2020, Läubli et al., 2020, Akhbardeh et al., 2021].
  - Document-level DA (DOC-DA) [Mathur et al., 2020] is the average score of DA from consecutive sentences. The motivation for this method was to take the broader context of evaluated sentences or paragraphs into account. Although this approach was discontinued after WMT20, the underlying method has been used to perform human evaluation at document-level [Kocmi et al., 2024] or to create or assess automatic global metrics [Gong et al., 2015, Deutsch et al., 2023b, Hendy et al., 2023, Raunak et al., 2024] (see Section 5.2).
  - DA with Scalar Quality Metric (DA+SQM) method [Kocmi et al., 2022] is a calibrated extension of DA, where evaluators score sentences on a scale from 0 to 100 and consult a seven-point labeled scale to ensure scoring consistency. Typically, evaluations using DA+SQM are performed at the sentence level, and paragraph-level scores are computed by averaging the scores of the sentences within that paragraph. Although it is feasible to evaluate entire paragraphs, doing so is more time-consuming, reduces the overall number of scores collected, and increases the cognitive effort required from evaluators, which can negatively impact inter-annotator agreement [Castilho, 2020, Kocmi et al., 2024]. To better account for contextual information, DA+SQM was implemented with document-level context, incorporating up to 10 preceding and following source sentences, and resulting in more consistent and insightful assessments.

- The **Error Span Analysis** method, as used in Multidimensional Quality Metrics (MQM) [Lommel et al., 2014], focuses on identifying specific errors within a translation and categorizing them into predefined dimensions (e.g., *Accuracy*, *Fluency*, *Terminology*, *Style*, and *Locale*). Human evaluators highlight, without using reference translations, error spans within the translation and assign severity levels to each error (e.g., *Minor*, *Major*, *Neutral*), providing a detailed and structured evaluation of translation quality. In their MQM variant, Freitag et al. [2021a] introduces additional error categories and customizes error severity weights. For example, *Non-translation* errors, used to tag entire segments that are too garbled for detailed error identification, carry a significantly higher weight, equivalent to multiple *Major* errors, to reflect the severity of such issues. Conversely, minor punctuation errors are assigned lower weights to prevent trivial issues from disproportionately affecting the overall score. Unlike the Likert-style scheme commonly used in many studies [Graham et al., 2020a], MQM raters do not directly assign scalar scores which are calculated after the annotation step by applying a weighting scheme that considers both the severity and the category of errors. This fine-tuned weighting system makes MQM particularly effective for evaluating broad-coverage

[5] As the context only consisted of a couple of previous sentences, this evaluation should better be termed as a context-aware evaluation.

MT systems and contexts where careful calibration of error severity is critical. The MQM framework explicitly accounts for the fact that segments can consist of several sentences and instructs evaluators to pay particular attention to document context when annotating, ensuring that errors are identified and categorized with an understanding of the surrounding content. In the setup of evaluating translated document, Freitag et al. [2021a] also seeks to temper the impact of long segments by imposing a maximum of five errors per segment, instructing raters to select the five most severe errors when segments contain more than five errors. This constraint helps maintain a balanced evaluation approach and reduces the disproportionate influence of lengthy segments. These indications force annotators to work on evaluation tet sets with source-translation alignment-then document-level MQM scores are calculated by averaging segment-level scores-or by limiting themselves to the most serious errors in a given translated document. This method allows for in-depth analysis at both sentence and document levels, capturing more granular quality aspects that continuous scoring methods like DA might overlook. Moreover, MQM has proven more capable of distinguishing between different systems compared to DA+SQM in paragraph-level setup [Kocmi et al., 2024, Riley et al., 2024a]. While MQM is the most widely used human evaluation metric within the MT community, [Riley et al., 2024a] found that there is less agreement among raters when assessing individual documents compared to when evaluating an MT system as a whole. Also Zhang et al. [2024] showed that MQM was inadequate for literary translation (as part of document-level MT). This suggests that MQM, despite its strengths, may not always offer the most robust solution for document-level evaluation due to these inconsistencies.

- The **Error Span Annotation** (ESA) method [Kocmi et al., 2024], combines elements from DA+SQM and MQM. In ESA, annotators highlight spans of text with errors and classify them as *Minor* (e.g., grammatical issues) or *Major* (e.g., meaning-altering errors). After marking the errors, they assign a overall score for the entire translation segment, similar to DA+SQM's 0-100 scale. Unlike MQM, ESA does not require categorizing errors into specific types, which simplifies the annotation process. This method aims to offer a balance between the depth of MQM and the efficiency of DA, allowing for quicker yet still informative human evaluations. Moreover, the incorporation of an additional DA score in ESA helps to counteract certain shortcomings of error span annotations when transformed into a single overall score like with MQM. At the paragraph level, for instance, if a mistranslated source element is repeated several times in the source and consequently in the translation, the single error score could accumulate this error disproportionately. The inclusion of a DA+SQM score mitigates this issue by providing a broader contextual assessment, ensuring that the overall quality score reflects both localized and repeated errors without overemphasizing them. So ESA also supports document-level evaluation by considering both local errors and the overall translation quality within the context of a document. While the data used for ESA annotation in the study—sourced from the WMT23 general shared task [Kocmi et al., 2023]—consisted of relatively short documents, with no document exceeding five sentences, feedback from annotators highlighted a potential limitation of the method for longer texts. This could impact the efficiency of ESA for evaluating "long documents".

4.2 Document-aware meta-evaluation

As the evaluation of MT systems progresses from sentence-level to document-level, a critical question arises: how reliable are human scores when used as the objective reference for automatically evaluating entire documents? The transition to document-aware evaluation not only demands more sophisticated automatic metrics but also requires a deeper understanding of the complexities involved in relying on human judgments as the ultimate target for these metrics. As exposed in Section 4.1, in most studies, human scores are treated as the gold standard, assuming they reflect translation quality accurately across contexts. However, this assumption is often implicit and rarely questioned, despite the known challenges of human evaluation, particularly in document-level settings.

One of the primary challenges is the subjectivity and variability in human scores. Annotators may provide different evaluations depending on the amount of context available, the complexity of the text, and their familiarity with the domain. Document-aware human evaluations, which incorporate larger spans of text for judging translation quality, aim to mitigate these issues by providing the broader context needed for accurate assessments. However, even with an additional context, human evaluations can be inconsistent, as shown in studies that highlight variance in annotator agreement when assessing context-dependent phenomena, including contextual errors,

such as pronoun resolution, discourse relations, or terminology consistency [Läubli et al., 2018, Castilho et al., 2020]. This variability complicates the use of human scores as a reliable target for automatic metrics, particularly in document-level evaluations where such context-dependent elements play a significant role.

Meta-evaluation studies —evaluations of the evaluation methods themselves— must explicitly address these challenges. When human scores are used as the objective reference, it is crucial to critically analyze how well they capture the full range of translation quality factors, especially those related to contextual errors at document-level. For example, while sentence-level evaluations might focus primarily on fluency or fidelity, document-level evaluations must consider additional factors like cohesion and coherence, which are not always reflected in individual sentence assessments. The degree to which human evaluators incorporate these factors into their judgments can vary significantly, raising questions about whether human scores alone should be the primary target for automatic evaluation metrics in document-level settings.

Moreover, the complexity of document-level translation introduces challenges regarding the granularity of human evaluations. Traditional methods, such as segment-based evaluations, can overlook the interactions between sentences that contribute to the overall coherence of a document. In contrast, document-based methods provide a more holistic view but may struggle with statistical robustness due to the limited number of documents typically evaluated. Studies have shown that document-level evaluations can lead to higher variance in scores, making it harder to establish strong correlations between human evaluations and automatic metrics [Graham et al., 2020b, Akhbardeh et al., 2021]. This raises the issue of how to balance the depth of evaluation with the need for reliable, reproducible results.

Another key issue in measuring contextual errors is the cognitive load placed on human annotators. As the complexity of the document increases, so does the difficulty in detecting and scoring context-dependent errors. Annotators may struggle to maintain consistency when faced with longer texts or more intricate discourse structures, leading to potential gaps in identifying contextual errors. This further complicates the use of human evaluations as a benchmark for automatic metrics that aim to capture document-level quality. Automatic metrics, particularly those designed for document-level translations, must be capable of detecting these subtle, context-dependent issues to provide a more accurate assessment of MT quality.

Finally, the use of human scores as an evaluation objective has implications for the development of automatic metrics. Most automatic metrics aim to reproduce human judgments by providing a single, global quality score, as discussed in Section 5. However, the inherent subjectivity in human evaluations suggests that automatic metrics may benefit from a more nuanced approach, focusing not only on replicating overall human scores but also on capturing specific document-level phenomena. Metrics that can disentangle different aspects of translation quality, such as adequacy, fluency, cohesion and coherence, may offer a more reliable evaluation framework than those that aim solely to match human scores.

In conclusion, while human scores are a valuable benchmark for evaluating MT systems, they are not without limitations, especially in document-aware contexts. Contextual errors, which are often invisible at the sentence level, require a deeper, more holistic evaluation approach. Meta-evaluation studies should explicitly address the complexities of relying on human judgments as the objective target for automatic metrics, considering factors such as variability, context-dependency, and the challenges of document-level evaluation. By doing so, researchers can develop more robust and reliable metrics that better capture the true quality of document-level translations and the subtle, yet impactful, contextual errors that may arise.

5 Global Automatic Evaluation

Automatic evaluation metrics provide a means to evaluate MT quality using reproducible, automatic methods that best imitate the judgments provided by human annotators. They are useful both for MT model development (in order to choose the most promising orientations) and for comparison and analysis of MT models. The computations required to perform such automatic evaluations can take different forms depending on how the metrics are designed and the judgments they wish to imitate. The most common strategy is for a metric to provide a single, global quality score encompassing all aspects of a translation (e.g., its fidelity, fluency, coherence, etc.), without the metric necessarily separately modeling each of these aspects (in reality, most metrics do not). Having a single score representing a system's translation quality is practical because it enables simple numerical comparisons between different MT systems. It also dispenses to explicitly capture all aspects leading to a translation's quality and to disentangle them, which would be a theoretically

difficult endeavor [Denkowski and Lavie, 2010, Kocmi et al., 2021, Agrawal et al., 2024]. However, this implies that it will not be possible to identify which aspects of a translation contribute to the quality score. This may be an issue in particular for document-level phenomena. For this, metrics that either target specific linguistic issues [Hardmeier and Federico, 2010, Wong and Kit, 2012, Gong et al., 2015, Miculicich Werlen and Popescu-Belis, 2017] (see Section 6.2 for a discussion of metrics that target specific phenomena) or that rely on an error analysis [Guerreiro et al., 2023] have been proposed.

Another important distinction between metrics is between those that are based on an aggregation of local decisions, versus those that compute their assessment in a holistic manner. To illustrate this distinction, note that BLEU [Papineni et al., 2002] or METEOR [Banerjee and Lavie, 2005] would fall in the first category, as the associated score aggregate multiple local matches, while COMET [Rei et al., 2020a] or BLEURT [Sellam et al., 2020], would belong to the second category."Aggregative" metrics are arguably easier to understand, as the local decisions that make up the global score can be inspected and used for the analysis.

Historically, the basic unit of analysis (the segment) for most existing metrics is a single sentence. To generate robust estimations of translation quality, multiple sentence-level scores must be aggregated, ideally on a random, representative corpus of inputs. For most metrics, such a corpus-level score is a simple statistic of segment-level scores (e.g., the empirical average), whereas for some others (e.g., BLEU [Papineni et al., 2002]), the computation of a corpus-level score is a more complex function of segment-level evaluations.

Extending metrics that were designed to evaluate (a collection of) short spans of texts to scores that could represent a fair assessment of the quality of a full document's translation has consequences for the metrics concerned and we discuss some of the associated challenges below.

In this section, we first list the document-level corpora (i.e. with consecutive sentences and document boundaries) that have been used to evaluate document-level translations. We then discuss how existing sentence-level automatic metrics have been extended to evaluate the quality of document-level translations.

5.1 Document-level test sets

Evaluating an MT's system's quality typically requires using the system to translate a set of sentences that were not seen during model development (a *test set*) and judging the quality of those translated sentences. Test sets are a staple of MT development; they are used when designing new MT systems and evaluating the final performance of a model. The performance of the model is either based on human evaluation judgments (see Section 4) or on automatic metric scores (see Section 5). In the development of automatic metrics, test sets are used to evaluate metrics on their ability to imitate human judgments of translation quality.

Here we summarize all datasets that were indirectly or directly created to test MT systems at the document level. We distinguish two subtypes, depending on whether they just contain aligned source-target documents, to be used as references, or if they also include automatic translations with error annotations.

5.1.1 Parallel documents

Several datasets provide aligned source and target documents suitable for evaluating document-level MT systems. These datasets maintain document boundaries and preserve discourse structures necessary for testing document-level translation. We mention here the most recent and most widely used corpora in the MT community. For a more exhaustive list, we invite the reader to look at the surveys of Peng et al. [2024b], Abdul Rauf and Yvon [2020].

- **WMT General MT**:[6] The Conference on Machine Translation (WMT) shared task introduces new test sets specifically designed for document-level MT evaluation. These test sets focus on multiple domains, including news, social media, speech (with audio and automatic speech recognition transcripts), and literary texts. They are provided at the paragraph level, ensuring that document context is preserved. The language pairs include translations from English to Chinese, Czech, German, Hindi, Icelandic, Japanese, Russian, Spanish (Latin America), and Ukrainian, as well as Czech to Ukrainian and Japanese to Chinese. The test sets aim to assess general MT capabilities across various domains and encourage the use of document-level context in translation.

[6] https://www2.statmt.org/wmt24/translation-task.html

- **IWSLT TED Talks Corpus**: Transcribed and translated TED talks[7] from the International Workshop on Spoken Language Translation (IWSLT) [Cettolo et al., 2012].
- **OpenSubtitles Corpus**: A large collection of movie and TV subtitles in multiple languages, focused on casual dialogue and language rich in cultural references. It provides a large source of conversational text, making it useful for evaluating how well MT systems handle informal language, register shifts, and the maintenance of context across multiple sentences. [Lison and Tiedemann, 2016]. The corpus provides substantial data for context-aware-level translation in dialogue-heavy contexts intead of pure document-level transaltion as we don't have a strick document boundaries.
- **WMT Biomedical Translation Task**:[8] The WMT Biomedical Translation Task provides datasets consisting of aligned biomedical documents across multiple languages. These datasets include Medline abstracts [Bawden et al., 2019, Névéol et al., 2020], Scielo scientific publications [Neves et al., 2016], EDP scientific publications, and clinical trial reports from ReBEC. Language pairs cover English with French, Spanish, Portuguese, German, Chinese, Romanian, Italian, and Russian. The test sets are designed to evaluate MT systems on real-world biomedical texts, maintaining document boundaries and context essential for assessing translation quality in this specialized domain.
- **News Commentary Corpus**:[9] A collection of aligned political and economic news articles and opinion pieces, with a focus on formal writing covering politics, economics, and international relations. The corpus, which is available in multiple language pairs, is regularly updated as part of the WMT shared tasks, and it provides well-structured, coherent text typical of journalistic content. This makes it valuable for evaluating how well MT systems perform in translating news and editorial articles at the document level.

5.1.2 Test sets with error annotations

In this section, we present a table of test sets consisting of parallel documents with translations that are annotated by labels. These labels can take the form of human judgments, as described in Section 4, or specific error annotations designed during the creation of the dataset, such as those found in the BWB (Bilingual Web Books) corpus [Jiang et al., 2022] where the authors identified and categorized the discourse errors[10] made by MT systems that are not captured in sentence-level evaluation. These labeled test sets serve a dual purpose: they provide a benchmark for evaluating translation quality and offer valuable resources for the development and refinement of automatic metrics, as outlined in Section 5.2. By associating errors with specific translation outputs, these test sets enable more targeted evaluations of MT systems, particularly in capturing document-level phenomena.

[7] https://ted.org

[8] https://github.com/biomedical-translation-corpora/corpora?tab=readme-ov-file

[9] https://opus.nlpl.eu/News-Commentary/corpus/version/News-Commentary

[10] Eight translators classified errors as sentence-level (affecting sentence fluency or adequacy) or document-level (coherence issues across sentences), with document-level errors categorized by linguistic phenomena.

Test Set	Languages	Evaluation	Domain	Reference
From WMT general MT shared task				
WMT20 General MT Shared Task	Zh–En, Cs–En, De–En, Iu–En, Ja–En, Pl–En, Ru–En, Ta–En	DA*†	News, Wikipedia	[Barrault et al., 2020]
WMT21 General MT Shared Task	Ru–En, En–Is, Ha–En, En–Zh, De–En, Zh–En, En–Ja, En–Cs, Fr–De, En–Ru, Bn–Hi, Zu–Xh, En–De, Ja–En, De–Fr	DA*	News, Wikipedia	[Wenzek et al., 2021]
WMT22 General MT Shared Task	Cz–En, Cz–Uk, De–En, De–Fr, En–Cz, En–De, En–Ja, En–Liv, En–Ru, En–Zh, Fr–De, Ja–En, Ru–En, Ru–Sah, Sah–Ru, Uk–Cz, Zh–En	DA and DA+SQM*†	News, Social, E-commerce, Conversation	[Freitag et al., 2023]
WMT23 General MT Shared Task	Zh–En, De–En, He–En, Ja–En, Ru–En, Uk–En, Cs–Uk, En–Cs	DA+SQM*†	News, Social/UGC, Manuals, E-commerce, Meeting notes, Speech	[Blain et al., 2023]
WMT24 General MT Shared Task	Cs–Uk, Ja–Zh, En–Zh, En–Cs, En–De, En–Hi, En–Is, En–Ja, En–Ru, En–Es-LA, En–Uk	ESA*†	News, Literary, Speech, Social	[Freitag et al., 2024]
From WMT metric shared task				
WMT21 Metric Shared Task Test Set	En–De, En–Ru, Zh–En	MQM*	News and TED talks	[Freitag et al., 2021b]
WMT22 Metric Shared Task Test Set	Zh–En, En–Ru, En–De	MQM*	News, Social, E-commerce, Chat	[Freitag et al., 2022]
WMT23 Metric Shared Task Test Set	En–De, He–En, Zh–En	MQM*†	News, Conversational, User Reviews, Manuals, and Social	[Freitag et al., 2023]
WMT24 Metric Shared Task Test Set	En–De, En-Es, Jz–Cz	MQM*†	News, Literary, Speech, Social	[Freitag et al., 2024]
Other Test Sets				
Adapted WMT20 Shared Task Test Set	En–De, Cz–En	MQM*	News	[Freitag et al., 2021a]
BWB	Cz–En	Doc-level errors	Sci-fi, Romance, Action, Fantasy, Comedy	Jiang et al. [2022]
Bio MQM dataset	Pt→En, En–De, En–Es, En–Ru, En–Fr, Zh–En	MQM*	Abstracts from crawled academic papers	[Zouhar et al., 2024]
GeneralMT2022 Multi-Segment Annotations	En–De, En-Cz	MQM*	News, Social, E-commerce, Conversation	[Riley et al., 2024b]
LitEval-Corpus	En–Zh, De–En, De–Zh	MQM, SQM, and BWS *†	Literary	[Zhang et al., 2024]

Table 6: Overview of test sets with human evaluations at the document level. For a given document, evaluations may include direct scoring of the entire document (denoted by †) or scoring each segment or sentence, which can then be aggregated into a document-level score (denoted by *).

5.2 Global automatic evaluation metrics

There is a vast literature discussing global sentence-level automatic metrics in MT (e.g, Koehn, 2020, chap. 4, Chatzikoumi, 2020, Lee et al., 2023), and many implementations of these scores. These metrics can be classified into four categories:

1. *surface-form metrics* rely on a crude comparison of an automatic translation with one or several gold references. Such comparisons take into account the surface forms of words, subwords or characters, e.g., by using n-grams, word sets, taken alone or in combinations with a metric addressing a specific to a particular linguistic phenomenon. Examples metrics include BLEU [Papineni et al., 2002], TER [Snover et al., 2006], METEOR [Banerjee and Lavie, 2005] and ChrF [Popović, 2015] (Section 5.2.1).

2. *metrics involving comparisons performed in embedding (continuous) spaces*, such as BERTscore [Zhang et al., 2020b], Prism [Thompson and Post, 2020a] or MoverScore [Zhao et al., 2019] (Section 5.2.2).

3. *metrics involving fine-tuning language models* on human judgments of MT quality (MQM, pairwise reranking, or DA), as exemplified by COMET [Rei et al., 2020a], MetricX-23 [Juraska et al., 2023], and MaTESe [Perrella et al., 2022] (Section 5.2.3).

4. *metrics involving prompting large language models* to generate judgments regarding the quality of MT outputs. This category includes metrics such as GEMBA-MQM [Kocmi and Federmann, 2023a] and AUTOMQM [Fernandes et al., 2023] (Section 5.2.4).

While a majority of metrics rely on the comparison of system outputs with one or several reference translations (particularly the case of metrics relying on the comparison of surface forms), some of the metrics mentioned in this section also exist in reference-free versions (also known as *quality estimation metrics*) [Specia et al., 2017], relying solely on a comparison of system outputs with the source texts. Some examples include CometKiwi [Rei et al., 2022b], MetricX-23-QE [Juraska et al., 2023], KG-BERTScore [Wu et al., 2023], and GEMBA-MQM.

A key distinction in this context is between segment-level and system-level scores. A segment-level score refers to the evaluation of a single translated segment, which could be a document, paragraph, or sentence, depending on the corpus or test set. In contrast, a system-level score aggregates scores across all segments translated by an MT system, providing an overall quality measure for the system as a whole.

A basic requirement of an automatic document-level metric is to provide a single quality score for an entire document (for a segment-level setup). With the exception of BLEU, discussed at length below, this is traditionally achieved by aggregating segment-level scores (often sentence-level) across an entire document [Mohammed and Niculae, 2024], using for example simple averaging. However, this aggregation approach introduces several challenges when applied to document-level evaluation.

First, in many document-level evaluation setups, the use of sentence-level metrics is not always straightforward, as 1-to-1 sentence alignments between the translation hypothesis, the source, and/or the reference text are not always present. This is because, unlike sentence-level MT systems, document-level MT systems are not constrained to produce one output sentence for each input sentence, making alignment non-trivial and sentence-level scoring difficult. It also make the evaluation of different systems non-comparable at segment-level [Wicks and Post, 2022]. This lack of alignment can compel evaluators to either work with the entire document or to re-align sentence spans,[11] adding complexity to the evaluation process and potentially diminishing the effectiveness of sentence-level metrics in accurately reflecting the true quality of document-level translations [Wicks and Post, 2022].

Second, while this method allows for a global assessment of a document's quality, it inherently fails to capture inter-sentential context and lacks the ability to evaluate broader discourse phenomena that are crucial to the quality of longer texts [Maruf et al., 2021, Abdul Rauf and Yvon, 2020]. Automatic metrics designed at the sentence level may therefore not precisely capture the quality of translations when assessed at the document level [Libovický et al., 2018, Deutsch et al., 2023a, Jin et al., 2023, Post and Junczys-Dowmunt, 2023] and therefore limiting their effectiveness in producing a comprehensive system-level assessment..

In the remainder of this section, we discuss how these challenges have been addressed for each type of metric listed above.

[11] As is necessary when evaluating speech translation systems [Matusov et al., 2005].

5.2.1 Surface metrics

Surface-based metrics compare the surface form of an MT system's output with one or more reference (or gold) translations (produced by human translators), the idea being that better automatic translations are more likely to overlap with ideal translations. This is for example the basis of the BLEU score [Papineni et al., 2002], and the metrics that sought to improve on it such as METEOR [Banerjee and Lavie, 2005] and ChrF [Popović, 2015]. BLEU is the most commonly used metric in the MT community for its ease of use, linguistic independence, and low computational cost. This metric can handle very long segments (in terms of the number of tokens), which is a very relevant criterion for documents' evaluation. However, in addition to the common criticisms of BLEU such as it failing to measure semantic similarity in the absence of lexical overlap (e.g., [Callison-Burch et al., 2006]), the standard BLEU score is insensitive to the contextual enhancements targeted by document-level MT systems. Improvements made by a document-aware translation system are often restricted to a small group of words (pronouns, connectives markers, etc., see Section 3.2), and, while being crucial to understanding a text, may remain undetected by statistical indicators based on n-gram counts. Another weakness of surface-based metrics is their reliance on an exact match between reference and hypothesis, when the correctness of a translation sometimes depends on agreement/consistency between target words: this is, for instance, the case of the translation of referential *it* into French, which depends on the gender of translation of the antecedent in the hypothesis, and may not necessarily match the reference translation).

Furthermore, document-level translation systems, particularly LLMs, seem to be able to take more liberties with word choice and to paraphrase more than conventional NMTs [Raunak et al., 2023]. This distances translation hypotheses away from their references, and BLEU may unfairly penalize this.

Document-level variants of BLEU The standard BLEU score ("vanilla" BLEU) assesses n-gram matches between sentence-aligned system outputs and reference translations, with n-gram counts being calculated across the entire corpus of sentences. It also incorporates a length penalty to penalize translations that are too short, such that the complete formulation of BLEU is as follows:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log\left(p_n(C_{sent})\right)\right) \tag{1}$$

where $p_n(C_{sent})$ is the modified n-gram precision evaluated on the sentence-level corpus C_{sent}:

$$p_n(C_{sent}) = \frac{\sum_{C \in \{C_{sent}\}} \sum_{n\text{-gram} \in C} Match_{clip}(n\text{-gram, C}}{\sum_{C' \in \{C_{sent}\}} \sum_{n\text{-gram}' \in C'} Count(n\text{-gram',C'})} \tag{2}$$

where $Count(s, C)$ counts the number of occurrences of s in C, and $Match_{clip}(s, C)$ counts the (clipped) number of occurrences of s that are both in C and in the reference translations of C. The BP (brevity penalty) is defined as follows:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp(\frac{c-r}{c}), & \text{otherwise.} \end{cases} \tag{3}$$

where c and r are the total length (in tokens) of the system outputs and the reference translations respectively.

At an abstract level, this standard *corpus-level* version of BLEU involves three main design choices:[12]

1. the level at which matches are counted, here sentences; for this, a one-to-one sentence alignment with the reference is necessary;
2. the level at which matches are accumulated to compute modified precisions, here the full corpus (cf. Equation 2); this is also the level used to evaluate and compare lengths, and for which Equation (1) is computed;
3. the level at which BLEU scores are averaged, here again the corpus, which means that we have only one term in the average.

[12] Notwithstanding the setting of the value of N.

A *sentence-level* variant of BLEU also exists, dubbed **sentBLEU**, which relies on different settings: (1) matches are computed at the sentence level, (2) BLEU scores are also computed at the sentence-level (as if each sentence were a mini corpus), (3) the averaging of individual scores can then optionally be performed at the corpus level to derive a single score (although the vanilla BLEU score is preferable as an aggregated corpus-level score). Smoothing techniques must be used to avoid null values for n-gram matches [Lin and Och, 2004, Gao and He, 2013, Chen and Cherry, 2014], which often happens for short segments, which would be problematic for the computations of the log() function.

Assuming the sentences in a corpus are further grouped into documents, this sentence-level version of BLEU can readily be aggregated at the document level – with the caveat that it would not evaluate global errors involving several segments and provided that sentence-level alignments can be calculated. There are however several other ways to turn BLEU into a document-level metric (illustrated in Figure 1):

- The first (illustrated in Figure 1d) (1) computes matches at the document-level but (2) aggregates these values once for the entire corpus to derive one single BLEU (that does not need any averaging). This would be equivalent to computing the vanilla BLEU as if each document were a single sentence. This is the approach of Liu et al. [2020b], introduced in their work (and several follow-ups) as **d-BLEU score**.

- An alternative (illustrated in Figure 1f) (1) compute matches in aligned documents (as if each document were a single long segment), and then (2) compute one BLEU score per document, and (3) average at corpus-level. This is akin to performing sentence-level BLEU with very long sentences and may yield optimistic values for matches count. This variant is used by Libovický et al. [2018] to extend sentBLEU to perform BLEU computations at the *paragraph level*. It is further elaborated and used to evaluate document-level systems by Peng et al. [2024a], who proposed the name **ds-BLEU** for this specific variant.

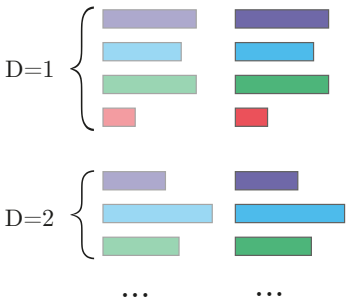
- Alternatively, as illustrated on Figure 1g, it is possible to (1) computes matches at the sentence level, which are then (2) aggregated into document-level BLEU scores and (3) an average over documents can then be performed to derive a single value. In this view, we compute vanilla-BLEU separately for each document.

As they compute a BP for each document, the first two versions imply a stricter control of the target length, which needs to match the reference length for each document. The last two versions, contrarily to the first one and standard versions of BLEU, only require alignment at the document level, and can handle cases where the reference and automatic translations differ in their number of output sentences. As each segment contains several sentences, the likelihood of obtaining a null value for n-gram matches is reduced compared to when a segment is just a single sentence. A downfall is that as the length increases, the reliability of short n-gram precision decreases, as the most common words in a language are more likely to be matched by mere chance [Libovický et al., 2018]. This means that these variants might not be directly appropriate for very long documents, unless the value of N increases with the document length. In any case, even though these variants seem better suited to handle documents than the simple averaging of sentBLEU scores, it is dubious that they will correctly capture contextual dependencies in the translation, and might be plagued by the general weaknesses of BLEU. Figure 1 illustrates the differences between these extensions of BLEU the document-level. Many more variants could be entertained, notably using weighted averages instead of the empirical mean, so as to mitigate the variance in document length.

Finally, it should be noted that similar extensions could be defined for other surface or neural metrics, and while some work in this direction exists [Gong et al., 2015, Libovický et al., 2018], they have not yet been exhaustively explored, particularly due to the criteria mentioned in the introduction, which make surface metrics very insensitive to phenomena at the document level.

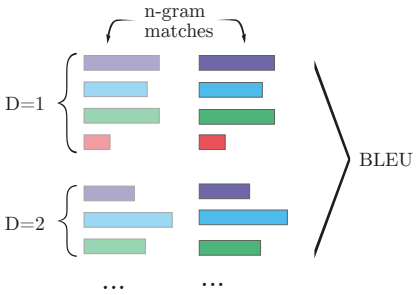
5.2.2 Embedding-based metrics

Embedding-based metrics capture the similarity between hypothesis and reference translations using embedding vectors computed by language models [Lee et al., 2023]. These models, trained on vast amounts of textual data, produce vector representations for words that encapsulate their context and the relationships between them. Unlike traditional fixed word-embeddings, these are often referred to as contextual-embedding vectors due to their ability to capture meaning in context [Liu et al., 2020a]. Contextual embedding representations generated by these models have been used to improve MT metrics by reflecting semantic similarity, rather than relying solely on lexical matching as observed with surface-level metrics (Section 5.2.1).

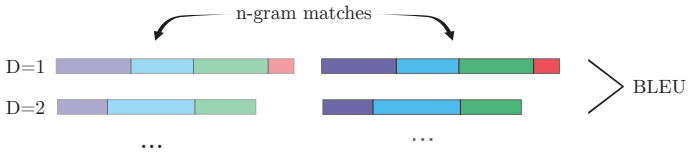


(a) A two-document corpus, with sentence alignments

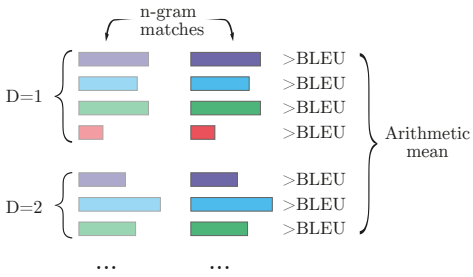
(b) Two documents viewed as long sentences



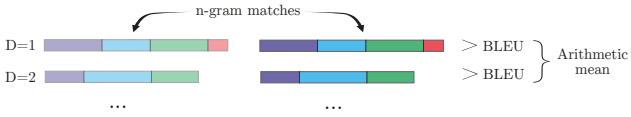
(c) Vanilla BLEU scores



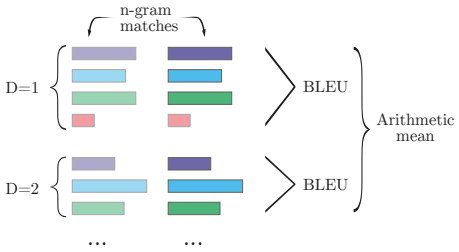
(d) d-BLEU is Vanilla BLEU, with document-as-sentences



(e) Vanilla SentBLEU



(f) ds-BLEU is SentBLEU, with document-as-sentences



(g) Vanilla BLEU scores, computed document-wise

Figure 1: Various ways to compute BLEU scores for document-level systems. Opaque colors correspond to translations, while transparent colors represent references.

The two main embedding-based metrics in the literature are BERTScore [Zhang et al., 2020a] and PRISM [Thompson and Post, 2020b]. BERTScore is an unsupervised metric that simply measures the cosine similarity between the contextual word embeddings of a pre-trained BERT-based model Devlin et al. [2019]. BERTScore encodes tokens from both the reference and the hypothesis, then computes soft alignments based on token similarities. These alignments are used to calculate precision, recall, and F1 scores of the hypothesis compared to the reference.

Prism, another unsupervised embedding-based metric, adopts a sequence-to-sequence neural paraphrasing approach to evaluate how well an MT hypothesis paraphrases a human reference translation. The evaluation process involves feeding the reference into the encoder and force-decoding the hypothesis in the decoder. The token-level probabilities of the output are aggregated to produce a score, and this process is repeated with the roles of the reference and hypothesis swapped. The final score is the average of the two resulting scores.

Both metrics have a high correlation with human judgment at the sentence level, even though they are outperformed by fine-tuned metrics (discussed in Section 5.2.3). Nevertheless, since the pre-trained language models underpinning these metrics are typically trained on sentence-level data and impose a maximum input length (as the fine-tuned metrics), embedding-based metrics seem inherently designed for sentence-level evaluation.

Doc-BERTScore and doc-Prism A context-aware extension of the BERTScore and PRISM metrics (respectively dubbed **doc-BERTScore** and **doc-Prism**) is introduced by Vernikos et al. [2022]. To extend BERTScore to the context-aware level, they provide the reference context while encoding the hypothesis or the reference with a pre-trained BERT model. However, they align only the tokens of the reference and hypothesis sentences to compute the alignment score, as illustrated in Figure 2. This method is very generic, and, as the authors point out, could also apply to other embedding-based metrics. For Prism, they concatenate the reference context to both the reference and hypothesis. The context is used as a prompt: that is, they only aggregate token-level probabilities for the sentence being evaluated. In their extension of Prism to the context-aware level, they use mBART-50 [Tang et al., 2020], a multilingual encoder-decoder language model, trained on document fragments, instead of retraining Prism specifically for the context-aware level.

The evaluation setup is the same as for doc-COMET (see Section 5.2.3) and involves computing correlations with human judgements. For both metrics, adding document-level contexts leads yields an increase in performance.

For Prism, they further observe that the sentence-level results obtained with Thompson and Post's multilingual model (m39v1) are better than the sentence-level results with mBART-50. However, by using document-level context, they are able to improve over the sentence-level Prism with mBART-50 in every language pair and domain. This narrows the gap between the mBART-based version of Prism and the one based on m39v1, even outperforming the stronger m39v1 model for two TED language pairs.

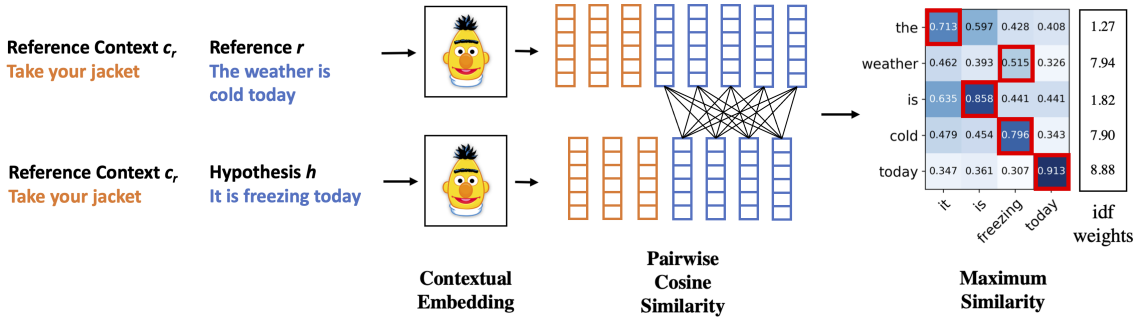


Figure 2: Borrowed from Vernikos et al. [2022] and adapted from [Zhang et al., 2020a]: To extend BERTScore to the document-level, the reference context is added to both the reference and the hypothesis sentences. This context is used to improve the contextual embeddings of the reference and hypothesis sentences, but it is not used when performing alignment and scoring, which follows standard sentence-level BERTScore.

The extension of Vernikos et al.'s work to other metrics is presented in Section 5.2.3.

To the best of our knowledge, no attempts have been made to extend embedding-based metric, such as the BERTScore, beyond a context-aware adaptation to a clear document-level metric. This could potentially involve creating an adapted dataset specifically for training, or leveraging sentence embedding models tailored to capture document-wide relationships. Such an extension might allow

BERTScore to more effectively account for larger-scale contextual information, further enhancing its ability to reflect document-level coherence and semantic consistency.

5.2.3 Fine-tuned metrics

Fine-tuned metrics consist of a pre-trained language model that has been fine-tuned as a regression model or a ranking model with human quality judgments, such as MQM or DA (see Section 4.1). Typically, these models are trained using datasets with human judgments of quality, like those from the WMT shared tasks. These metrics generally encode the source, hypothesis, and reference segments via a multilingual pre-trained language model (LM), and compute the representation of each sentence as the average of its output token embeddings. These representations are further combined via an aggregation function and then fed to a regressor that predicts a score for each output sentence.

Fine-tuned metrics like COMET [Rei et al., 2020a], MetricX [Juraska et al., 2023], MaTESe [Perrella et al., 2022], BLEURT [Sellam et al., 2020], Unite [Wan et al., 2022], and their variants have been the top performers in recent WMT metrics shared task evaluations [Freitag et al., 2022, 2023] and are recommended by the shared task organizers.

Nevertheless, fine-tuned metrics cannot be readily applied at the document level. This is primarily because they are fine-tuned with sentence-level data (as most training data is at the sentence level): segments are sentences, and associated labels are human evaluations carried out at the sentence level, and the metrics therefore may have difficulties to generalize to sentences in context, paragraphs and documents. This means that their scores will not take into account errors related to discourse phenomena. Additionally, due to the computational cost of training such architectures, a maximum input length is imposed, which restricts direct inference on long documents. In the remainder of this section, we discuss approaches aiming to mitigate such issues.

doc-COMET, Doc-COMET-QE Vernikos et al. [2022] introduce a straightforward method to make sentence-level pre-trained MT metrics context-aware. They propose to enrich the hypothesis-reference-source or hypothesis-source comparisons by including additional context from the reference document. This context, consisting of the two preceding sentences from the source and/or the reference, is used to enhance the contextual embeddings from the encoder model of both the hypothesis and the reference sentences. After embedding, the extra context is discarded, and the score is computed in the same manner as for sentence-level metrics. The classifier used for the regression part remains the same, and no additional context vector is added as input. Notably, only the context that is actually available and applicable is taken into account; for instance, the first sentence of a document will have no context, and the second sentence will have one single preceding sentence as its context. These authors apply this method to two fine-tuned metrics (COMET and COMET-QE) to create what they respectively call **doc-COMET**, and **doc-COMET-QE**.[13]

In their evaluation, the authors compare the performance of their context-aware metrics with human-generated MQM annotations from the 2021 WMT metrics shared task [Freitag et al., 2021b], providing an assessment across different benchmarks. Specifically for doc-COMET-QE, they conduct a targeted evaluation of discourse phenomena using contrastive test sets such as ContraPro (EN-DE), ContraPro (EN-FR) and DiscEvalMT (Section 6.1.2), showing that the metrics can handle context-sensitive phenomena. Moreover, they compare the document-level adaptations of each metric against their sentence-level counterparts and also include comparisons with other recent metrics like BlonDe [Jiang et al., 2022] (6.2). Results (reported in Table 7) suggest that incorporating some document-level context in a pre-trained metrics can enhance correlation with human judgments. The study also revealed that BlonDe, a metric specifically designed for the document level (see Section 6.2) has lower performance than both pre-trained metrics and their proposed context-aware extensions.

Doc-COMETkiwi Hendy et al. [2023] adapt the COMETkiwi metric (the referenceless version of COMET) to the document level by using a sliding window approach to handle sentence alignments that are not one-to-one. This adaptation, dubbed **doc-COMETkiwi**, averages COMETkiwi scores across segments to compute a global translation quality score. This method evaluates sentences across multiple contexts, overcoming the limitations of traditional sentence-level metrics without requiring any re-training or fine-tuning.

This straightforward adjustment offers three distinct benefits compared to standard sentence-level evaluations. First, it allows for the assessment of each sentence within its surrounding context.

[13] As discussed in Section 2, a more appropriate term would be context-level-COMET, as the evaluation is still performed at the sentence-level (with limited previous context), rather at the level of a complete document.

Model	TED talks			News		
	En→De	En→Ru	Zh→En	En→De	En→Ru	Zh→En
BlonDe	-	-	-0.232	-	-	0.212
Prism (m39v1)	0.656	0.867	0.272	0.841	0.799	0.558
COMET	**0.818**	0.841	0.266	0.772	0.659	**0.628**
Doc-COMET	0.816	**0.849**	**0.297**	**0.802***	**0.676**	0.513
COMET-QE	0.694	0.818	**-0.209**	0.711	0.688	**0.529**
Doc-COMET-QE	**0.724**	**0.830**	-0.255	**0.733**	**0.733***	0.462

Table 7: Table adapted from [Vernikos et al., 2022]: System-level Pearson correlations with WMT 2021 MQM annotations for several metrics (and their document-level variant). Within each document/sentence-level pair, **bold** denotes the best correlation and "*" denotes a statistically significant ($p < 0.05$) difference.

Model	En→De			En→Fr				
	Intra	Inter	Total	Intra	Inter	Total	Anaphora	WSD
Lopes et al. [2020]	-	-	70.8	-	-	83.2	82.5	55.0
COMET-QE	78.2	40.9	48.4	76.3	76.6	76.5	50.0	50.0
Doc-COMET-QE (this work)	**80.5**	**72.6**	**74.2**	**88.7**	**88.0**	**88.3**	**83.5**	**68.0**

Table 8: Accuracy (percentage correct) for targeted evaluation of contextual phenomena. Our document-level version of COMET-QE substantially outperforms the sentence-level COMET-QE, and also outperforms the best methods proposed by Lopes et al. [2020], demonstrating that it is successfully incorporating contextual information.

Second, the sliding window approach entails that sentences will be evaluated in multiple contexts, thanks to the overlapping nature of the windows. Finally, this approach overcomes the constraints of limited context windows in evaluation models, which might otherwise impair the assessment of quality in longer sections of text.

The authors defined this generalization of COMETkiwi to assess the MT ability of GPT models at the document-level. They directly assess the performance of the model by comparing COMETkiwi, d-BLEU (Section 5.2.1), and doc-COMETkiwi on the WMT22 test set. However, they evaluate document-level systems without conducting a meta-evaluation of doc-COMETkiwi upstream, notably to test their hypothesis that the average score over several sentences of a sentence-level metric (fine-tuned mainly on sentence-level data) actually captures error scores of targeted document level phenomena.

PARA-UNIF and PARA-STRAT Deutsch et al. [2023b] introduce PARA-UNIF and PARA-STRAT, two BLEURT-style fine-tuned regression models [Sellam et al., 2020] based on the mT5 encoder-decoder language model of Xue et al. [2021]. These metrics are based on the construction of paragraph datasets from sentence-level corpus and are designed to evaluate the quality of paragraph translations in a way that accounts for cross-sentence dependencies and document-level phenomena.

The authors use WMT data from 2019 onwards, given that both the DA and MQM sentence-level annotations were carried out in the context of their surrounding paragraphs and could therefore been seen as a proxy for paragraph-level ratings. Paragraph-level sores are derived either by averaging the DA scores or summing MQM scores assigned to individual sentences within a paragraph block.

PARA-UNIF involves training a metric using data that is uniformly sampled from paragraph instances constructed from sentence-level translation data. In uniform sampling, paragraphs are selected randomly regardless of their length (i.e., their number k of sentences). This means that shorter paragraphs (e.g., $k = 1, 2$) are likely to appear more frequently in the training data because they are more common, while longer paragraphs (e.g., $k = 10$) are less frequent. In contrast, for PARA-STRAT, stratified sampling is used to ensure a balanced representation of paragraphs of different lengths in the training data. The authors create a dataset where there is an equal number of paragraphs for each k value from 1 to 10. This approach helps the metric evaluate paragraphs of varying lengths more effectively, addressing the issue that longer paragraphs are rarer and might therefore not be well represented in a training set built with uniform sampling.

They compare their performances with BLEU, COMET-22, and AutoMQM (using the PaLM-2 model [Anil et al., 2023]) as sentence-level metrics applied to paragraphs, with document-level

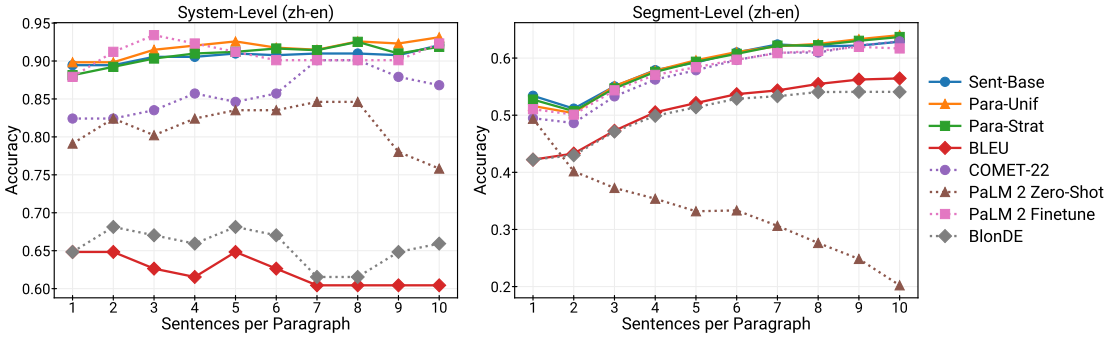


Figure 3: Figure borrowed from [Deutsch et al., 2023b]: System- and segment-level accuracy results for the zh-de language pair on the paragraph-level WMT'22 MQM data for different numbers of k sentences per paragraph. As the number of sentences per paragraph increases, the accuracy scores of the metrics appear to either not decrease (system-level, left) or increase (segment-level, right). This suggests that accurately scoring a paragraph is an easier task than an individual sentence, even for metrics that are not trained on paragraph-level examples.

metric (BlonDE) and with sentence-level baseline variants of the PARA- metrics.Contrary to expectations, accuracy improved significantly as the paragraph length increases, particularly at the segment level, as shown in Figure 3. This suggests that longer texts might be easier to score due to noise reduction, resulting in a more reliable signal.

Additionally, training on paragraph-level examples does not necessarily enhance performance; metrics trained on sentence-level data perform comparably when applied to paragraphs. Surprisingly, even sentence-level metrics not specifically trained on longer data effectively generalize to paragraph-level scoring. Two methods of applying sentence-level metrics to paragraphs—direct scoring and averaging individual sentence scores—show similar performance and high agreement with human ratings, indicating that more complex training on paragraph-level data may not yield significantly better results.

The authors conclude that PARA-UNIF and PARA-STRAT do not necessarily outperform metrics trained on sentence-level data, possibly due to bias in the construction of their paragraph-level datasets, which are based on human judgments collected using sentence-by-sentence translations. Nevertheless, this opens the way for methods to extend standard metrics while preserving existing work and exploiting existing datasets.

SLIDE The SLIDE (SLIding Document Evaluator) metric [Raunak et al., 2024] is a referenceless metric which processes blocks of sentences through a sliding window that moves across each test document. The window covers chunks of consecutive sentences and inputs them into an unmodified, off-the-shelf quality estimation model such as COMETKiwi. Given that COMET's underlying encoder is trained on broader contexts, it may demonstrate consistent evaluation behavior beyond typical sentence-level lengths. Also, since a sentence's evaluation may vary depending on its position within a document, it could be beneficial to assess each sentence in various contexts. This approach is very similar to the doc-COMETkiwi metric [Vernikos et al., 2022] described above, but its authors also performed a complete meta-evaluation. Additionally, SLIDE extends the work by Deutsch et al. [2023b] (Section 5.2.3), which did not perform any meta-evaluation of reference-free metrics.

SLIDE was evaluated with different versions of COMET: COMET20-QE [Rei et al., 2020b], COMET22-QE [Rei et al., 2020b], as well as the reference-based version COMET22 [Rei et al., 2022a], against traditional reference-based metrics including doc-COMET (Section 5.2.3), using pairwise system ranking accuracy with WMT22-MQM annotations.

The results demonstrate that incorporating context into COMET-QE metrics make them better at distinguish between systems. Even a single sentence of context already enhances the model's discrimination capabilities. This improvement occurs despite the scores being accumulated over groups of sentences, which differs from COMET's pure sentence-level training.

Moreover, SLIDE performs better than doc-COMET when evaluating paragraphs rather than individual sentences. However, this improvement does not occur with the reference-based version COMET22, consolidating the findings of Deutsch et al. [2023b] regarding the reference-based metrics. Additionally, it was found that some reference-based metrics such as BLEURT and MetricX were still outperforming SLIDE (see Table 9). Finally, this method of extending evaluation to the document level remains a coherent option for reference-free metrics, as no changes to the underlying

Metric	MQM	DA+SQM
Ⓡ metricx_xl_DA_2019	0.865	0.850
Ⓡ metricx_xxl_MQM_2020	0.850	0.861
Ⓡ BLEURT-20	0.847	0.827
Ⓡ metricx_xl_MQM_2020	0.843	0.859
SLIDE (6,6)	0.843	0.838
ⓇCOMET-22	0.839	0.839
SLIDE (7,1)	0.839	0.814
Ⓡ COMET-20	0.836	0.823
Ⓡ Doc-COMET(2)	0.836	
Ⓡ UniTE	0.828	0.847
Ⓡ MS-COMET-22	0.828	0.830
Ⓡ UniTE-ref	0.818	0.838
Ⓡ MATESE	0.810	
SLIDE (2,1)	0.807	0.825
Ⓡ YiSi-1	0.792	0.782
COMETKiwi (WMT-22)	0.788	0.832
Doc-COMET(2)	0.737	0.810
COMETKiwi (Public)	0.770	0.816
Ⓡ chrF	0.734	0.758
Ⓡ BLEU	0.708	0.704

Table 9: Adapted from Raunak et al. [2024]: Pairwise system accuracy for WMT22-MQM annotations. Metrics that use a reference are tagged with a Ⓡ symbols. The entries are of the form **SLIDE** (w, s). w is the window and s the stride.

sentence-based evaluators are needed, making this enhancement essentially cost-free, provided that document boundary annotations are available.

5.2.4 Zero-shot prompting metrics

Recent years have seen a growing interest in applying LLMs to MT [Xu et al., 2024], particularly at the document level [Wang et al., 2023]. Given the multi-tasking capabilities of these models, researchers have also begun investigating their potential to evaluate the quality of automatic translations [Huang et al., 2024].

Initial evaluation work using LLMs has been promising, as seen in the WMT23 metric shared task, where GEMBA-MQM [Kocmi and Federmann, 2023a] ranked as one of the best metrics [Freitag et al., 2023]. It builds on the earlier GEMBA metric [Kocmi and Federmann, 2023b], a GPT-based metric for MT evaluation, based on zero-shot prompting an LLM to produce quality scores for individual segments, which are then aggregated over segments to produce system-level scores. Four different prompts were tested for, simulating different types of human evaluation: direct assessment (GEMBA-DA, scores ranging from 0-100), scalar quality metrics (GEMBA-SQM, scores from 0-100), star ratings (GEMBA-stars, scores from 1-5), and quality class labels (GEMBA-classes, discrete labels). Scores range from 0-100 for GEMBA-DA and GEMBA-SQM, 1-5 for GEMBA-stars, and discrete labels for GEMBA-classes. Seven GPT models were tested, with GPT-4 [OpenAI et al., 2024] used as the default for most experiments. Although the authors did not evaluate GEMBA at the document level, they suggest that GPT-enhanced evaluation metrics could advance document-level evaluation due to their ability to utilize much larger context windows.

In follow-up work, Kocmi and Federmann [2023a] released GEMBA-MQM, which uses three-shot prompting with the GPT-4 model to mark error spans using the MQM framework. This few-shot learning extension helps adapt GEMBA to any language pair. Even though GEMBA-MQM evaluates isolated sentences, it was found at WMT23 that GEMBA-MQM achieved higher system-level correlation scores than metrics specifically designed for document-level evaluation, such as doc-COMET (Section 5.2.3), in the EN-DE paragraph-level evaluation [Freitag et al., 2023]. Fernandes et al. [2023] also designed a prompt technique for MT evaluation, AutoMQM, to replicate MQM scores, as did Lu et al. [2024] with EAPrompt, which uses a single prompt. For deeper analyses on LLMs applied to MT evaluation tasks at the sentence level, we refer readers to the study by Huang et al. [2024].

As highlighted, the use of LLMs brings both advancements and challenges to the MT evaluation field. The rapid development of LLM-based metrics demonstrates the potential of these models to outperform traditional evaluation metrics, particularly at the system level Huang et al. [2024].

Moreover, these same models can be used for other evaluation protocols such as error explanation with x-Tower [Treviso et al., 2024] and even GEMBA. However, the ongoing pursuit of optimized strategies for leveraging LLMs in MT evaluation suggests this will remain a dynamic area of research. Despite these promising developments, significant gaps remain, particularly at the document level. Currently, there are no LLM-based metrics specifically designed to evaluate translation quality across entire documents, and there is a lack of comprehensive studies on how these models perform when handling longer texts or detecting discourse-level translation errors. Addressing these gaps will be crucial for fully realizing the potential of LLM metrics that can effectively handle complex, document-level phenomena.

6 Targeted Evaluation of Specific Phenomena

In contrast to global evaluation metrics, which provide overall scores treating translations as "black-box" outputs, more targeted methods are needed to assess how well MT systems handle specific linguistic phenomena. Global metrics, while useful for system-level evaluations to rank systems, often fail to provide insights into how discourse-level issues, such as coherence and cohesion, are managed. As a result, they may overlook critical document-level errors that impact the naturalness and accuracy of translations.

Given that the frequency of phenomena requiring document-level contexts is not always very high (most sentences do not require document context to be correctly translated) [Bawden et al., 2018a], and that the frequency of these phenomena in standard corpora is not known [Post and Junczys-Dowmunt, 2023], the evaluation of a random segment, even of considerable length such as a paragraph, does not necessarily allow for the identification of translation errors at the document level. Moreover, many of the context-dependent phenomena mentioned in Section 3 are difficult to evaluate because they are dependent on the MT system's previous translation choices, which are not predictable. For example, if translating the two English sentences *The owl was in the barn. It was sleeping*, the correct French translation of *It* (as either *il* (masc.) or *elle* (fem.)) depends on how the antecedent *owl* is translated (*hibou* (masc.) or *chouette* (fem.)). Traditional test sets, whereby a score is given depending on a reference translation will in most cases not correctly reward all correct translations because it does not take into account such choices. To address these limitations, two complementary evaluation methods have emerged for assessing specific linguistic phenomena in MT systems:

- *Test suites*: These are are custom-designed evaluation sets tailored for the evaluation of specific aspects of MT output, allowing for a more focused evaluation [King and Falkedal, 1990]. Faced with the complexity above, test suites exist as an alternative solution increasingly adopted by the MT community to evaluate specific aspects of translation [King and Falkedal, 1990, Isabelle et al., 2017, Bojar et al., 2018].

- *Automatic metrics for specific phenomena*: Recent research has led to the development of automatic metrics that explicitly model and evaluate discourse-related features. These metrics aim to capture coherence, cohesion, and document-level dependencies, providing a more detailed view of an MT system's handling of context-dependent translations.

Both test suites and automatic metrics are designed to complement global metrics (discussed in Section 5) and can be combined with them to provide a more comprehensive evaluation of an MT system's ability to manage document-level phenomena. In this section, we provide an overview of both standard test suites (Section 6.1), then, of alternative automatic metrics to evaluate specific linguistic phenomena (Section 6.2). We focus on evaluating the specific phenomena discussed in Section 3, excluding terminology. The work on terminology constitutes a substantial and separate topic deserving of focused attention and is beyond the scope of this survey. For a detailed discussion on terminology, we refer readers to [Semenov et al., 2023].

6.1 Test suites for document-level phenomena

Test suites are evaluation sets targeting a particular phenomenon or set of phenomena for evaluation. Traditionally, they consist of sets of sentences that contain the targeted phenomena and that may be accompanied by reference translations, which can be used to evaluate outputs generated by an MT system.

Within this framework, for the document-level evaluation, a particular subset of test suites has gained attention: contrastive test suites. These focus not just on the generation of translations, but

also on evaluating the model's ability to differentiate between correct and incorrect translations, targeting specific linguistic phenomena. Contrastive test suites can be classified into two types:

- *Discriminative contrastive test suites*: These are designed to assess the system's ability to distinguish between correct and incorrect translations. The model is given pairs of translations—one correct and one incorrect—and is evaluated based on its ability to assign a higher probability to the correct translation. This method is useful because it avoids the need to account for all possible correct translations, focusing instead on the model's ability to identify the best option. This approach emphasizes error detection rather than generation, allowing for automated evaluation without needing to account for all possible correct translations and solving the reliance on using the translation of past context.

- *Generative contrastive test suites*: These extend the evaluation by testing both the system's ability to generate correct translations and its ability to distinguish between correct and incorrect alternatives. After generating a translation, the system is required to assess the correctness of its output. Generative contrastive test suites are often incorporated into shared tasks, such as those in the WMT evaluation campaign, where systems are first required to generate translations and then to discriminate between correct and incorrect ones. So these sets typically require at least some manual evaluation, as they involve assessing the adequacy and fluency of the generated translations. Generative sets can reflect a system's real-world performance but are more complex to implement and assess.

Contrastive test suites—both discriminative and generative—have been applied to a range of context-dependent phenomena, such as word-sense disambiguation [Rios Gonzales et al., 2017, Bawden et al., 2018b], coreference and anaphora resolution [Bawden et al., 2018b, Müller et al., 2018, Lopes et al., 2020], gender disambiguation [Gete et al., 2022], register selection [Gete et al., 2022], and deixis and ellipsis [Voita et al., 2019]. Some of the discriminative (contrastive) test suites also exist in generative versions, notably when they are submitted as test suites to the WMT shared tasks; test suites are added to the main test set to be translated by all systems and therefore only generative-style test suites can be included.

Both types of test suites have their advantages and limitations. Generative-style test suites evaluate a system's ability to produce correct translations, but often require manual evaluation to go beyond simple global scores. In contrast, discriminative contrastive test suites rely on probability scores, but this does not necessarily reflect the system's ability to generate a correct translation. Therefore, as some systems may be good at discriminating, yet poor at generating acceptable translations [Post and Junczys-Dowmunt, 2023], it is recommended to use both types of evaluation to gain a comprehensive understanding of system performance.

In this section, we present the main test suites—both generative and contrastive—targeting document-level phenomena and discuss how they help assess specific translation challenges. The following test suites we present focus on evaluating document-level phenomena related to coherence and/or cohesion, addressing either multiple aspects simultaneously or concentrating on a single phenomenon. These test suites emphasize the evaluation of document-level aspects related to coherence and/or cohesion, either by focusing on individual elements or by addressing multiple facets at once.

6.1.1 Manually-designed test suites

Document-Level Phenomena test suite at WMT19 (EN–CS) [Rysová et al., 2019] is specifically designed to assess document-level coherence by examining discourse linguistics phenomena. This test suite focuses on three main aspects: topic-focus articulation, discourse connectives, and alternative lexicalizations of these connectives, particularly multi-word discourse connectives. On purpose, this corpus can not be used to address errors in coreference, pronoun and gender translation, although these are very important phenomena for assessing textual coherence. This test suite comprises 101 documents, totaling 3,500 source sentences derived from the Penn Discourse Treebank [Prasad et al., 2019], categorized under "essay" or "letter" types. Trained linguists manually evaluated these documents, analyzing both the English source text and one of the MT outputs. The evaluation involved identifying targeted phenomena in the source text, with annotators marking whether these phenomena were accurately reflected in the MT output.

DELA Challenge Set (EN–$\mathbf{PT_{BR}}$) Castilho et al. [2021] developed an English-Brazilian Portuguese document-level corpus annotated with context-aware issues, namely: gender, number, ellipsis, reference, lexical ambiguity, and terminology. The corpus contains 60 full documents

and was compiled across six different domains: subtitles, literary, news, reviews, medical, and legislation. Jointly, three annotators compiled and annotated each sentence of the corpus regarding the six context-aware issues previously mentioned and found in [Castilho et al., 2020], using a guideline represented by a decision tree. Although the evaluation of each sentence is conducted within the context provided by the entire document, the evaluation judgments are collected at the sentence level and rely on a sentence-by-sentence alignment between the source and the translation hypothesis.

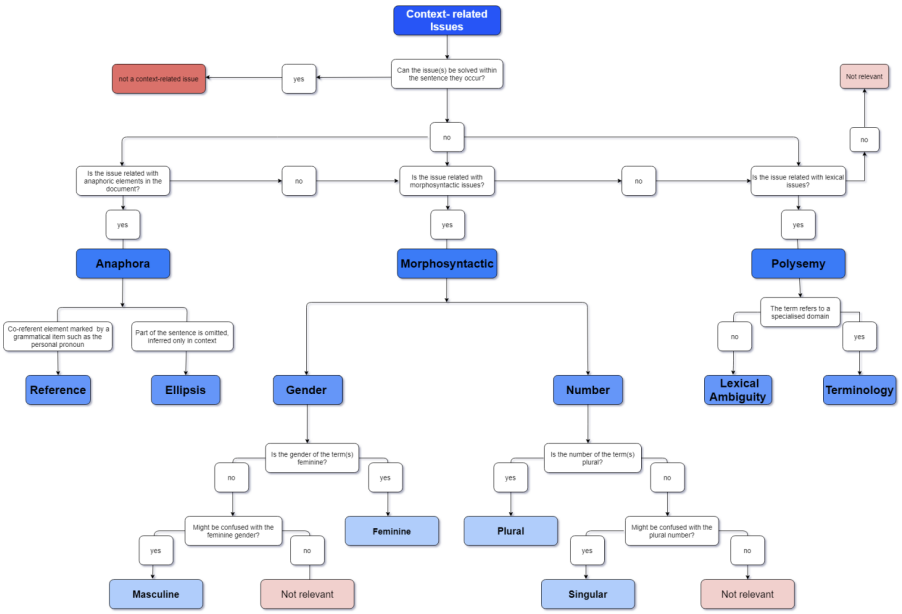


Figure 4: Decision tree used to guide the annotation of document-level issues of the DELA challenge set. Figure from [Castilho et al., 2021].

ELITR Markables Test Suite (EN–CS) [Zouhar et al., 2020] focuses on evaluating the consistency and clarity of domain-specific terms, referred to as "*markables*", across English–Czech and Czech–English translations. This test suite involves six documents, examining 215 occurrences of these markables across various specialized domains. The evaluation process is divided into two phases, involving more than 10,500 assessments to compare how 13 competing MT systems (from WMT20) handle these critical terms against the backdrop of document-level context.

6.1.2 Constrastive test sets

ContraWSD (DE–EN, DE–FR) [Rios Gonzales et al., 2017] is a contrastive word sense disambiguation test set for the German-English and German-French directions. It consists of 7,359 sentences sourced from translated books, news articles, Global Voices, and UN assembly transcripts containing ambiguous German words. Each sentence is accompanied by a (correct) reference translation and several contrastive (incorrect) translations, in which the ambiguous word has been translated with an incorrect meaning (See Figure 10 for an example). This dataset is designed to evaluate how often an MT model can use the context surrounding the ambiguous words to prefer the correct over the incorrect translations, measuring total accuracy, accuracy per word sense, and accuracy across frequency classes of word usage. Although the dataset provides contextual cues within sentences, it is not designed to test document-level translation capabilities, focusing instead on context-aware, sentence-level outputs. A generative version of the test suite for DE–EN was developed in 2018 for the WMT test suites task [Rios et al., 2018], testing the system's ability to generate the correct translation in its top (1-best) output. The test set, aimed at black-box MT systems, features 3,249 sentence pairs and shifts towards a semi-automatic evaluation process that uses both automatic detection and manual checks to assess translation accuracy.

Source: *Er hat zwar schnell den Finger am* ***Abzug****, aber er ist eben neu.*
Reference: *Il a la* ***gachette*** *facile mais c'est parce qu'il debute.*
Contrastive output: *Il a la* ***soustraction*** *facile mais c'est parce qu'il débute.* Contrastive output: *Il a la* ***déduction*** *facile mais c'est parce qu'il débute.* Contrastive output: *Il a la* ***sortie*** *facile mais c'est parce qu'il débute.* Contrastive output: *Il a la* ***rétraction*** *facile mais c'est parce qu'il débute.*

Table 10: Sample from the ContraWSD test set [Rios Gonzales et al., 2017], displaying the translation of the ambiguous word *Abzug* from German into French.

DiscEvalMT (EN–FR) [Bawden et al., 2018b] is a contrastive test set designed to focus on anaphora and lexical choice, which includes 200 pairs of manually crafted, two-sentence examples for each category. These examples are uniquely constructed such that the correct translation depends on the preceding sentence context, effectively balancing each correct translation with an incorrect counterpart to establish a non-contextual baseline with a 50% precision rate. For the anaphora subset, the test involves choosing the correct singular or plural personal and possessive pronouns, with the translation challenging the system to maintain consistency with previously translated antecedents, even when these are translated in a non-standard manner. For the lexical choice set, the challenge lies in addressing lexical ambiguity and repetition (lexical repetition involves cases where the same source word appears in the context and the current sentence, and must be translated in the same way to maintain the cohesive nature of the text). The test set has also been translated into Czech to test English–Czech translation [Jon, 2019].

ContraPro (EN–DE) [Müller et al., 2018] is a large-scale contrastive dataset focusing on the translation of English anaphoric pronouns into German (*er*, *sie*, *es*), drawn from the OpenSubtitles2018 corpus [Lison et al., 2018]. This automatically created set identifies sentences containing instances of the English pronoun *it* present in coreference chains and their corresponding German translations, generating contrastive erroneous translations. The dataset includes 4,000 examples for each target pronoun type, paired with two incorrect variants where the correct pronoun has been replaced, for a total of 24,000 test cases. The disambiguating context for these translations (i.e., the antecedent) can be present in the current sentence or extend across any number of preceding sentences.

ContraPro (EN–FR) [Lopes et al., 2020], an extension of the (EN–DE) ContraPro dataset of Müller et al. [2018], is a large-scale pronoun test set derived again from OpenSubtitles. The dataset contains 14,000 examples, equally distributed among the French pronouns *il'*, *elle*, *ils*, and *elles*. The creation process includes using Neuralcoref[14] to detect English pronouns and their antecedents, and FastAlign [Dyer et al., 2013] to align them to their French counterparts. Contrastive translations are generated by inverting the gender of pronouns (and of words within the sentence that are marked for the gender of the pronoun) using the Le*fff* lexicon [Sagot, 2010].

Contrastive test sets for deixis, ellipsis, lexical cohesion (EN–RU) [Voita et al., 2019] addresses a range of issues such as lexical cohesion (2k instances), anaphora (3k instances), verb selection (500 instances), and morphology (500 instances) in the context of source-side ellipsis. Following [Bawden et al., 2018b], each test instance comprises a true example—sequences of sentences along with their reference translations—and several contrastive translations that are correct at the sentence level but reveal errors only when placed in full context. The provided segments typically extend across three sentences, ensuring adequate contextual information for evaluation. As previously, the system's performance is assessed based on its ability to prioritize the true translation over the contrastive alternatives.

MuCoW [Raganato et al., 2019] is a multilingual contrastive test suite developed to evaluate the word sense disambiguation (WSD) capabilities systems across 16 language pairs. It uses over 200,000 contrastive sentence pairs, automatically constructed from word-aligned parallel corpora and the comprehensive multilingual sense inventory provided by BabelNet [Navigli and Ponzetto, 2012]. The construction of MUCOW involved three steps: (i) identifying ambiguous source words with multiple translations, (ii) clustering target words using BabelNet to group potential synonyms, and refining these clusters with sense embeddings to ensure accurate sense granularity and (iii) creating

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

contrastive sentence pairs by substituting target words with different lexicalizations from the refined clusters.

TANDO Contrastive Test Set (ES–EU) [Gete et al., 2022] addresses specific translation challenges between Basque and Spanish. It primarily focuses on the translation of pronouns, adjectives, and nouns that have gender-specific forms in Spanish but not in Basque, as well as nuances in formal and informal registers. These linguistic features require a deep understanding of context to ensure accurate and coherent translations. The dataset is organized into two main subsets, each consisting of context blocks from various domains like books, TED talks, and parliamentary proceedings. The contrastive translations were compiled using high-quality NMT models and were professionally post-edited. The final selections underwent manual review to verify the accuracy of the contextual information, ensuring that the datasets effectively challenge the translation models to handle complex, real-world linguistic variations. Overall, the test sets were balanced to mitigate potential biases in gender or register and to reflect a variety of contexts, including blocks with fewer than five preceding sentences.

Generative Adaptation of Contrastive Test Sets [15] While originally designed as discriminative contrastive test sets, Post and Junczys-Dowmunt [2023] adapted several of these contrastive test set, previously introduced, into generative versions. The core idea of this adaptation involves translating the source sentence and evaluating whether the generated output includes the correct target word or phrase (such as a pronoun, word sense, or gendered noun) without any of the incorrect alternatives. For some datasets, such as ContraPro (EN–DE and EN–FR), the correct target word is provided as an annotation, making this task straightforward. In cases where the correct and incorrect translations are not annotated, a phrase-level comparison between positive and negative examples is performed to identify the intended correct translation. The generative test score records success when the correct word or phrase appears in the system's output and none of the incorrect options are present, providing an accuracy measure that reflects the system's real-world generative abilities. This shift to a generative framework enables a more comprehensive evaluation of a system's ability to generate fluent and contextually accurate translations, moving beyond simple error detection to directly assess translation generation. Notably, this approach has been applied to a variety of datasets described above, including ContraPro (EN–DE, EN–FR), DiscEvalMT (EN–FR), and ContraWSD (DE–EN, DE–FR) and the contrastive test sets for deixis, ellipsis, lexical cohesion (EN–RU).

In summary, evaluations with test suites, particularly contrastive test suites, have been widely used across various contexts and language pairs to assess translation issues associated with specific discourse phenomena. However, many of them typically focus more on discrete cohesion elements rather than on exclusive coherence aspects. This emphasis arises from the inherent complexities of assessing coherence, which is significantly more challenging than assessing cohesion. The contrastive test sets presented are among the most cited in the literature, but they share common limitations. First, they concentrate on a narrow range of contrastive errors involving only a few words, as the segments evaluated are relatively short and consist of 2 or 3 sentences rather than complete paragraphs or documents. As a result, these tests may not adequately reflect global coherence errors related to the global context in document translation. Second, test suites are fixed in time, and as translation models evolve, particularly with the advent of LLMs, these test suites may become obsolete, and some translation systems might be trained using these same test suites. These issues suggest that contrastive evaluations, although locally useful, must be used in conjunction with other evaluation tools if one wants to obtain an overview of a document translation system's performance.

6.2 Automatic metrics for evaluating specific linguistic phenomena

We mentioned in Section 5.2 that global automatic metrics do not provide insights into how they handle coherence, which discourse phenomena they capture, and which they do not. In parallel to the development of these metrics, automatic evaluation metrics for specific document phenomena [Maruf et al., 2021], that could explicitly identify incoherent and incohesive elements in MT outputs, have been very employed by the MT community. In this section, we discuss automatic metrics for evaluating specific discourse phenomena recently developed or used in the MT literature. These metrics are complementary to the global metrics of Section 5.2 and can therefore be aggregated with them to incorporate specific document phenomena into a global evaluation score.

[15] https://github.com/marian-nmt/docmt23/tree/master

+RC and +LC Wong and Kit [2012] enhanced traditional sentence-level evaluation metrics, such as BLEU, METEOR, and TER by integrating lexical cohesion ratios calculated at the document level. The lexical cohesion devices analyzed involve the repetition of content words, such as nouns, adjectives, adverbs, and main verbs, within a document. This also includes the use of hypernyms and synonyms (detected using WordNet). Two ratio are computed: (i) LC (number of lexical cohesion devices divided by the total number of content words) and (ii) RC (number of repetitions divided by the number of content words). These ratios are then merged into the sentence-level metrics using a weighted average, improving their correlation with human judgments.

The authors examine, through experiments, the effectiveness of using LC and RC ratios when used alone or in conjunction with other evaluation metrics for MT evaluation at the document and system levels using the MetricsMATR [Dobrinkat et al., 2010] and MTC4 datasets [Ma, 2006] as evaluation data and Pearson's correlation coefficient to compare to the human judgment. They outlined the correlation rates of various evaluation metrics, the LC and RC metrics demonstrated strong correlations with human assessments at the system level, showing comparable results to metrics like BLEU and TER. However, at the segment-level (document-level), LC and RC were less effective on their own, yet seemed to help when integrated with BLEU (BLEU+RC and BLEU+LC) and TER. However the same is not observed for METEOR. The added value of these rations lies in its incorporation into a global metric. Although this does not seem to work for all metrics, the use of these ratios remains a strong baseline when introducing new metrics.

H- Gong et al. [2015] propose new metrics that incorporate document-level features, such as a "gist consistency" score and a text cohesion score, alongside existing evaluation metrics like BLEU and METEOR, to measure text cohesion. "Gist consistency" measures how well the main topics of the MT output align with those in the reference text. This is achieved using a topic model trained via Latent Dirichlet Allocation (LDA) [Blei et al., 2001], where the "document-topic" distributions of both the MT output and reference texts are compared using Kullback-Leibler (KL) divergence [Shannon, 1948]. The score is then calculated. Text cohesion is assessed using simplified lexical chains, which track the repetition of content words across sentences in both the MT output and the references. The cohesion score accounts for the correct and incorrect repetition of words, with the final score averaged across all chains. The combined metric, denoted as H, is defined as a weighted average of a document-level extension of BLEU or METEOR (e.g., d-BLEU [5.2.1]) and the gist consistency score or text cohesion score.

The authors assess H-BLEU and H-METEOR by evaluating their correlation with human judgments on two datasets: MTC2 and MTC4. Since document-level human assessments were not available, the study averages sentence-level human scores, weighted by sentence length, to approximate document-level assessments. The performance of the H-metrics is evaluated using Pearson and Kendall correlation coefficients, with results showing that incorporating gist consistency improves the correlation with human judgments. However, the impact of text cohesion was less pronounced than gist consistency. The authors highlight that the improvements in METEOR are likely due to its ability to utilize synonym information, which aligns with the LDA model's focus on major topics. The results suggest that while both measures improve MT evaluation metrics, gist consistency is more effective than text cohesion.

Term Consistency Metric [Semenov and Bojar, 2022] present a term consistency metric for evaluating the consistency and unambiguity of terminology in MT, particularly in professional domains like legal texts. This metric assesses (i) consistency—ensuring each source term is translated into the same target term throughout the document—and (ii) unambiguity, where different source terms map to distinct target terms. The evaluation process involves automatic term extraction from the source text and alignment with their translations using alignment algorithms. The translated terms are compared to "pseudo-references," which can either be the first occurrence of the term or its most frequent translation in the document. The consistency and unambiguity of the translations are then evaluated through – multiclass precision, recall, true positive rate, and other data science metrics applied to term occurrences. They also define their "own" evaluation score by grouping the lists by the source terms and counting the percentage of the correct occurrences of the exact term. This allows the metric to capture both how consistently terms are translated and whether each source term is mapped to a distinct target term. The metric was tested on the Czech-to-English ELITR agreement corpus and evaluated several MT systems from WMT21 and WMT22. Results showed that this metric provided a better assessment of document-level term consistency compared to traditional sentence-level metrics like BLEU or chrF, showing higher correlation with human judgment at the document level. While it does not explicitly address translation adequacy, which

is left to mainstream metrics, it offers significant insights into the handling of domain-specific terminology, making it valuable for specialized contexts.

BlonDe BlonDe (Bilingual Evaluation of Document Translation) [Jiang et al., 2022] is a surface-form document-level automatic metric that focuses on discourse phenomena. Distinct from traditional metrics, BlonDe assesses translations by computing a similarity-based F1 measure across discourse-related spans, focusing on discourse coherence. BlonDe's approach involves categorizing spans that embody specific discourse phenomena, such as inconsistencies in named entity translation, verb tense, pronoun accuracy, discourse markers, and lexical accuracy. The metric automatically annotates these spans both in the target and reference texts. BlonDe is then defined based on automatic comparisons between the vector of spans in the hypothesis and in the reference, measured through precision and recall scores, culminating in an F1 measure that provides a view of translation quality. BlonDe is evaluated using the BWB test set (Section 5.1.2), where it is compared against standard metrics like BLEU and METEOR. This evaluation shows that BlonDe scores exhibit a good trend from sentence level to document-level, unlike the mostly linear trends observed in standard metrics. However, the authors did not compare BlonDe with SOTA metrics at the time of publication. [Vernikos et al., 2022] re-evaluated BlonDe by comparing it to their own metric designed at the document level (Section 5.2.3), and showed its lack of performance on more widespread context-aware test sets.

DiscoScore DiscoScore [Zhao et al., 2023] is a parametrized reference-based metric, used for MT evaluation, which uses a pre-trained BERT to model discourse coherence through the lens of readers' focus, based on the Centering theory [Grosz et al., 1995]. The authors define two variants that can be distinguished in how they use focus.[16] First, they model focus frequency and semantics and compare their difference between hypothesis and reference ("Focus Difference" and the associated metric DS-FOCUS). Second, they use focus transitions to model the interdependence between sentences through "Sentence Graphs" and the associated metric DS-SENT. Building upon this, they present a graph-based approach to compare hypotheses with references.

- "Focus Difference" studies suggest that focus transitions signal text coherence [Guinaudeau and Strube, 2013]. A coherent hypothesis should have multiple overlaps with the reference in terms of focus, with these overlaps being similar in meaning and frequency. The DS-FOCUS models relies on semantics and frequency to compare a hypothesis with a reference. It uses a bipartite graph where one set of vertices represents the foci and the other represents tokens in the text. An adjacency matrix links these foci to their corresponding tokens. Contextualized encoders like BERT generate token embeddings, which are then summed to create focus embeddings for both the hypothesis and the reference. The DS-FOCUS score then measures the distance between the embeddings of common foci in the hypothesis and reference, scaled by the number of foci in the hypothesis. This score provides an indication of how closely the hypothesis aligns with the reference in terms of focus.
- The "Sentence Graph" concept produces sentence embeddings within a document's context, noting that many contextualized encoders fail to model the interdependence between sentences. Drawing on Centering theory, it suggests that two sentences are considered continuous in meaning if they share at least one focus. Conversely, a meaning shift occurs when no focus is shared. DS-SENT compares sentence embeddings between a hypothesis and a reference by constructing a graph based on sentence connectivity. The method uses an adjacency matrix that reflects shared foci between sentences and their distances. The DS-SENT score is then calculated by measuring the cosine similarity between the hypothesis and reference graph embeddings, assessing the alignment of sentence structures.

In this work, the authors investigate two choices for the focus: nouns and semantic entities (lexical cohesion device in the form of repetition of entities) as exposed in Figure 5. They assess DiscoScore using WMT20 document-level data and find that BERT-based metrics are generally weak in system-level correlation with human ratings in terms of coherence. In contrast, DiscoScore strongly correlates well with human-rated coherence, outperforming BARTScore [Yuan et al., 2024] and two baselines, RC and LC. Although DiscoScore focuses solely on coherence, the authors

[16] The formal definition of focusing in discourse is given at two levels [Grosz, 1977]:

1. global focus on entities relevant to the overall discourse
2. local focus on the specific entity most relevant to an utterance.

Hypothesis

Chelsea have made an offer for FC Tokyo forward Yoshinori Muto. The 22-year-old will join Chelsea 's Dutch partner club Vitesse Arnhem on loan next season if he completes a move to Stamford Bridge. Chelsea signed a £200million sponsorship deal with Japanese company Yokohama Rubber in February.

Reference

Naoki Ogane says that Chelsea have made an offer for Yoshinori Muto. The 22-year-old forward has one goal in 11 games for Japan. Muto admits that it is an 'honour' to receive an offer from the Blues. Chelsea have signed a £200m sponsorship deal with Yokohama Rubber. Muto graduated from university with an economics degree two weeks ago. He would become the first Japanese player to sign for Chelsea.

	t_1	t_2	t_3	t_4	t_5	...
Chelsea	1	0	0	0	0	1
offer	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) FocusDiff

	s_1	s_2	s_3
s_1	0	1	0.5
s_2	0	0	1
s_3	0	0	0

(b) SentGraph

Figure 5: Borrowed from Zhao et al. [2023]. Sample hypothesis and reference are from SUMMEval. Each focus is marked in a different color, corresponding to multiple tokens as instances of a focus. Foci shared in Hypothesis and Reference are marked in the same color. (a)+(b) are adjacency matrices used to model focus-based coherence for Hypothesis; for simplicity, adjacency matrices for Reference are omitted. FocusDiff and SentGraph are variants of DiscoScore. For FocusDiff, they use (a) to depict the relations between foci and tokens, reflecting focus frequency. For SentGraph, they use (b) to depict the interdependence between sentences according to the number of foci shared between sentences and the distance between sentences.

note that combining DS-FOCUS with BERTScore brings a small improvement in correlating with adequacy. This underscores the importance of including discourse signals in the assessment of system outputs, as the discourse features derived from DiscoScore can effectively differentiate between a hypothesis and a reference. A limitation of this work is that they only investigate two types of foci — nouns and semantic entities — without considering other cohesion devices (Section 3.2). Furthermore, the construction of the Discoscore is geared towards the assessment of English as a target language.

7 Meta-evaluation

In this section, we turn to methods aimed at evaluating and comparing the performance and validity of metrics, a process known as *meta-evaluation*. Its main objective is to determine the ability of an automatic evaluation metric to replicate or replace human evaluation, which, as described above, is considered the gold standard. The most commonly used approach to evaluate metrics is to compute the agreement (or the correlation) between human judgments of translation quality and the scores assigned by the metrics to the same set of translations.

7.1 Computing correlations with human judgments

Metrics can either be meta-evaluated at the system- or segment-level. System-level correlations are calculated between corpus-level scores (one per MT system, averaged over the corpus' segments), whereas segment-level correlations evaluate metric scores for individual segments rather than on aggregated system scores. The correlation can be estimated using different correlation coefficients such as:

- Pearson coefficient, which captures a linear correspondence between two input vectors, defined as their covariance divided by the product of their variances.
- Spearman coefficient, which is equivalent to Pearson applied to the ranks of the inputs.
- Kendall rank coefficient, which is a type of rank correlation coefficient that counts how frequently the metric and human scores agree (i.e. are concordant) or disagree (i.e. are discordant) when ranking pairs of translations.

In the WMT20 metrics shared task [Mathur et al., 2020], a document-level correlation using the DOC-DA score (Section 4) was defined to explicitly differentiate segment-level approaches, which use sentences as segments, from the document-level approaches, which use documents as segments. An alternative way to measure correlation between automatic metrics and human judgments is through Pairwise Accuracy (PA) [Kocmi et al., 2021]. PA gauges the extent to which a metric ranks system pairs in the same order as human judges, relative to the total number of system pairs in the dataset. Recent work by [Mathur et al., 2020, Kocmi et al., 2021] underscores that the primary application of a metric is often to choose between competing systems. Thus, an effective metric should produce pairwise rankings that closely align with those made by humans. This perspective has led to the adoption of PA, or extension like Soft Pairwise Accuracy (SPA) [Thompson et al., 2024], in recent WMT Metrics Shared Tasks [Freitag et al., 2021b, 2022, 2023, 2024]. With the shift to document-level evaluations, there has been no clear change in the correlation methods compared to those used at the sentence level. Human scores have been adjusted (Section 4), but they are still processed by the same meta-evaluation "metrics". Although a document-level correlation was introduced in Mathur et al. [2020] (and later abandoned), this was more about providing a clear formulation to distinguish a sentence from a paragraph or document, but it remains the same procedure used at the sentence level. Perrella et al. [2024] highlight issues in meta-evaluation methods, noting that fine-tuned metrics often exploit "spurious correlations"—patterns tied more to source text structure than to translation quality. For example, these metrics may learn to penalize segments with many proper nouns or numbers, reflecting biases from observed human scores rather than true translation quality. The authors show that such biases are reinforced by segment-level meta-scoring methods and pose particular challenges for document-level evaluation, where specific linguistic features such as coherence, cohesion, and length introduce unique complexities, as discussed in Section 3. Although the study did not cover meta-evaluation at the document level (focusing only on discrete features of it) and its insights only exposed weaknesses in fine-tuned metrics, it underscores the need to adapt meta-evaluation methods for document-level metrics, as applying sentence-level practices risks misrepresenting true performance and reinforcing irrelevant correlations. Tailored approaches are essential for accurately meta-evaluating document-level metrics.

Currently, at the sentence level, neural-based metrics, particularly XCOMET-Ensemble [Guerreiro et al., 2023] MetaMetrics-MT [Anugraha et al., 2024] and MetricX-24-Hybrid [Juraska et al., 2024], are better correlated with human judgments than non-neural metrics, although performance varies across language directions and is also notably affected by the quality of the reference translations. Reference-free metrics also performed well, especially in cases where reference quality was low. For a detailed analysis of the performance of metrics, we refer readers to the WMT23 and WMT24 Metrics Shared Task paper [Freitag et al., 2023, 2024] and the survey of Lee et al. [2023].

7.2 Evaluating metrics using challenge sets

An alternative to correlation analysis is the use of challenge sets, in the same spirit as the assessment of MT systems using challenge sets (see Section 6.1). This method provides a more fine-grained analysis of the strengths and weaknesses of the metrics, and aids in understanding metrics' sensitive to certain types of translation error (for specific phenomena). Metric evaluation using challenge sets was introduced in the WMT21 metrics shared task [Freitag et al., 2021b] to target specific linguistic phenomena, such as sentiment polarity, antonym replacement, and named entities. Each challenge set is a contrastive test set consisting of two MT outputs (along with the corresponding source and reference), one of which contains the translation error of interest, and the other of which does not. A metric that is sensitive to the targeted phenomenon should give a lower score to the erroneous MT output. Kendall's tau-like correlation is used to estimate the metric performance, checking the number of times a metric assigns a higher score to the MT output without the error and vice versa.

There are no challenge sets specifically designed for the document (or paragraph) level. Although some datasets encompass various linguistic phenomena, they include sections that address document-level linguistic features, even if these are not the primary focus of the dataset. For example, a DE–EN challenge set was manually designed for the WMT22 shared task, based on the semi-automatic test suite TQ-AutoTest [Macketanz et al., 2018], including assessments of discourse-level phenomena, such as coordination and ellipsis, named entities, terminology, long-distance dependencies and interrogatives. Their findings indicate that embedding-based metrics such as BLEURT-20 and COMET-MQM-2021 seem to be less sensitive to these phenomena. Similarly, ACES (Translation Accuracy Challenge Sets) [Amrhein et al., 2022], introduced in WMT22 and reused in WMT23, also covered discourse-level errors, including pronoun usage, connectives, co-reference, and ambiguity. ACES, automatically constructed and are heavily influenced by the MQM framework, comprises 36,499 examples, spanning 146 language pairs and covering 68 phenomena

(3,698 examples specifically addressing discourse-level phenomena). The results, averaged across these different phenomena, indicated that XCOMET-Ensemble achieved the best performance for sensitivity to discourse-level errors.

In addition to these specific challenge sets, some contrastive test sets designed to evaluate MT systems can also be used to assess MT metrics. For instance, Vernikos et al. [2022] used ContraPro (EN–DE), ContraPro (EN–FR), and DiscEvalMT to meta-evaluate doc-COMET-QE (in comparison to COMET-QE) by confirming gains achieved through the use of document-level context. In this evaluation, each sentence in the contrastive test set was treated as a different hypothesis, and the researchers measured how often doc-COMET-QE ranked the correct translation higher. Since reference-based metrics would trivially succeed in such tasks by relying on the reference, the authors was limited on source-based metrics like COMET-QE, evaluating how well these metrics perform without relying on a reference translation.

However, it should be noted that the segments in those test sets, mentioned above, remain relatively short sentences, which makes it challenging for sentence-level data to adequately model discourse-level errors present in documents.

8 Conclusion

In this survey, we explored various aspects of document-level evaluation for MT, examining the techniques and challenges associated with assessing translation quality beyond the sentence level. We began by discussing key document-level linguistic phenomena, such as coherence and cohesion, which become increasingly important as text length grows. While research has made strides in addressing these challenges, most efforts remain focused on local phenomena that require additional context, often isolating linguistic issues without fully considering their interaction with other document-level traits. This compartmentalization provides a solution, but at the cost of overlooking the broader question: What is the quality of a translated document [Scarton and Specia, 2016]? Human evaluation techniques for document-level MT have evolved, particularly in their focus on context-aware and document-aware evaluations, which are seen as the gold standard. We observed the shift from DA to more fine-grained approaches like MQM, which allows for a more nuanced evaluation of segments within their broader context but which is more time-consuming [Kocmi et al., 2024] and highlighting alternatives such as ESA. Nevertheless, even with these advances, traditional evaluation methods are still adapting to document-level needs.As we discussed, global automatic metrics remain a foundation of translation evaluation. Their primary function—evaluating a translation or an entire system—has expanded into the document-level space. However, not all metrics have received the same level of attention. Surface-level metrics, such as BLEU, have shown limited efficacy for document evaluation, which has driven research toward neural metrics, especially fine-tuned metrics that leverage available labeled datasets. However, document-level corpora aligned with human evaluation remain scarce, limiting the advancement of these metrics. Emerging approaches, like prompt-based metrics, offer promising results but have yet to be fully tested at the document level. Parallel to these developments, targeted approaches for evaluating specific linguistic phenomena, such as contrastive test sets and automatic metrics, offer valuable insights. These approaches complement global metrics, which often struggle with errors related to specific linguistic challenges. Yet, the complexity of implementing these targeted evaluations, particularly contrastive test sets, often requires manual effort and can be difficult to scale.

The problems in document-level MT evaluation are far from solved. While notable progress has been made, from improving human evaluation methods to developing automatic metrics, much remains to be redefined and extended. A significant gap persists between context-aware evaluation methods and a truly document-level evaluation framework. Current approaches rely heavily on sentence-level structures, leaving open questions about how small paragraphs compare to large documents in terms of evaluation accuracy. One major challenge is the adaptation of human evaluation methods to better accommodate document-level traits, which could, in turn, pave the way for more adaptable automatic metrics. Though recent advances are promising, none have proven entirely convincing, and fundamental questions about the nature of evaluation persist. A key consideration in this ongoing discussion is the distinction between what we aim to measure. Should we continue to use global scores that also reflect local properties, such as fluency and fidelity, or should we develop metrics that focus solely on global document-level traits like cohesion and coherence, assuming that fluency and fidelity can be captured with more local contexts? The current reliance on human metrics that aggregate local scores raises the question of whether a more refined, purely document-level evaluation framework is possible.

New methodologies, such as question-answer metrics, gap-filling tasks [Forcada et al., 2018, Han

et al., 2022] or automatic error explanation using large language models [Treviso et al., 2024], offer a potential shift in how we evaluate document-level translations, but they have yet to be fully adapted for automatic evaluation. These approaches have been explored minimally in a human evaluation protocol [Scarton and Specia, 2016] but could also provide valuable insights applied automatically at the document level. Additionally, we may need to rethink the purpose of translation before refining our evaluation frameworks. Translation requirements vary depending on the context: whether the goal is to provide a gist for casual understanding, to do data filtering or to achieve precise, nuanced translations for high-stakes settings like the translation of contracts or other technical contents. These various purposes would naturally necessitate different evaluation criteria. Ultimately, the development of more robust document-level evaluation methods will depend on deeper theoretical reflection and innovative practical approaches. As MT continues to advance, the question of how best to assess its outputs at the document level remains an open and critical area for future research.

Bibliography

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

Sweta Agrawal, António Farinhas, Ricardo Rei, and André F. T. Martins. Can automatic metrics assess high-quality translations?, 2024. URL https://arxiv.org/abs/2405.18348.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.1.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.44.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam

Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 Technical Report, 2023. URL https://arxiv.org/abs/2305.10403.

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.32.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6402. URL https://aclanthology.org/W18-6402.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.1.

Rachel Bawden, Thomas Lavergne, and Sophie Rosset. Detecting context-dependent sentences in parallel corpora. In Pascale Sébillot and Vincent Claveau, editors, *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 393–400, Rennes, France, 5 2018a. ATALA. URL https://aclanthology.org/2018.jeptalnrecital-court.22.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118. URL https://aclanthology.org/N18-1118.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5403. URL https://aclanthology.org/W19-5403.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. Findings of the WMT 2023 shared task on quality estimation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.52. URL https://aclanthology.org/2023.wmt-1.52.

David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:601–608, 01 2001.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pages 27–34, May 2016. URL http://www.cracking-the-language-barrier.eu/mt-eval-workshop-2016/. LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem", LREC 2016 ; Conference date: 24-05-2016 Through 24-05-2016.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6401. URL https://aclanthology.org/W18-6401.

Maud Bénard, Natalie Kübler, Alexandra Mestivier, Joachim Minder, and Lichao Zhu. Étude des protocoles d'Évaluation humaine pour la traduction de documents. Technical report, Projet ANR MaTOS, livrable D4-1.1, 2024. URL https://hal.science/hal-04700009.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In Diana McCarthy and Shuly Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1007.

Sheila Castilho. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.137.

Sheila Castilho. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.4.

Sheila Castilho and Rebecca Knowles. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, pages 1–31, 2024. doi: 10.1017/nlp.2024.7.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. Approaches to human and machine translation quality assessment. In *Translation quality assessment*, pages 9–38. Springer, 2018.

Sheila Castilho, Maja Popović, and Andy Way. On context span needed for machine translation evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.461.

Sheila Castilho, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. DELA corpus - a document-level corpus annotated with context-related issues. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 566–577, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.63.

Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: Web inventory of transcribed and translated talks. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way, editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, May 28–30 2012. European Association for Machine Translation. URL https://aclanthology.org/2012.eamt-1.60.

Eirini Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161, 2020.

Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3346. URL https://aclanthology.org/W14-3346.

Hsin-Hsi Chen, Guo-Wei Bian, and Wen-Cheng Lin. Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 215–222, College Park, Maryland, USA, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034717. URL https://aclanthology.org/P99-1028.

Michael Denkowski and Alon Lavie. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA, October 31-November 4 2010. Association for Machine Translation in the Americas. URL https://aclanthology.org/2010.amta-papers.20.

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

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. Training and meta-evaluating machine translation evaluation metrics at the paragraph level, 2023b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Marcus Dobrinkat, Tero Tapiovaara, Jaakko Väyrynen, and Kimmo Kettunen. Normalized compression distance based measures for MetricsMATR 2010. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors, *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 343–348, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-1752.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *North American Chapter of the Association for Computational Linguistics*, 2013. URL https://api.semanticscholar.org/CorpusID:8476273.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.100. URL https://aclanthology.org/2023.wmt-1.100.

Mikel L. Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 192–203, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6320. URL https://aclanthology.org/W18-6320.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021a. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.73.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.2.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.2.

Jianfeng Gao and Xiaodong He. Training MRF-based phrase translation models using gradient ascent. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 450–459, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1048.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166, May 2023. ISSN 1076-9757. doi: 10.1613/jair.1.13715. URL https://www.jair.org/index.php/jair/article/view/13715.

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

Zhengxian Gong, Min Zhang, and Guodong Zhou. Document-level machine translation evaluation with gist consistency and text cohesion. In Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier, editors, *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2504. URL https://aclanthology.org/W15-2504.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. Is all that glitters in machine translation quality estimation really gold? In Yuji Matsumoto and Rashmi Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1294.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. Improving evaluation of document-level machine translation quality estimation. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 356–361, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2057.

Yvette Graham, Barry Haddow, and Philipp Koehn. Translationese in machine translation evaluation, 2019. URL https://arxiv.org/abs/1906.09833.

Yvette Graham, Christian Federmann, Maria Eskevich, and Barry Haddow. Assessing human-parity in machine translation on the segment level. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4199–4207, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.375. URL https://aclanthology.org/2020.findings-emnlp.375.

Yvette Graham, Barry Haddow, and Philipp Koehn. Statistical power and translationese in machine translation evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.6. URL https://aclanthology.org/2020.emnlp-main.6.

Barbara J. Grosz. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'77, page 67–76, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995. URL https://aclanthology.org/J95-2003.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection, 2023. URL https://arxiv.org/abs/2310.10482.

Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1010.

M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Routledge, 1976.

HyoJung Han, Marine Carpuat, and Jordan Boyd-Graber. SimQA: Detecting simultaneous MT errors through word-by-word question answering. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.378. URL https://aclanthology.org/2022.emnlp-main.378.

Sabien Hanoulle. *Translating documentaries: does the integration of a bilingual glossary of domain-specific terminology into the translation process reduce the translators' workload?* PhD thesis, Ghent University, 2017.

Christian Hardmeier. Discourse in statistical machine translation: A survey and a case study. *Discours*, 12 2012. doi: 10.4000/discours.8726.

Christian Hardmeier and Marcello Federico. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 283–289, Paris, France, December 2-3 2010. URL https://aclanthology.org/2010.iwslt-papers.10.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation, 2023.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. Lost in the source language: How large language models evaluate the quality of machine translation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.211. URL https://aclanthology.org/2024.findings-acl.211.

Md Mahfuz ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. On the evaluation of machine translation for terminology consistency, 2021. URL https://arxiv.org/abs/2106.11891.

Pierre Isabelle, Colin Cherry, and George Foster. A challenge set approach to evaluating machine translation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1263. URL https://aclanthology.org/D17-1263.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.16. URL https://aclanthology.org/2020.eval4nlp-1.16.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.111. URL https://aclanthology.org/2022.naacl-main.111.

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

Josef Jon. Exploring contextual information in neural machine translation. Master's thesis, Brno University of Technology, Faculty of Information Technology, 2019. URL https://www.fit.vut.cz/study/thesis/21979/.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.63. URL https://aclanthology.org/2023.wmt-1.63.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.35.

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

Saeed Ketabi and Ali Asghar Jamalvand. A corpus-based study of conjunction devices in english international law texts and its farsi translation. *International Journal of Linguistics*, 4(4):362, 2012.

Payal Khullar. Are ellipses important for machine translation? *Computational Linguistics*, 47(4): 927–937, December 2021. doi: 10.1162/coli_a_00414. URL https://aclanthology.org/2021.cl-4.30.

Margaret King and Kirsten Falkedal. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*, 1990. URL https://aclanthology.org/C90-2037.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez

Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June 2023b. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.57.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.1.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. Error span annotation: A balanced approach for human evaluation of machine translation, 2024. URL https://arxiv.org/abs/2406.11580.

P. Koehn. *Statistical Machine Translation*. Statistical Machine Translation. Cambridge University Press, 2010. ISBN 9780521874151. URL https://books.google.co.il/books?id=4v_Cx1wIMLkC.

Philip Koehn. *Neural Machine Translation*. Cambridge University Press, 2020.

Monika Krein-Kühle. Cohesion and coherence in technical translation: The case of demonstrative reference. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 1, 2002.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://aclanthology.org/D18-1512.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4), 2023. ISSN 2227-7390. doi: 10.3390/math11041006. URL https://www.mdpi.com/2227-7390/11/4/1006.

Yikun Lei, Yuqi Ren, and Deyi Xiong. CoDoNMT: Modeling cohesion devices for document-level neural machine translation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5205–5216, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.462.

Jindřich Libovický, Thomas Brovelli, and Bruno Cartoni. Machine translation evaluation beyond the sentence level. In Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Maja Popović, Celia Rico, André Martins, Joachim Van den Bogaert, and Mikel L. Forcada, editors, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 199–208, Alicante, Spain, May 2018. URL https://aclanthology.org/2018.eamt-main.18.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219032. URL https://aclanthology.org/P04-1077.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1275.

Qi Liu, Matt J. Kusner, and Phil Blunsom. A survey on contextual embeddings, 2020a. URL https://arxiv.org/abs/2003.07278.

Siyou Liu and Xiaojun Zhang. Corpora for document-level neural machine translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.466.

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

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12 2014. doi: 10.5565/rev/tradumatica.77.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6.

Allíson Beeby Lonsdale. *Cohesion and Coherence*, pages 215–230. University of Ottawa Press, 1996. URL http://www.jstor.org/stable/j.ctt1cn6sgb.21.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level neural MT: A systematic comparison. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco

Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.24.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. Error analysis prompting enables human-like translation evaluation in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.520. URL https://aclanthology.org/2024.findings-acl.520.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67, March 2020. ISSN 1076-9757. doi: 10.1613/jair.1.11371. URL http://dx.doi.org/10.1613/jair.1.11371.

Xiaoyi Ma. Multiple-Translation Chinese (MTC) Part 4 LDC2006T04. Web Download, 2006.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. TQ-AutoTest – an automated test suite for (machine) translation quality. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1142.

Belinda Maia. Terminology and translation — bringing research and professional training together through technology. *Meta*, 50(4), 2005. doi: https://doi.org/10.7202/019921ar.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation, 2021. URL https://arxiv.org/abs/1912.08494.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. Results of the WMT20 metrics shared task. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.77.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA, October 24-25 2005. URL https://aclanthology.org/2005.iwslt-1.19.

Alan Melby and Christopher Foster. Context in translation: Definition, access and teamwork. *The International Journal for Translation & Interpreting Research*, 2, 11 2010.

Elise Michon, Josep Crego, and Jean Senellart. Integrating domain terminology into neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.348. URL https://aclanthology.org/2020.coling-main.348.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://aclanthology.org/D18-1325.

Lesly Miculicich Werlen and Andrei Popescu-Belis. Validation of an automatic metric for the accuracy of pronoun translation (APT). In Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors, *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4802. URL https://aclanthology.org/W17-4802.

Vipul Mishra, Chenhui Chu, and Yuki Arase. Lexically cohesive neural machine translation with copy mechanism. *CoRR*, abs/2010.05193, 2020. URL https://arxiv.org/abs/2010.05193.

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

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6307. URL https://aclanthology.org/W18-6307.

Jonathan Mutal, Johanna Gerlach, Pierrette Bouillon, and Hervé Spechbach. Ellipsis translation for a medical speech to speech translation system. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 281–290, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.30.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193: 217–250, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.07.001. URL https://www.sciencedirect.com/science/article/pii/S0004370212000793.

Aurélie Névéol, Antonio Jimeno Yepes, and Mariana Neves. MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors' abstract writing practice. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3676–3682, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.453.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1470.

P. Newmark. *A Textbook of Translation*. English language teaching. Prentice-Hall International, 1988. ISBN 9780139125935. URL https://books.google.co.il/books?id=ABpmAAAAMAAJ.

B. Oommen and I. Reichstein. On translating ellipses amidst elliptic obstacles. In *Proceedings. 1986 IEEE International Conference on Robotics and Automation*, volume 3, pages 1755–1760, 1986. doi: 10.1109/ROBOT.1986.1087425.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny

Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

Hanting Pan. Translating conjunctive cohesion in legal documents. *Perspectives*, 22(1):1–20, 2014.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Ziqian Peng, Rachel Bawden, and François Yvon. À propos des difficultés de traduire automatiquement de longs documents. In Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-Dac, Julie Mauclair, Jose G. Moreno, and Julien Pinquier, editors, *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*, volume 1 : articles longs et prises de position, pages 2–21, Toulouse, France, July 2024a. ATALA & AFPC. URL https://inria.hal.science/hal-04623006.

Ziqian Peng, Rachel Bawden, and François Yvon. Handling Very Long Contexts in Neural Machine

Translation: a Survey. Technical Report Livrable D3-2.1, Projet ANR MaTOS, June 2024b. URL https://inria.hal.science/hal-04652584.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. MaTESe: Machine translation evaluation as a sequence tagging problem. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.51.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.856. URL https://aclanthology.org/2024.acl-long.856.

Andrei Popescu-Belis. Context in neural machine translation: A review of models and evaluations. Arxiv preprint 1901.09115., 2019. URL https://arxiv.org/abs/1901.09115.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Maja Popović. Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 translation hypotheses. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5353. URL https://aclanthology.org/W19-5353.

Maja Popović and Sheila Castilho. Are ambiguous conjunctions problematic for machine translation? In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 959–966, Varna, Bulgaria, September 2019. INCOMA Ltd. doi: 10.26615/978-954-452-056-4_111. URL https://aclanthology.org/R19-1111.

Matt Post and Marcin Junczys-Dowmunt. Escaping the sentence-level paradigm in machine translation. Arxiv preprint 2304.12959, 2023. URL https://arxiv.org/abs/2304.12959.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. Penn discourse treebank version 3.0. Online, 2019. URL https://catalog.ldc.upenn.edu/LDC2019T05.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5354. URL https://aclanthology.org/W19-5354.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. Do GPTs produce less literal translations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*

Papers), pages 1041–1050, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.90. URL https://aclanthology.org/2023.acl-short.90.

Vikas Raunak, Tom Kocmi, and Matt Post. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 205–211, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.18. URL https://aclanthology.org/2024.naacl-short.18.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Unbabel's participation in the WMT20 metrics shared task. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online, November 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.101.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. Finding replicable human evaluations via stable ranking probability. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4908–4919, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.275. URL https://aclanthology.org/2024.naacl-long.275.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. Finding replicable human evaluations via stable ranking probability, 2024b. URL https://arxiv.org/abs/2404.01474.

Wayne Rimmer. Beyond the Sentence: Introducing Discourse Analysis Grammar. *ELT Journal*, 60(4):392–394, 10 2006. ISSN 0951-0893. doi: 10.1093/elt/ccl033. URL https://doi.org/10.1093/elt/ccl033.

Annette Rios, Mathias Müller, and Rico Sennrich. The word sense disambiguation test suite at WMT18. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6437. URL https://aclanthology.org/W18-6437.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. Improving word sense disambiguation in neural machine translation with sense embeddings. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4702. URL https://aclanthology.org/W17-4702.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. A test suite and manual evaluation of document-level NMT at WMT19. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5352. URL https://aclanthology.org/W19-5352.

Benoît Sagot. The lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/701_Paper.pdf.

Carolina Scarton and Lucia Specia. A reading comprehension corpus for machine translation evaluation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1579.

Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. Searching for context: a study on document-level labels for translation quality estimation. In İlknur Durgar El-Kahlout, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hollowood, and Andy Way, editors, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 121–128, Antalya, Turkey, May 2015. URL https://aclanthology.org/W15-4916.

F.N. Scott, J.V. Denny, and J.V. Denney. *Paragraph-writing: A Rhetoric for Colleges*. Allyn and Bacon, 1909. URL https://books.google.co.il/books?id=vvoAAAAAYAAJ.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning robust metrics for text generation. arxiv preprint 2004.04696, 2020. URL https://arxiv.org/abs/2004.04696.

Kirill Semenov and Ondřej Bojar. Automated evaluation metric for terminology consistency in MT. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.41.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. Findings of the WMT 2023 shared task on machine translation with terminologies. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors,

Proceedings of the Eighth Conference on Machine Translation, pages 663–671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.54. URL https://aclanthology.org/2023.wmt-1.54.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Karin Sim Smith. On integrating discourse in machine translation. In Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors, *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4814. URL https://aclanthology.org/W17-4814.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. The trouble with machine translation coherence. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 178–189, 2016. URL https://aclanthology.org/W16-3407.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. *Quality estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2017.

Felix Stahlberg and Shankar Kumar. Jam or cream first? modeling ambiguity in neural machine translation with SCONES. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4950–4961, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.365. URL https://aclanthology.org/2022.naacl-main.365.

Dario Stojanovski and Alexander Fraser. Improving anaphora resolution in neural machine translation using curriculum learning. In Mikel Forcada, Andy Way, Barry Haddow, and Rico Sennrich, editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 140–150, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6614.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1: 5–21, 2020. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2020.11.001. URL https://www.sciencedirect.com/science/article/pii/S2666651020300024.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning, 2020. URL https://arxiv.org/abs/2008.00401.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. Exploring document-level literary machine translation with parallel paragraphs from world literature. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.672. URL https://aclanthology.org/2022.emnlp-main.672.

Brian Thompson and Matt Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.8. URL https://aclanthology.org/2020.emnlp-main.8.

Brian Thompson and Matt Post. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark

Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online, November 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.67.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.118.

S. Thornbury. *Beyond the Sentence: Introducing Discourse Analysis*. Macmillan books for teachers. Macmillan Education, 2005. ISBN 9781405064071. URL https://books.google.fr/books?id=SQC2PwAACAAJ.

Antonio Toral. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.20.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6312. URL https://aclanthology.org/W18-6312.

Marcos Treviso, Nuno M. Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André F. T. Martins. xTower: A Multilingual LLM for Explaining and Correcting Translation Errors, 2024. URL https://arxiv.org/abs/2406.19482.

Muriel Vasconcellos. Cohesion and coherence in the presentation of machine translation products. In *Proceedings of the Georgetown University Round Table on Languages and Linguistics*, 1989. URL https://api.semanticscholar.org/CorpusID:35404104.

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

Rob Voigt and Dan Jurafsky. Towards a literary machine translation: The role of referential cohesion. In David Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors, *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2503.

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

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. UniTE: Unified translation evaluation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.558. URL https://aclanthology.org/2022.acl-long.558.

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

Warren Weaver. Translation. In *Proceedings of the Conference on Mechanical Translation*, Massachusetts Institute of Technology, 17-20 June 1952. URL https://aclanthology.org/1952.earlymt-1.1.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.2.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA, October 5-8 1994. URL https://aclanthology.org/1994.amta-1.25.

Rachel Wicks and Matt Post. Does sentence segmentation matter for machine translation? In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.78.

Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1097.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, Shimin Tao, Hao Yang, and Yanfei Jiang. Empowering a metric with LLM-assisted named entity annotation: HW-TSC's submission to the WMT23 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 822–828, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.70. URL https://aclanthology.org/2023.wmt-1.70.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lv, and Qun Liu. Modeling lexical cohesion for document-level machine translation. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024. URL https://arxiv.org/abs/2401.08417.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTSCORE: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393.

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

Ran Zhang, Wei Zhao, and Steffen Eger. How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs, 2024. URL https://arxiv.org/abs/2410.18697.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=SkeHuCVFDr.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, 2020b. URL https://arxiv.org/abs/1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1053. URL https://aclanthology.org/D19-1053.

Wei Zhao, Michael Strube, and Steffen Eger. DiscoScore: Evaluating text generation with BERT and discourse coherence. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.278. URL https://aclanthology.org/2023.eacl-main.278.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. WMT20 document-level markable error exploration. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 371–380, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.41.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. Fine-tuned machine translation metrics struggle in unseen domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.45. URL https://aclanthology.org/2024.acl-short.45.