Evaluation and large-scale training for contextual machine translation

Matt Post and **Marcin Junczys-Dowmunt**
Microsoft
{mattpost,marcinjd}@microsoft.com

Abstract

Despite the fact that context is known to be vital for resolving a range of translation ambiguities, most traditional machine translation systems continue to be trained and to operate at the sentence level. A common explanation is the lack of document-level annotations for existing training data. This work investigates whether having such annotations would be helpful for training traditional MT systems at scale. We build large-scale, state-of-the-art contextual MT systems into German, French, and Russian, fixing the datasets while comparing the effect of sourcing contextual training samples from both parallel and back-translated data. We then evaluate these contextual models across a range of contextual test sets from the literature, where we find that (a) document annotations from both mined parallel and back-translated monolingual data are helpful, but that the best contextual MT systems do not draw *contextual* training samples from the parallel data. We also make two points related to evaluation: (b) contrastive score-based metrics on challenge sets are not discriminative; instead, models must be tested directly on their ability to generate correct outputs, and (c) standard corpus-level metrics such as COMET work best in settings that are dense in contextual phenomena.

1 Introduction

By nature of its sentence-based design, traditional machine translation (MT) is unable to correctly translate any sentence with extra-sentential dependencies, such as pronouns in languages with grammatic gender, except by chance (Table 1). Despite significant prior work on the topic (§ 2), and general acknowledgment of the need to move on (Sennrich, 2018), contextual translation has never managed to overtake MT research, and sentence-level systems continue to dominate. This "sentence-level ceiling" leaves a gap between them and their increasingly powerful LLM counterparts, and raises the question of whether this gap can be narrowed or closed, if traditional MT systems could be trained properly with context.

English	German
I lost my hat. *Have you seen it?*	Ich verlor meinen Hut. *Hast du **es** gesehen?*

Table 1: The sentence-level translation ceiling. Selecting the correct pronoun (*ihn*, masc.) requires context.

A common explanation for the lack of context in MT has to do with the relative dearth of document-level annotations that are available for mined parallel and even monolingual data. At the same time, it has long been understood (Venugopal et al., 2011) and recently corroborated (Thompson et al., 2024) that crawled bitext is rife with machine translation output, which—though high quality at the sentence level—may attenuate the contextual signal. We explore this central problem by building the first large-scale, state-of-the-art translation systems trained on data with complete document annotations. We are able to do this because instead of public data, we use a private, in-house dataset (§ 3) that we have crawled ourselves. This crucially allows us to explore the effects of document annotations sourced from both parallel and monolingual (backtranslated data), together and in isolation, in order to quantify their effects. We find that:

- **It is best to source contextual training examples from backtranslated data only**. We find gains in contextual metrics from systems trained with contextual signals from both parallel and backtranslated data. However, the best systems source these samples from backtranslated data only.
- **Generative evaluation is crucial**. Contrastive metrics, where the task is to discriminate good and bad translations using model

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 1125–1139
November 15–16, 2024.

scores, are often used to evaluate contextual MT. We show that contextual systems that are trained on mined parallel documents do well on this task, but perform poorly when asked to generate correct translations. Only generative evaluation, which looks at whether correct words were produced, distinguishes good from bad contextual systems.

- **Standard metrics are most useful on discourse-dense datasets**. Standard sentence-level metrics like COMET are much more discriminative between sentence- and contextual systems when applied to datasets that are dense in discourse phenomena.

Together, these results raise important considerations for the construction and evaluation of contextual translation systems.

2 Background and Related Work

The transition to neural architectures was a paradigm enabler for document translation, since it eliminated the Markov limitations of statistical MT (Maruf et al., 2019). Much work has focused on special architectures and input encodings. This includes cache models (Tu et al., 2018; Kuang et al., 2018), hierarchical attention (Miculicich et al., 2018), separately encoding context (Voita et al., 2018; Zhang et al., 2018), allowing attention across a batch of pseudo-documents (Wu et al., 2023), encoding sentence position (Bao et al., 2021; Lupo et al., 2023), and sparse attention mechanisms (Guo et al., 2019). A number of approaches work on base systems outputs, such as post-editing with contextual language models (Voita et al., 2019a) and using contextual language models to rerank sentence-level system output Yu et al. (2020). Junczys-Dowmunt (2019) built one of the earliest contextual systems to perform well at WMT. Sun et al. (2022) also proposed to use standard transformer models, testing small architectures with no backtranslated data, and using a "multi-resolutional" training approach that creates overlapping documents. **We focus instead on standard architectures, judging them to be sufficient at large enough sizes.**

The lack of document-annotated parallel data is a longstanding problem. Datasets with document annotations are relatively small and specialized: they include OpenSubtitles (Lison and Tiedemann, 2016), WIT^3 (Cettolo et al., 2012), News Commentary, and Europarl (Koehn, 2005). Liu and Zhang (2020) provide a nice survey, and release a small amount of government-crawled new data for Chinese–Portuguese. Very recently, document annotations on Paracrawl data have become available Pal et al. (2024); Wicks et al. (2024). These annotations are available for only a relatively small subset of the data, however; even so, their results corroborate what we find here. (2024, Table 2) see *drops* in performance from systems trained with their parallel data annotations, unless the gold target context is provided; (2024) see small but consistent gains when the parallel data has been sufficiently filtered. The Conference on Machine Translation (WMT) began releasing limited document-level data for DE-EN and CS-EN in 2019 (Barrault et al., 2019). This limitation has forced researchers to get creative. Voita et al. (2019b) built a monolingual post-editing system that took the output of a baseline system and used it for document-level "repair". Sugiyama and Yoshinaga (2019) also used target-side data for backtranslation, evaluating in small-data settings with BLEU and contrastive metrics. **Our work is unique in that we have complete document annotations on very large web-crawled datasets**, and shows that these annotations on parallel data, as a whole, are not as useful.

Contextual metrics work has been important. PROTEST (Guillou and Hardmeier, 2016) used hand-designed pronoun test cases and also evaluated generatively. Many special test sets have been developed isolating important contextual phenomena and largely evaluating discriminatively (more in § 4). Läubli et al. (2018) provided early evidence that document-level metrics would be helpful. Several document-level metrics have been proposed, including BlonDe (Jiang et al., 2022), which compares automatically-identified phenomena in the output to those in a reference, and Doc-COMET (Vernikos et al., 2022), which incorporates contextual sentence representations. Both metrics are interesting but await deeper evaluation and we did not explore them in this paper. Vamvas and Sennrich (2021) have also noted the problem with the disconnect between contrastive evaluation and generative ability for machine translation. Both Fernandes et al. (2023) and Wicks and Post (2023) developed rules to identify contextually-dependent sentences. In this work, we show that **datasets dense in contextual phenomena are important for evaluating contextual ability**, and that **discriminative contextual evaluation is of limited use**.

3 The data challenge

Large publicly-available parallel datasets do not have document annotations. While the Conference on Machine Translation (WMT) has made overtures in this direction,[1] including ensuring that test data is source-language-natural and contains document information, parallel and monolingual data is limited to a small subset of all data[2] for which such information is easily retained. This is also true of recent work extracting document annotations from Paracrawl (Pal et al., 2024; Wicks et al., 2024).

We wish to experiment with and compare annotations sourced from both parallel and back-translated monolingual datasets. We therefore turn instead to a state-of-the-art, large collection of in-house data. We work with three language pairs: English→German, English→French, and English→Russian, which were selected because of the availability of good contextual evaluation data in each of them (§ 4). Our data comprises the following sources (Table 2):

- Monolingual data, crawled from expected-native sites: news (10%), data linked from the Open Directory Project[3] (40%), filtered webcrawl (40%), and Wikipedia and its outlinks (10%).
- Crawled parallel web data (similar to ParaCrawl)
- CCMatrix parallel data (Schwenk et al., 2021b), which has no document information.

Datasets have been filtered using bicleaner (Ramírez-Sánchez et al., 2020), with additional boilerplate and document deduplication.

Although the dataset is private, there is nothing in it that would surprise any researcher; the data was crawled from the web using standard techniques. The parallel data sources include a rough equivalent of ParaCrawl (Bañón et al., 2020) and also CCMatrix (Schwenk et al., 2021b). The monolingual data sources focus on sites where we expect data to have been written natively.

We emphasize that experiments at the scale presented in this paper are only possible with our private dataset, since document annotations are only available for small-data training settings like the TED talks data (Cettolo et al., 2012) used by IWSLT.[4] In a nod to the importance of repeatable work, we include results on the subset of our experiments that are possible on English–German public data and show that they corroborate corresponding results on private data (Section 7.6).

4 Contextual evaluation

A basic hurdle in the path to contextual translation is the difficulty of evaluation. We expect that contextual systems will produce improved translations of discourse-level phenomena, however, the frequency of these phenomena in standard corpora is not known, and we expect them to be relatively rare. This paper includes three types of evaluation.

4.1 Corpus-level metrics

The conventional way to test system performance is with standard metrics such as chrF (Popović, 2015) or COMET (Rei et al., 2020), which accumulate sentence-level scores to compute a single score for a test set. If the test set is organized into documents (as many are, including those from WMT), its sentences can be translated contextually and then split back out to sentences for evaluation. The expectation is that contextual translation will produce gains. However, a key consideration is whether the dataset is dense enough with contextual phenomena. Attempts to automatically identify sentences requiring context have shown the task to be difficult (Bawden et al., 2018) though possible with hand-created rules (Fernandes et al., 2023; Wicks and Post, 2023), but are often rare. Consequently, improvements may be invisible without the right test set.

We compare the performance of contextual systems using a standard corpus-level metric, COMET[5], on the following two test sets:

- WMT15. We use newstest2015 (Bojar et al., 2015) for EN→FR, and newstest2022 for EN→DE and EN→RU (Kocmi et al., 2022).
- OpenSubtitles (Lison and Tiedemann, 2016). We use the CTXPro (Wicks and Post, 2023) gender dataset, which is large and focuses on pronouns and anaphora.

[1] statmt.org

[2] Parallel: europarl, news-commentary, CzEng, and Rapid; Monolingual: news-crawl (en, de and cs), europarl, and news-commentary. Source: http://www2.statmt.org/wmt23/translation-task.html

[3] https://odp.org

[4] iwslt.org

[5] wmt20-comet-da

	English–French			English–German			English–Russian		
source	lines	docs	mean	lines	docs	mean	lines	docs	mean
mono	166.4	5.5	29.7	205.4	7.0	29.1	202.7	6.5	31.1
parallel									
→ crawled	123.1	3.7	33.0	116.7	4.7	16.6	72.4	4.7	13.2
→ ccmatrix	65.1	0	-	45.4	0	-	2.4	0	-

Table 2: Statistics of the training data used in our experiments (lines and docs in millions). The *mean* column is the mean document length in sentences of documents with ≥ 2 sentences.

Because the CTXPro dataset was constructed to select them, we expect it to be much denser in discourse phenomena. Data sizes are listed above the results in Table 3.

4.2 Contrastive test sets

The dominant paradigm for evaluation of long-tail document phenomena has been so-called *contrastive evaluation* (Sennrich, 2017), in which a system is tested on its ability to discriminate (via assigned model score) between correct and incorrect translation pairs. The correct examples are usually taken from found text; the incorrect ones are created by inserting an error of some sort. We look at three such test sets, examples of which can be found in Appendix A.

ContraPro (EN-DE) Müller et al. (2018) focus on the German pronouns *es*, *er*, and *sie*. They pair sentences containing naturally-found instances of pronouns drawn from OpenSubtitles with two variants where the incorrect pronoun has been used.

ContraPro (EN-FR) Lopes et al. (2020) extended ContraPro for EN-FR; the main difference is that there is only one incorrect example, since French has only two grammatical genders.

GTWiC (EN-RU) (Voita et al., 2019b) *Good Translation, Wrong in Context* (GTWiC) tests verb selection (500 instances) and morphology (500) in the presence of source-side ellipsis.

4.3 Testing generative ability

The challenge sets above test whether a model can discriminate between good and bad examples with using model score. However, this is at best a proxy for the true test of a machine translation system, which is to determine whether it generates the correct word or phrase. As we will show, many document models perform extremely well on these tasks, but when asked to actually translate the source sentence, produce the wrong word (Table 5). The contrastive nature of these test sets is at odds with the actual task: what is needed are metrics that directly evaluate a model's *generative*, rather than its *discriminative*, ability.

Fortunately, because these test sets were distributed with rich annotation information, we can transform them into generative test sets, where we test for the correct word in the output. A test set $\mathcal{T}$ comprises a set of test examples in the form of tuples (S, R, w), where S is the source sentence, R the reference, and $w \in R$ the target word or phrase that is expected to be found in the translation output. Let $\{T_i\}$ be the set of translations of the source sentences $\{S_i\}$. We compute accuracy[6] as

$$\text{acc}(T, \mathcal{T}) = \frac{1}{|T|} \sum_{i=1}^{|T|} \delta(w_i \in T_i)$$

This is not a perfect metric, since a correct translation may have paraphrased around the pronoun, but we do not expect that to systematically favor any particular system.

We have further opportunity to test this kind of accuracy with **CTXPro** (Wicks and Post, 2023), which expands ContraPro's coverage to many other languages and linguistic phenomena (auxiliaries, formality, gender, and inflection). CTXpro is evaluated only generatively, and has been been tested only on a single system, DeepL,[7] which is known to make use of context.

5 Experimental setup

We train and compare five models on the exact same data from two sources: parallel ($\mathcal{P}$) and back-translated monolingual ($\mathcal{B}$) data; the only difference among the models is whether document sam-

[6] Here accuracy is the same as both precision and recall.

[7] deepl.com

ples are drawn from neither, one, or both of the datasets.

Training All models are transformers trained with Marian (Junczys-Dowmunt et al., 2018a,b). We create two classes of models: first, those for backtranslation, and second, a set of models that constitute our primary comparative evaluation. For each language pair, we build a single joint unigram subword model (Kudo, 2018) with a vocabulary size of 32k that is used for both sets of models. Models are trained on random permutations over the training data for a predetermined number of updates. We use a batch size of 500k target-side tokens and a maximum sample length (whether sentences or pseudo-documents) of $L = 256$ tokens.

Backtranslated data The monolingual data is backtranslated (Sennrich et al., 2016) using sentence-level transformer systems (Vaswani et al., 2017) with 12 encoder and 6 decoder layers, an embedding size of 1024, and a feed-forward dimension of 8192. These models are trained for 20 virtual epochs.

This backtranslated data will be used to train contextual systems, but we note that this is not a problem, for two reasons. The major reason is that the target-side contextual signal is unaffected by backtranslation; since the original document boundaries are retained, any mistakes introduced by sentence-level backtranslation will appear just as normal source-side noise that the model must learn to overcome. Losses will be computed against the original, intact, target-side context. Second, even if this were not the case, our backtranslation models are into English, which is morphologically simpler than the evaluated translation direction.

Models For our contextual models, we also train transformers with a 12-layer encoder, a 6-layer decoder, and an embedding dimension of 1,024, but increase the feed-forward network size to 16,384. These models are trained for 40 virtual epochs to reflect the larger amounts of training data.

All of our models are trained on the complete parallel ($\mathcal{P}$) and backtranslated ($\mathcal{B}$) data. They vary only in whether the training procedure is permitted to construct multiple-sentence samples (also called *pseudo-documents* or *chunks*) from both, neither, or exactly one of these two pools of data. We compare the following systems, using the syntax NAME(pool$_1$, pool$_2$) to denote the pools of data each draws from; the presence of a box around the data source notes that pseudo-documents were drawn from it.

- SENT($\mathcal{P}$,$\mathcal{B}$). A sentence-level baseline.
- RAND($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$). A contextual system, but trained with completely random contexts.
- DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$). A contextual system, with documents from parallel and back-translated data.
- DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$). A contextual system, with documents drawn from parallel data only.
- DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$). A contextual system, with documents drawn from backtranslated data only.

Creating samples We create our training data on the fly using SOTASTREAM (Post et al., 2023), which iterates over $\mathcal{P}$ and $\mathcal{B}$. At each iteration, each data source is permuted randomly at the document level. To generate each sample, SOTASTREAM first chooses randomly between the two data pools. If documents are disabled on the pool, it simply returns the next sentence pair. If documents are enabled, it then chooses a maximum token length, and concatenates sentences on both sides until this length is reached on the source side, or the document's end is reached. Concatenated sentences are joined with a special ⟨SEP⟩ token, which facilitates sentence alignment at inference time for evaluation. Contextual samples are *chunked*, our term for the 1:1 concatenative construction described in Tiedemann and Scherrer (2017).[8] The training toolkit is then responsible for buffering as many samples as are needed to sort and form batches for training.

Inference For inference, we use the *last sentence* approach as defined in Herold and Ney (2023): each input sentence (the *payload*) is prepended with left sentence context, up to a maximum token length, L, which includes the payload. The translation system translates this as a single unit. The ⟨SEP⟩ token is then used to extract the payload's translation. This is repeated for all sentences in a test set, allowing standard sentence-level metrics to be applied to the results.

6 Results

Sentence-level metrics We begin by establishing baseline scores with a standard corpus-level metric, COMET, in Table 3. We include a commercial

[8]This can be contrasted with the "multi-resolution" approach of Sun et al. (2022), which creates training samples of different lengths from many overlapping sub-sequences of each input document

		EN→DE		EN→FR		EN→RU	
		WMT	CTXPro	WMT	CTXPro	WMT	CTXPro
	#lines	1,500	31,640	2,307	43,375	2,307	32,948
	Microsoft	62.0	27.7	67.6	36.4	67.3	39.1
sent-level	SENT($\mathcal{P}$,$\mathcal{B}$)	61.1	24.4	67.4	34.5	70.0	38.5
sent-level	RAND($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	59.2	22.7	67.6	33.6	68.9	36.6
sent-level	DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	60.2	23.4	67.0	33.5	70.5	38.8
sent-level	DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$)	59.7	22.6	68.8	34.1	70.0	37.8
sent-level	DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	60.9	24.5	67.8	34.7	70.3	38.2
context	RAND($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	58.8	20.4	66.8	32.1	68.7	35.4
context	DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	60.7	26.9	67.2	37.8	69.2	43.2
context	DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$)	60.2	25.4	67.9	37.6	68.5	40.3
context	DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	60.8	31.6	68.7	42.2	70.6	45.8

Table 3: COMET20 scores on WMT (22/15) and OpenSubtitles (CTXPro/gender) test sets translating alone (top block) and with context (bottom block). Numbers within a column are comparable. The gains from DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) (with context) over SENT($\mathcal{P}$,$\mathcal{B}$) (without it) are much larger for the discourse-dense OpenSubtitles data.

baseline (Microsoft, accessed via API). As another baseline, we present sentence-level results for the sentence-level system trained on all of our data. We then present results for all our models translating the test corpora (WMT and OpenSubtitles, using the CTXPro/gender dataset) in two modes: at the sentence level (top block), and with context (bottom block). In this way, we can look at the effect of context at both training and inference time.

Accuracy-based generative evaluation Next, we look at the broader CTXPro datasets and evaluate them using word accuracy on their relevant phenomena. Table 4 contains results for all three language pairs for all CTXPro datasets.

Contrastive suites Finally, we turn to the document-level contrastive and generative metrics described in § 4.2–4.3. Table 5 contains results for all three language pairs.

7 Discussion

7.1 Standard sentence-level metrics show gains if the dataset is dense enough

Table 3 shows state-of-the-art performance for all models when translating at the sentence level (without context), compared to the commercial system. This confirms the large-scale, state-of-the-art nature of our experiments. On the WMT datasets, we see a fairly a regular small drop on sentence-level translation with SENT($\mathcal{P}$,$\mathcal{B}$) (first row top sent-level section), that is slowly regained as we move down to DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$). We note that we do not expect the contextual translation systems to perform *better* at sentence-level translation, but hope they retain performance there.

Next, Table 3 allows comparison of sentence-level translation to contextual translation (top versus bottom section). On the WMT datasets, the effects gains are fairly small (-0.3 for EN→DE, +0.6 for EN→RU). Looking at the CTXPro columns, however, we observe fairly large, consistent gains when translating contextually with nearly all the (non-randomized) DOC systems, but especially for the DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) system across all three languages (+7.2 for EN→DE, +7.7 for EN→FR, and +7.3 for EN→RU). The CTXPro dataset is the OpenSubtitles gender-identified portion, so it is extremely dense in phenomena that require context to resolve compared to the WMT datasets, and is better able to discriminate systems with contextual abilities.

7.2 Domain and context both play a role

The DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) system showed large gains in Table 3 when translating CTXPro contextually. One explanation is that CTXPro is, by construction, "discourse dense". But it also represents a domain shift, from news to conversational domains. We would like to have an idea of how much of the gain is due to each.

We therefore conduct a followup experiment in EN→DE that compares two datasets in the OpenSubtitles domain: the CTXPro/gender "dense" test

	EN→DE			EN→FR		EN→RU			
	AUX	FORm	GEN	FORm	GEN	AUX	FORm	GEN	INFl
#lines	3,180	45,000	31,640	30,000	43,375	8,667	40,075	32,948	30,000
SENT($\mathcal{P}$,$\mathcal{B}$)	4.7	42.1	44.4	38.2	38.9	5.3	51.2	37.5	
RAND($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	4.7	39.6	42.4	36.9	38.2	5.5	51.4	36.7	32.2
DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	4.9	41.7	50.7	38.7	47.6	20.9	58.6	45.5	39.8
DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$)	4.2	41.4	47.2	42.7	45.2	16.7	56.8	39.5	37.4
DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	7.5	45.0	66.0	43.8	54.8	25.2	58.7	53.5	42.6

Table 4: Generative accuracy on CTXPro datasets, where the task is to translate a source sentence and then determine whether an exact form of the required target word is in the output. The contextual systems trained on documents from mined parallel data perform notably worse than the DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) system.

	EN→DE		EN→FR		EN→RU			
	gender		gender		NP ellipsis		VP ellipsis	
model	contr.	gen.	contr.	gen.	contr.	gen.	contr.	gen.
RAND($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	43.3	35.5	71.2	40.1	18.0	24.8	52.6	4.8
DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	77.0	40.9	91.2	56.2	20.9	58.6	45.5	39.8
DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$)	75.1	37.0	92.5	52.5	16.7	56.8	39.5	37.4
DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	**80.8**	**66.8**	**93.4**	**68.5**	**25.2**	**58.7**	**53.5**	**42.6**

Table 5: Document contrastive test scores (contr.) and their generative (gen.) variants. All accuracies are over items with extra-sentential antecedents only. DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) consistently performs best on generative metrics by wide margins, while for contrastive metrics, other contextual systems are often similar or exhibit no consistent pattern.

context	Dense true	Sparse true	Dense rand
SENT($\mathcal{P}$,$\mathcal{B}$)	24.4	30.5	24.4
DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$)	26.9	31.4	24.8
DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$)	25.4	32.4	25.4
DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	31.6	31.7	21.8

Table 6: EN→DE COMET scores on a dense dataset (OpenSubtitles CTXpro/gender) with true and random contexts; next, a sparse dataset (random sample of OpenSubtitles) with true contexts. DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) gains most over the sentence baseline on dense with true contexts and is harmed most on dense with random contexts. The doc systems are similar on the sparse dataset.

set, and another test set, which contains a random sample of 500 ten-sentence documents from OpenSubtitles 2016, yielding a corpus size of 4,973 sentences. We label this second one "sparse": since it was selected randomly, it is likely to be much less dense in contextual phenomena. For contextual systems, we translate each of these as a single chunk, and then split them out for evaluation with COMET. The results are in Table 6.

The differences between the first two columns shows that the DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) gains over the sentence system are much larger on the "dense" dataset (+7.2 vs. +1.2). Performance among the contextual systems is closer, as we saw with WMT datasets. This suggests that the flat performance with WMT data was likely due to it, too, being sparse with contextual phenomena. **For standard, sentence-based metrics like COMET to separate these systems, dense test sets are needed.**

Table 6 (column 3) contains the results of another experiment, where we replace the context of each sentence in the "dense" dataset with a random context. This hurts performance, and the effect is most pronounced on the DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) system, suggesting that this model is most dependent on a reliable contextual clue.

7.3 Generative word-based accuracy corroborates these differences

Table 4 presents the results of word-based accuracy on the CTXPro datasets, across a range of linguistic phenomena. With word-based accuracy, we

are testing whether a word is present in the output. This leaves open the possibility of metric mistakes. For example, if the pronoun *er* is expected in the output, a system could be penalized for translating the sentence correctly with no pronouns, or it could be rewarded for generating a semantically unrelated instance of *er*. We do not expect this to systematically favor any one system.

Here, we see a similar gap between (a) contextual systems versus a random context and (b) especially, a gap between DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) and the other contextual systems. For EN→DE and EN→FR, the gender categories are similar to the ContraPro test sets for those languages, but much larger. This is most true for the GENder category (with gains of +23.6, +16.6, and +10.4), but also for other categories, including auxiliaries (+19.7 for EN→FR) and EN→RU inflection (+10.4).

7.4 The general trend favors BT-only contextual data

Figure 1 visualizes the metric score gains from Tables 3 and 5 for all four contextual models over the sentence-level baselines. The x-axis is arranged by the percentage of the contextual examples that are drawn from parallel data. This makes clearer the observations from the discussion so far: contextual annotations from parallel data are better than nothing, but they are inferior to those from the backtranslation monolingual data, and removing them is preferable.

7.5 Contrastive test sets are less discriminative

Table 5 contains results that pair contrastive accuracies (§ 4.2) with their generative counterparts. Across all three language pairs, there is an interesting pattern: in the contrastive metrics, the document systems improve over the sentence baseline, as a block. However, *the generative metrics see their best results with* DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)*, often by a large margin*. Together with the observations in the previous section, we believe this calls into question the reliability of contrastive metrics. What we really care about in an MT system is its ability to *generate* the correct results at inference time. Discriminative ability is at best a proxy for this ability; if its results do not correlate with such metrics, it calls into question its reliability.

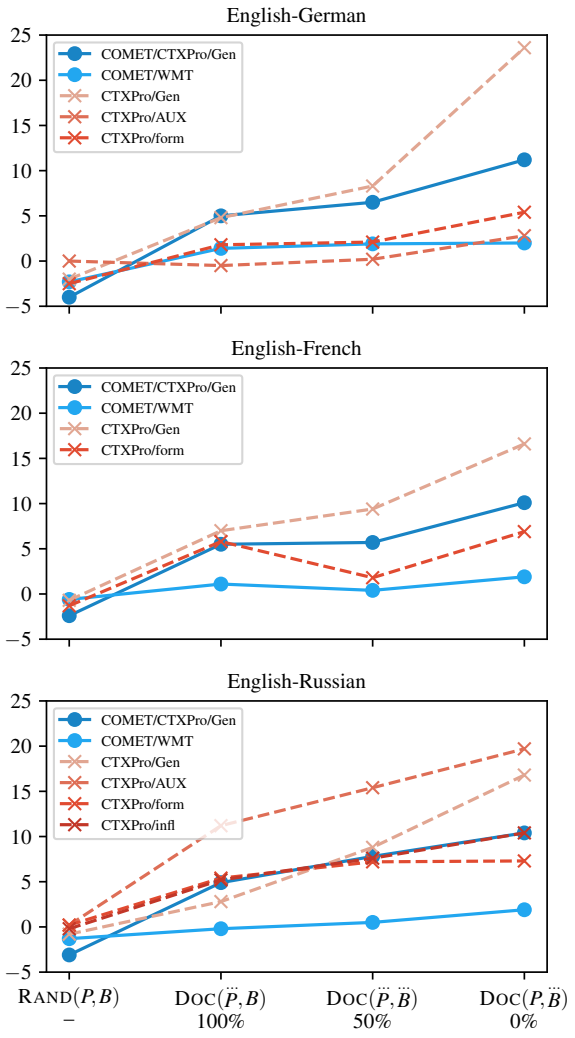

Figure 1: Contextual metric gains over the sentence baseline for COMET and accuracy metrics for the four systems, arranged by the percentage of contextual samples sourced from parallel data.

7.6 Experiments with public data provide some corroboration

Since complete document annotations for publicly available large-scale parallel data do not exist, we were unable to build DOC($\ddot{\mathcal{P}}$,$\ddot{\mathcal{B}}$) and DOC($\ddot{\mathcal{P}}$,$\mathcal{B}$) on open data. However, we can build the SENT($\mathcal{P}$,$\mathcal{B}$) and DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$) systems with a subset of the WMT22 EN→DE data with monolingual document annotations, and see whether they exhibit the same pattern.

We use all available parallel data provided for WMT22 (Kocmi et al., 2022):[9] Europarl v10 (Koehn, 2005), Paracrawl v9 (Bañón et al., 2020), Common Crawl,[10] News Commentary, Wiki Ti-

[9] statmt.org/wmt22/translation-task.html
[10] https://commoncrawl.org/

context	But let's not give in just yet.⟨SEP⟩ Right now, this is our one chance to be different.⟨SEP⟩ We could do something great with it.⟨SEP⟩ Like save the science museum.⟨SEP⟩ We grew up going to that place our whole lives.⟨SEP⟩ It's gave us so much.⟨SEP⟩ This is an opportunity to give something back.⟨SEP⟩ Besides, aren't you curious?⟨SEP⟩ So, three wishes are granted to whoever discovers the box.
source	But we all found **it**. And touched it at the same time.
SENT($\mathcal{P}$,$\mathcal{B}$)	Aber wir haben **es** alle gefunden. Und **es** gleichzeitig berührt.
RAND($\dddot{\mathcal{P}}$,$\dddot{\mathcal{B}}$)	Aber wir haben **es** alle gefunden und gleichzeitig berührt.
DOC($\dddot{\mathcal{P}}$,$\dddot{\mathcal{B}}$)	Aber wir haben **es** alle gefunden. Und haben **es** gleichzeitig berührt.
DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	Aber wir haben **sie** alle gefunden und gleichzeitig angefasst.
ref	Aber wir haben **sie** alle gleichzeitig entdeckt und berührt.
context	Mark it.⟨SEP⟩ If Mr. Wick isn't dead already, he soon will be.⟨SEP⟩ Will you mark it, sir?⟨SEP⟩ You have no idea, what's coming do you?⟨SEP⟩ I have everyone in New York looking for him.⟨SEP⟩ I doubt we will see him again.⟨SEP⟩ Do you now?⟨SEP⟩ You stabbed the devil in the back, and forced him back into the life that he had just left.⟨SEP⟩ You incinerated the priest's temple.
source	Burned **it** to the ground.
SENT($\mathcal{P}$,$\mathcal{B}$)	Verbrannte **es** bis auf die Grundmauern.
RAND($\dddot{\mathcal{P}}$,$\dddot{\mathcal{B}}$)	Verbrannte **es** zu Boden.
DOC($\dddot{\mathcal{P}}$,$\dddot{\mathcal{B}}$)	Hast **es** zu Boden gebrannt.
DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	Sie haben **ihn** niedergebrannt.
ref	Und **ihn** niedergebrannt.

Table 7: Translation examples from the CTXPro gender dataset demonstrating DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)'s superior performance. Pronouns are **in bold** with antecedents underlined. For all but SENT($\mathcal{P}$,$\mathcal{B}$), the source is translated together with the context, and then the context is discarded.

.

tles v3, Tilde MODEL Corpus (Rozis and Skadiņš, 2017), and Wikimatrix (Schwenk et al., 2021a). A few of these resources have document-level information, but we do not use any of it. For monolingual data, the only data available with document metadata is News Crawl.[11] We used all even years from 2008–2020, backtranslating it from German to English with an internal system. No filtering is applied. From this data, we train the only two of our systems supported by this setup: SENT($\mathcal{P}$,$\mathcal{B}$) and DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$). These are trained for 40 virtual epochs each using the same settings described in Section 6.[12]

Results can be found in Table 8. They are encouraging: we see the same pattern of improvement between SENT($\mathcal{P}$,$\mathcal{B}$) and DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$), although the absolute numbers are lower. Compared to our in-house data, the document metrics are even better for SENT($\mathcal{P}$,$\mathcal{B}$).

[11] https://data.statmt.org/news-crawl/de-doc/

[12] Mono data: 311.2m lines, 14.1m docs, with a mean sentence length of 21.9 sentences. Parallel data: 297.6m lines.

system	COMET	gender contr.	gender gen.
SENT($\mathcal{P}$,$\mathcal{B}$)	60.6	56.7	23.9
DOC($\dddot{\mathcal{P}}$,$\dddot{\mathcal{B}}$)	x	x	x
DOC($\dddot{\mathcal{P}}$,$\mathcal{B}$)	x	x	x
DOC($\mathcal{P}$,$\ddot{\mathcal{B}}$)	59.4	83.4	64.3

Table 8: Metrics on the only two models we are able to build on public data. Similar patterns are observable to those seen in Tables 3 and 5.

7.7 MT output in crawled parallel data

We do not undertake an exploration of the causes for the results and analysis discussed in Figure 1 and throughout this section, but there is an obvious explanation: we suspect that parallel web-crawled data is full of machine-translated output. Widespread use of translation across the web, especially since the release of Google Translate in 2006, is a commercial success story that has unfortunately produced a kind of "poisoning of the

English	German
Unique Moorish style **villa** set in a tropical oasis with pool, guest accommodation and amazing views. ⟨SEP⟩ Property Reference 1846 ⟨SEP⟩ **It** was built by the current owner...	Einzigartige maurische **Villa** in einer tropischen Oase mit Pool, Gästeunterkunft und herrlicher Aussicht. ⟨SEP⟩ Referenznummer 1846 ⟨SEP⟩ **Es** wurde vom jetzigen Besitzer gebaut...

Table 9: An example of bad data drawn from the parallel data pool. While the sentence-level translations are fine, the incorrect pronoun *Es* in the third sentence suggests sentence-level machine or low-quality human translations.

well", where machine translation outputs are later collected as training data for new systems (Venugopal et al., 2011). Recent work has corroborated how extensive this is in multi-way parallel data (Thompson et al., 2024).

Quantifying this awaits further work, but it is easy to source examples from our parallel data (Table 9). While we don't know if this was generated by machine or a human, we do know that even large NMT systems are sensitive to small amounts of poor data.[13]. This data may still be of high quality at the sentence level; it is only *inter-sentence contextual* information that is affected. If true, this suggests that **contextual translation introduces a new quality dimension that is invisible** in the standard sentence-level training paradigm, and the problem may in fact be quite large, since all machine translation content in the wild will have been generated at the sentence level.

We suspect that our monolingual data—which by design was sourced from known target-native sites, such as newspapers—is largely immune from these problems. Training on sentence-level translations is primarily a problem for data translated in the *forward* direction. Backtranslation introduces noise into the *source language text*, while preserving the target-language contextual signal.

We leave to future work an investigation into detecting and removing machine translation output from parallel data at high enough precision.

8 Conclusions

Machine translation research and production systems continue to be dominated by sentence-level approaches. A common explanation for this shortcoming is the lack of document-annotated parallel data. We have compared the effectiveness of constructing contextual translation models for three translation directions in large-data settings. Our results suggests that while mined parallel data is of high-enough quality for building sentence systems and contains some contextual signal, **it is best to construct contextual training samples from back-translated data only**. Although we have not investigated the reasons for this, we consider it a strong possibility that our parallel data, which is mostly crawled from the web and has had only sentence-level filtering applied, contains large amounts of data that was machine-translated at the sentence level, a finding that is very likely to hold for publicly available data, as well. This suspicion makes sense a priori, and is bolstered in other recent work (Thompson et al., 2024; Wicks et al., 2024; Pal et al., 2024).

We have also shown the importance of **evaluating contextual machine translation output in its generative capacity**, rather than in its ability to discriminate good outputs from bad ones. This can be done by using provided challenge sets like CTXPro or converting existing contrastive metrics like ContraPro and its variants, or by using standard corpus-level metrics like COMET on test sets that are sufficiently dense with contextual phenomena.

A fruitful avenue for followup work is to automatically identify sentences that require context to translate correctly, which could be used to filter training data and also in the construction of new test sets. Though we have focused on "traditionally"-trained MT, it will also be useful to learn how LLMs perform on these tasks.

Acknowledgments

We thank the authors of the contrastive test sets (particularly ContraPro, which set the standard) for their fantastic work creating well-designed, usable, and easily-extendable datasets, along with code. Thanks also to Rachel Bawden, Salvador Mascarenhas, Christian Federmann, Tom Kocmi, Vikas Raunak, Arul Menezes, and Huda Khayrallah for helpful comments and feedback, and to Fai Sigolov for help with Russian.

[13] A classic example is source-copy data (Ott et al., 2018)

Limitations

With respect to reproducibility, the deepest limitation of our paper is our use of private data. There is therefore a risk that our findings might not be reproducible by other teams working with (necessarily) different datasets. Finally, although we suspect our results will hold for language pairs beyond the three we investigated, it is possible they will not generalize.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

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. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Thomas Lavergne, and Sophie Rosset. 2018. Detecting context-dependent sentences in parallel corpora. In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 393–400, Rennes, France. ATALA.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

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

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1315–1325, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023. On search strategies for document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12827–12836, Toronto, Canada. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018a. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018b. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135, Melbourne, Australia. Association for Computational Linguistics.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation. *CoRR*, abs/2006.10369.

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. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Siyou Liu and Xiaojun Zhang. 2020. Corpora for document-level neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France. European Language Resources Association.

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

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

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019. A survey on document-level machine translation: Methods and evaluation. *CoRR*, abs/1912.08494.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *CoRR*, abs/1803.00047.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

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

Matt Post, Thamme Gowda, Roman Grundkiewicz, Huda Khayrallah, Rohit Jain, and Marcin Junczys-Dowmunt. 2023. SOTASTREAM: A streaming approach to machine translation training. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 110–119, Singapore. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

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

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich. 2018. Why the time is ripe for discourse in machine translation. Talk given at NGT 2018: https://aclanthology.org/volumes/W18-27/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

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

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1763–1775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch. 2011. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. Recovering document annotations for sentence-level bitext. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9876–9890, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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

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

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

A Dataset examples

Examples from the datasets used for generative and contrastive evaluation can be found in Tables 10 and 11.

The prototype has passed every test, sir. It's working. | Der Prototyp hat jeden Test erfolgreich durchlaufen, Sir. {Er,Es,Sie} funktioniert.

(a) ContraPro example. Contrastive examples are formed by substituting incorrect pronouns.

Veronica, thank you, but you **saw** what happened. We all did. | Вероника, спасибо, но ты видела, что произошло. Мы все **хотели**.

(b) GTWiC example. The first Russian sentence uses the formal register.

Table 10: Examples from contrastive test sets.

(AUX) I just figured you need to know. And now you do. → Je pensais que tu méritais de savoir. Et maintenant tu *sais*.
(INF) My friend had some mech work done here. Industry stuff. → Вы ставили имплант моей подруге. Промышленную штуковину.
(FORm) I don't know you, but.. → Ich kenne Sie nicht, aber...

Table 11: Examples of contextually-sensitive auxiliary and inflection elision from the CTXPro dataset.

B Model capacity

Much work in investigating document-level machine translation has been limited to standard-size Transformer architectures (cf. Zhang et al. (2018); Sun et al. (2022); Lopes et al. (2020)). Yet it stands to reason that modeling longer-range phenomena will require increased model capacity, and in fact, the base model size we chose for our experiments (12 layer encoder, 16k FFN) reflects this. Here, we provide more detail, varying two model parameters only: (i) the number of encoder layers, and (ii) the width of the model feed-forward layer (encoder and decoder side). We keep all other parameters the same, including fixing the decoder depth to 6. Focusing on changes to the encoder depth helps limit grid search and is justified by prior work showing that (relatively cheap) encoder layers can be traded for (relatively expensive) decoder layers with no penalty (Kasai et al., 2020). We alternate between increasing the number of encoding layers, and increasing the dimension of the Transformer feed-forward layer.

Table 12 contains English–German results. Unsurprisingly, all scores continue to rise, up to the wide 18-layer model. Both increasing the number of encoder layers, and increasing the size of the FFN, contribute to better performance. This suggests that the common approach of working with 6-layer Transformer base models is not enough

arch	params	BLEU	COMET	C/Pro	G/Pro
6/1k	146m	27.0	48.7	65.2	58.4
6/2k	171m	27.4	49.7	66.2	58.7
6/4k	221m	28.0	51.0	69.7	62.9
12/4k	297m	28.4	51.8	70.6	66.0
6/8k	322m	27.8	51.0	71.7	62.8
12/8k	448m	28.6	52.5	74.2	67.1
6/16k	523m	28.4	51.7	74.5	64.9
18/8k	574m	28.8	53.0	75.0	67.1
12/16k	750m	28.9	52.8	75.8	68.5
18/16k	977m	29.3	53.3	75.5	69.4

Table 12: Model capacity (encoder layers / FFN / # params) for an EN-DE document model, ordered by param. count. Decoder depth is always 6 layers. Scores were computed on a checkpoint after 30k updates. BLEU and COMET scores are on WMT21, translating as sentences. C/Pro is over the complete test set, while G/Pro is over only sentences with external anaphora.

for document-context MT. There is more to gain by moving to larger models and likely, to larger datasets and context lengths, as well.