Generative evaluation for contextual machine translation

Anonymous ACL submission

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. This limitation is an inherent performance ceiling that is increasingly glaring compared to their natively-contextual LLM counterparts. 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. Working with a private parallel and monolingual data set, we build large-scale, state-of-the-art contextual MT systems into German, French, and Russian. We find that these systems are harmed when including contextual training examples sourced from mined parallel bitext. We also show that these improvements are invisible when using contrastive score-based test sets; instead, models must be tested directly on their ability to generate correct outputs, or with standard metrics on discourse-dense test sets. This provides evidence that mined parallel bitext does not contain reliable contextual signals—perhaps because it was translated in a sentence-level manner. Where possible, we repeat our results on public data.

1 Introduction

Large language models (LLMs) have transformed the field of natural language processing, providing high-quality working solutions to problems in many domains (e.g., question answering, summarization, multi-step reasoning) that even a few years ago had no solution in clear sight. LLMs have also proven capable in subtasks that had already experienced substantial commercial success: most notably, machine translation (MT). While they have not supplanted the "traditional" MT paradigm (i.e., using sequence-to-sequence models trained on mined-parallel and backtranslated-monolingual data), they outperform them in many high-resource language pairs (Xu et al., 2023), and also introduce new capabilities, such as easy stylistic adaptation (Moslem et al., 2023). At the same time, MT systems trained in this way retain their own advantages, including small model sizes and corresponding inference-time efficiency.

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

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

An important advantage that LLMs possess is that they are natively document-level, meaning they can easily handle contextual phenomena. By nature of its sentence-based design, traditional MT is unable to correctly translate any sentence with extra-sentential dependencies, such as pronouns in languages with grammatical gender, except by chance (Table 1). Despite significant prior work on the topic (§ 7), and general acknowledgment of the need to move on (Sennrich, 2018), contextual translation has never managed to take hold in MT research, and sentence-level systems continue to dominate. This 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.

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 (§ 2) 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:

- **Sourcing contextual training examples from parallel data is harmful**. Parallel text mined from the web is a key component in constructing *sentence-based* translation systems, but attempts to use it *contextually* fail. We suspect this has to do with the prevalence of machine translation output (Thompson et al., 2024), which may be high quality at the sentence level, but which has a weakened contextual signal. We get around this by sourcing contextual samples only from backtranslated data.
- **Generative evaluation is crucial**. Contrastive metrics, where the task is to discriminate good and bad translations using model 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 require 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 The challenge of data

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.

We wish to experiment with and compare annotations sourced from both parallel and backtranslated monolingual datasets. We therefore turn instead to a state-of-the-art, large collection of in-house data.

2.1 Data description

We work with three language pairs: English→German, English→French, and English→Russian. We chose these languages because of the availability of good contextual evaluation data in each of them (§ 3). 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 proprietary, 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 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

[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

	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.

data and show that they corroborate corresponding results on private data (Appendix C). We also include 1000-document EN-DE samples with this submission.

2.2 MT output in crawled parallel data

Translation is a core facilitator of cross-cultural communication, and also an expensive one, when undertaken by humans. It is therefore not surprising that automated machine translation has long been one of the success stories from the field of natural language processing, with widespread commercial adoption and popularization, especially with the release of Google Translate in 2004. Unfortunately, one consequence of this success has been a "poisoning of the well", where machine translation outputs are later collected as training data for new systems (Venugopal et al., 2011).

It is standard practice to filter out the worst quality translations with various techniques. At the same time, not all machine-generated data is bad for training. An example, sourced from our parallel data, can be found in Table 3. The individual sentence pairs are fine for training sentence-level systems, and *only become problematic when training contextual ones*. 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.[5] This fact, together with recent reports on the prevalence of MT output in multi-way parallel datasets (Thompson et al., 2024), suggest there may be a big problem. This is all to say that **contextual translation introduces a new quality dimension that is invisible** in the standard training paradigm, and the problem may in fact be quite large, since all machine translation content in the wild will have been generated by sentence-level systems.

We do not expect to see this problem for our monolingual data. It is drawn selectively from sites and sources which are most likely to produce target-side native data, such as news sites. We do not have direct proof that there are not elements of translated data, but the experiments and discussion in this paper help establish this.

3 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.

3.1 Corpus-level metrics

The conventional way to test system performance is with corpus-level 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.

- WMT (2015 for EN→FR, and 2022 for the others). We expect that these are sparse.

- OpenSubtitles (Lison and Tiedemann, 2016). We use the CTXPro/gender dataset (§ 3.3),

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

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 3: 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.

which is large and discourse-dense, namely with pronouns and anaphora.

We compute a standard corpus-level COMET score[6] on these test sets, in two settings: translating (i) without context and (ii) with up to 10 sentences or 250 tokens of left context.

3.2 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 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. Systems are evaluated on the percentage of time they correctly score the positive example above its incorrect variant, by way of model score.

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.

Examples of sentences in these test sets can be found in Appendix A.

3.3 Testing generative ability

The challenge sets above test whether a model can discriminate between good and bad examples with using model score. 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 6). 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 as

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

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 also use **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.

4 Experimental setup

We train and compare four 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 samples are drawn from neither, one, or both of the

[6] Model wmt20-comet-da

[7] deepl.com

datasets. 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, trained for 20 virtual epochs[8] on the parallel data.

Models All of our models are transformers trained with Marian (Junczys-Dowmunt et al., 2018a,b). For each language pair, we build a single joint unigram subword model (Kudo, 2018) of size 32k. Our experiments with different model capacities (Appendix B) led us to use a 12-layer encoder, a 6-layer decoder, an embedding dimension of 1,024, and a feed-forward network size of 16,384. We train for 40 virtual epochs. We use a batch size of 500k target-side tokens. Our maximum document sample length is $L = 256$ tokens.

Our models vary based on whether they are trained on multi-sentence samples (compared to just single sentences) from the backtranslated data, the parallel data, both datasets, or neither. We compare the following variants, using the syntax NAME(pool$_1$, pool$_2$) to denote the pools of data each draws from:

- SENT($\mathcal{P}$,$\mathcal{B}$). A sentence-level baseline.
- SENT⋆($\mathcal{P}$,$\mathcal{B}$). An inference-only baseline that abuses SENT($\mathcal{P}$,$\mathcal{B}$) to translate contextually.[9]
- DOC($\mathcal{P}_d$,$\mathcal{B}_d$). A contextual system, with documents from parallel and back-translated data.
- DOC($\mathcal{P}_d$,$\mathcal{B}$). A contextual system, with documents drawn from parallel data only.
- DOC($\mathcal{P}$,$\mathcal{B}_d$). 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 randomized permutations of $\mathcal{P}$ and $\mathcal{B}$. To generate each sample, SOTASTREAM first chooses randomly between the two data pools (parallel and backtranslated). A run-time flag determines whether contextual samples are enabled for each pool (denoted $\mathcal{P}_d$ and $\mathcal{B}_d$, respectively). If not, it simply returns the next sentence pair. If so, 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*, meaning they are formed from adjacent, non-overlapping sequences of sentences in the training data, in contrast to the "multi-resolution" approach (Sun et al., 2022), which creates training samples from many overlapping sub-sequences of each input document. 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 an *overlapping* approach. 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.

5 Results

Sentence-level metrics We begin by establishing baseline scores with a standard corpus-level metric, COMET, in Table 4. We include a commercial baseline (Microsoft, accessed via API). 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.

We also conduct a followup experiment in English→German designed to investigate the importance of (a) having the true context at inference time and (b) comparing "contextually dense" and "sparse" datasets in the same domain. In Table 5, we first compare the OpenSubtitles CTXPro/gender dataset. Column 1 reports results translating with the true context (i.e., a repeat of the contextual results from Table 4), whereas column 2 randomly shuffles the contexts of these 31,640 test sentences, testing how important having the true context is for translation. Column 3 reports results from a new, random selection of 500 ten-sentence documents from OpenSubtitles 2016, yielding a corpus size of 4,973 sentences. These documents are each translated in single chunks. We call this subset "sparse": since it was selected randomly, it is likely to be much less dense in contextual phenomena.

[8] Updates from one billion target-side tokens.

[9] In this setting alone, no ⟨SEP⟩ token is used when combining sentences, since the sentence model has not seen them.

		EN→DE		**EN→FR**		**EN→RU**	
		WMT	CTXPro	WMT	CTXPro	WMT	CTXPro
	model/#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($\mathcal{P}$,$\mathcal{B}$)	61.7	24.7	69.1	35.4	70.0	38.5
	DOC($\mathcal{P}_d$,$\mathcal{B}_d$)	62.0	25.4	70.0	35.7	70.5	38.8
	DOC($\mathcal{P}_d$,$\mathcal{B}$)	61.3	24.3	69.2	35.0	70.0	37.8
	DOC($\mathcal{P}$,$\mathcal{B}_d$)	62.2	25.8	69.8	35.7	70.3	38.2
context	DOC($\mathcal{P}_d$,$\mathcal{B}_d$)	62.1	30.8	69.2	40.4	69.2	43.2
	DOC($\mathcal{P}_d$,$\mathcal{B}$)	62.1	29.2	67.6	39.4	68.5	40.3
	DOC($\mathcal{P}$,$\mathcal{B}_d$)	62.2	34.3	70.2	44.1	70.6	45.8

Table 4: 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($\mathcal{P}$,$\mathcal{B}_d$) (with context) over SENT($\mathcal{P}$,$\mathcal{B}$) (without it) are much larger for the discourse-dense OpenSubtitles data.

	Dense		Sparse
context	true	rand	true
SENT($\mathcal{P}$,$\mathcal{B}$)	24.7		30.5
DOC($\mathcal{P}_d$,$\mathcal{B}_d$)	30.8	24.8	31.4
DOC($\mathcal{P}_d$,$\mathcal{B}$)	29.2	25.4	32.4
DOC($\mathcal{P}$,$\mathcal{B}_d$)	34.2	21.8	31.7

Table 5: 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($\mathcal{P}$,$\mathcal{B}_d$) 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.

Contrastive suites Next, we turn to the document-level contrastive and generative metrics described in § 3.2–3.3.

For generative document metrics, we took special care with SENT⋆($\mathcal{P}$,$\mathcal{B}$). It was not trained with the separator token, making it hard to identify the payload sentence's translation. We work around this by applying the Moses sentence splitter.[10] Spot-checking suggests this to be a reasonable heuristic that likely *overestimates* accuracy, since the identified sentence is often longer than it should be. Table 6 contains results for all three language pairs.

Accuracy-based generative evaluation Finally, in Table 7, we present accuracy results on the relevant CTXPro datasets for each language.

6 Discussion

6.1 Standard sentence-level metrics work if the dataset is dense enough

Table 4 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 consistent gain of roughly a COMET point when moving from the baseline sentence-level translation with SENT($\mathcal{P}$,$\mathcal{B}$) (first row top sent-level section) to DOC($\mathcal{P}$,$\mathcal{B}_d$); however, these gains are observed in nearly all contextual systems. Looking at the CTXPro columns, however, we observe a large, clear separation between the DOC($\mathcal{P}$,$\mathcal{B}_d$) system and all the other contextual systems, across all three languages.

We believe the explanation for this is two-fold: first, the CTXPro dataset is the OpenSubtitles gender-identified portion, so it is extremely dense in discourse phenomena. Second, while three systems have contextual training data, it is only DOC($\mathcal{P}$,$\mathcal{B}_d$) whose contextual training signal has not been muddied by unreliable contextual data from the mined parallel training data pool.[11]

Table 5 contains results that suggest these score differences are not random effects between the two test sets. In the first experiment (first two results columns), we randomly swap the contexts provided to each payload sentence in the CTXPro EN→DE/gender dataset. The effect is most pronounced on the DOC($\mathcal{P}$,$\mathcal{B}_d$) system, suggest-

[10] github.com/mediacloud/sentence-splitter

[11] We note that OpenSubtitles is not in our training data.

	EN→DE		**EN→FR**		**EN→RU**			
model	C/Pro	G/Pro	C/Pro	G/Pro	C/ell$_{\text{infl}}$	G/ell$_{\text{infl}}$	C/ell$_{\text{VP}}$	G/ell$_{\text{VP}}$
Literature	70.8	-	83.2	-	76.2	-	80.0	-
Sent($\mathcal{P}$,$\mathcal{B}$)	50.0	33.2	71.6	22.5	51.8	24.8	19.8	4.6
Sent$\star$($\mathcal{P}$,$\mathcal{B}$)	69.0	46.3	93.1	62.3	77.0	32.8	55.0	19.2
Doc($\mathcal{P}_d$,$\mathcal{B}_d$)	76.5	47.8	**95.1**	62.5	84.2	35.8	**68.0**	26.0
Doc($\mathcal{P}_d$,$\mathcal{B}$)	71.6	41.9	94.3	60.4	76.2	31.8	66.2	26.4
Doc($\mathcal{P}$,$\mathcal{B}_d$)	**77.9**	**70.5**	94.8	**77.3**	**84.6**	**39.6**	66.0	**28.4**

Table 6: Document contrastive test suites and their generative variants. Contrastive accuracies (C/*) are over the entire dataset in order to compare with the literature, while generative accuracies (G/*) are over extra-sentential items only. Literature scores are taken from Lopes et al. (2020, EN→FR,EN→DE), and Voita et al. (2019b). Feeding documents to Sent$\star$($\mathcal{P}$,$\mathcal{B}$) (which it wasn't trained on) increases contrastive scores over the sentence baseline and generally brings generative scores within line of doc systems trained with parallel data.

	EN→DE			EN→FR		EN→RU			
	AUX	FORm	GEN	FORm	GEN	AUX	FORm	GEN	INFl
Sent($\mathcal{P}$,$\mathcal{B}$)	4.5	42.5	44.5	39.0	38.9	4.7	52.2	37.6	32.9
Doc($\mathcal{P}_d$,$\mathcal{B}_d$)	7.8	46.0	55.7	45.4	48.0	20.9	58.6	45.5	39.8
Doc($\mathcal{P}_d$,$\mathcal{B}$)	7.6	44.4	52.0	45.7	46.2	16.7	56.8	39.5	37.4
Doc($\mathcal{P}$,$\mathcal{B}_d$)	11.7	46.3	69.8	46.3	56.4	25.2	58.7	53.5	42.6

Table 7: 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($\mathcal{P}$,$\mathcal{B}_d$) system.

ing that this model is most dependent on a reliable contextual clue. The final column shows performance on a random subset of OpenSubtitles (§ 2), rather than the carefully-selected CTXPro/gender. Here, we see the performance among the document systems is quite similar, 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.

6.2 Contrastive test sets are problematic

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($\mathcal{P}$,$\mathcal{B}_d$), often by a large margin*. Additionally, the Sent$\star$($\mathcal{P}$,$\mathcal{B}$) system *improves* over the Sent($\mathcal{P}$,$\mathcal{B}$) system when measured contrastively, but these gains are not reflected in the generative metric. This calls into question the reliability of contrastive metrics, since we know this system has no generative document capacity.

We direct special attention to Sent$\star$($\mathcal{P}$,$\mathcal{B}$). This system was trained on sentences only, yet it performs on par with—and even above—the best literature results. Yet we know from the generative experiments that when asked to produce these translations, it is often unable to do so. Discriminative ability is not the same as generative.

We repeat the note from Section 3.3 that generative evaluation may penalize a system that produces a correct sentence not containing the pronoun, or unfairly credit a system that happens to generate the pronoun by accident, but do not expect that this will favor any particular system. Spot-checking suggests to us that the large differences reported in Table 6 capture actual improvements.

6.3 Generative accuracy captures differences

Finally, Table 7 shows a similar gap between the Doc($\mathcal{P}$,$\mathcal{B}_d$) and other systems when testing for word-based accuracy. For EN→DE and EN→FR, the gender categories are similar to the ContraPro test sets for those languages, but much larger. The other categories show that the gains continue across a range of linguistic phenomena.

7 Related Work

A good early survey of work in contextual neural MT is Maruf et al. (2019), who cover work with both RNN and Transformer frameworks along a rich taxonomy.

The transition to neural architectures was a paradigm enabler for document translation, since it eliminated the Markov limitations of statistical MT. 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.

Datasets with document annotations 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. 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 scales to very large web-crawled datasets and shows that parallel data, as a whole, may be harmful.

Contextual metrics work has been important. PROTEST (Guillou and Hardmeier, 2016) used hand-designed pronoun test cases and also evaluated generatively. Läubli et al. (2018) provided early evidence that document-level metrics would be helpful. BlonDe (Jiang et al., 2022), evaluated for Chinese–English, automatically identifies discourse-relevant phenomena in the output and compares to a reference, optionally combined with an n-gram fluency component. Doc-COMET (Vernikos et al., 2022) is simpler and builds sentence representations from context. 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. Fernandes et al. (2023) developed rules to identify contextually-dependent sentences.

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. Our results suggests that parallel data may be of high enough quality for building sentence systems, but may be **harmful when used to build contextual ones**. As an explanation, we consider it a strong possibility that web-crawled parallel data contains too much machine translation output, contaminating the contextual signal. This suspicion makes sense a priori, and is confirmed in other recent work(Thompson 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. In fact, the **failed contextual signal was invisible without such evaluation**. This can be done by using challenge sets (which often mark the expected word) at sufficient scale to cover the noise of the accuracy metric, 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.

Limitations

With respect to reproducibility, the deepest limitation of our paper is our use of private, rather than public, data. As we explained, this was a necessity, since public data does not contain the annotations we need. There is therefore a risk that our findings might not be reproducible by other teams working with (necessarily) different datasets. We have attempted to mitigate this problem by reproducing a subset of our results on publicly available data (Appendix C), where our findings stood. We hope that this corroboration, together with the the fact that harvesting data from the web is itself a well-understood science, help mitigate this risk. Finally, although we suspect our results will hold for language pairs beyond the three we investigated, further complications could arise, and it is possible they will not generalize.

We have focused on adding a missing capability to the traditional MT training pipeline, still actively used in both research and deployment. We leave to future work an in-depth comparison to LLMs on contextual performance using these measures.

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.

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.

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 *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.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

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.

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 Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism.

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.

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.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

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 8 and 9.

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.

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 8: 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 9: Examples of contextually-sensitive auxiliary and inflection elision from the CTXPro dataset.

Table 10 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 for document-context MT. There is more to gain by moving to larger models and likely, to larger datasets and context lengths, as well.

C Results on public data

The full breadth of this paper's experiments was not possible on public datasets; due to the lack of document annotations on large-scale parallel

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 10: 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.

system	COMET	C/Pro	G/Pro
Sent($\mathcal{P}$,$\mathcal{B}$)	60.6	56.7	23.9
Doc($\mathcal{P}$,$\mathcal{B}_d$)	59.4	83.4	64.3

Table 11: Metrics on the only two models we are able to build on public data. Similar patterns are observable to those seen in Tables 4 and 6.

data, we are unable to build Doc($\mathcal{P}_d$,$\mathcal{B}_d$) and Doc($\mathcal{P}_d$,$\mathcal{B}$) systems. However, we can build the Sent($\mathcal{P}$,$\mathcal{B}$) and Doc($\mathcal{P}$,$\mathcal{B}_d$) 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):[12] Europarl v10 (Koehn, 2005), Paracrawl v9 (Bañón et al., 2020), Common Crawl,[13] News Commentary, Wiki Titles 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.[14] 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($\mathcal{P}$,$\mathcal{B}_d$). These are trained for 40 virtual epochs each using the same settings described in Section 5.[15]

Results can be found in Table 11. They are encouraging: we see the same pattern of improvement between Sent($\mathcal{P}$,$\mathcal{B}$) and Doc($\mathcal{P}$,$\mathcal{B}_d$), although the absolute numbers are lower. Compared to our in-house data, the document metrics are even better for Sent($\mathcal{P}$,$\mathcal{B}$).

[12] statmt.org/wmt22/translation-task.html

[13] https://commoncrawl.org/

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

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