Sourcing trustworthy documents for training contextual machine translation systems

Anonymous ACL submission

Abstract

Despite the fact that document context is known to be vital for resolving a range of translation ambiguities, most 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. In this paper, we investigate whether having such annotations would be helpful, even with the knowledge that much of bitext mined from the web may have been translated poorly by humans or by (sentence-level) MT. Working with large-scale parallel and monolingual data sets that we produced in-house, we build large-scale contextual MT systems into German, French, and Russian. We find that contextual MT systems benefit most when document samples are constructed from high-quality back-translated monolingual data *only*. We also show that these improvements are only visible when the systems are evaluated on their *generative* ability on dense test sets, as opposed to *contrastive* discrimination between good and bad examples. The results confirm our suspicion that bitext crawled from the web may be of a quality that is too low to reliably maintain contextual cues for training MT.

1 Introduction

There are two key components to the remarkable advances in the field of natural language processing over the past few years. The first of these is the architecture, which is the original transformer (Vaswani et al., 2017) with a number of incremental tweaks, but importantly, scaled to larger model sizes. The second is the data: training on more and more of it, and extending the basic unit from the single sentences to documents. Encoder-only models such as BERT (Devlin et al., 2019) allowed up to 512 tokens of context, bringing training sample lengths well over ones employed for machine translation even today. Decoder-only large language models (LLMs) extended this to thousands of tokens (Brown et al., 2020), and make use of documents as their basic unit of training.

Instead of simple architectures and document-level data used to train LLMs, machine translation models make almost no use of contextual data, and research tends to focus on complex architecture changes (e.g., Lopes et al. (2020); Yu et al. (2020)). Despite significant prior work on the topic (§ 7), and general acknowledgment of the need to move on (Sennrich, 2018), contextual translation has not managed to take hold, and sentence-level systems continue to dominate. This leaves a gap between them and their increasingly powerful LLM counterparts, which are expanding to larger and larger contexts.

One reason for this is the lack of document-level annotations for MT training data. Although most bitext originates in documents, the typical extraction pipeline drops this information in the cleaning and deduplication process, such that dataset releases remain sentence-based. Similarly, monolingual data used for backtranslation is commonly released without this information. This hampers contextual efforts from the beginning, and the resulting small-data scenarios invite architectural experimentation. While it is assumed that having this information would resolve the problem, the answer may not be that simple. It has long been known that data quality matters as much as quantity (Koehn and Knowles, 2017; Ott et al., 2018), and it has been known even longer that much of the parallel data available on the web is of low quality, whether produced by amateur or underpaid translators or MT systems.

We explore this central problem by building the first large-scale, state-of-the-art standard Transformer model 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:

- Document annotations sourced from parallel and back-translated monolingual data produce large gains in document-level contrastive metrics, where the task is to discriminate correct from manipulated translations (§ 5.2).
- Crucially, however, if we instead evaluate contextual systems generatively (i.e., whether their output correct disambiguates context-sensitive words), the accuracy gains are significantly smaller, but only for systems trained from crawled bitext (§ 5.3).
- Standard sentence-level metrics are much more discriminative between sentence- and contextual systems when applied to datasets that are dense in discourse phenomena (§ 5.4).

Our findings suggest that sourcing documents from crawled parallel data may not be reliable, at least without heavy filtering for quality. In a nod to the importance of open research, we repeat a subset of our experiments on English–German public data (§ 6), corroborating our main result, and suggesting a path for future work.

2 The challenge of data

Publicly available translation datasets typically do not come with document annotations. While the Conference on Machine Translation has made overtures in this direction, 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[1] 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 1):

- Monolingual and parallel data crawled from the web, all containing document metadata.
- CCMatrix parallel data (Schwenk et al., 2021b), which has no document information.

Although the dataset is proprietary, we can say the following about it. There is nothing in it that would surprise any researcher who has experience assembling machine translation datasets for high-resource languages; data has been crawled from the web using standard techniques. The monolingual data sources focus on sites where we expect data to have been written natively. This includes large collections of news data (10%), data linked from the Open Directory Project[2] (40%), filtered webcrawl (40%), and Wikipedia and its outlinks (10%). The parallel data sources include a rough equivalent of Paracrawl as well as CCMatrix. Statistics for the data can be seen in Table 1.

We emphasize that our full set of experiments are not possible with public data, but that we corroborate the subset that are with open data (§ 6).

2.2 Problems with 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 2. The individual sentence pairs are fine for training sentence-level systems, but become problematic when training contextual ones. While we don't know if this was generated by machine or a (tired or underpaid) human, we do know that even large NMT systems

[1] 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

[2] https://odp.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 1: 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.

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, who put love and care into every detail.	Einzigartige maurische **Villa** in einer tropischen Oase mit Pool, Gästeunterkunft und herrlicher Aussicht. ⟨SEP⟩ Referenznummer 1846 ⟨SEP⟩ **Es** wurde vom jetzigen Besitzer gebaut, der Liebe und Sorgfalt in jedes Detail legte.

Table 2: 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.

are sensitive to small amounts of poor data[3] This is all to say that **contextual translation introduces a new quality dimension that is invisible to standard filtering pipelines**, 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, which is target-side native.

3 The challenge of 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. Attempts to automatically identify sentences requiring context have shown the task to be difficult (Bawden et al., 2018) but possible with hand-created rules (Fernandes et al., 2023; Wicks and Post, 2023). Consequently, improvements may be hard to measure and observe with standard metrics.

Fortunately, there exist a range of test sets that have been developed to capture extra-sentential phenomena. By and large, these test sets are *contrastive* ones, where the task is to use a model's score to discriminate between good and bad examples. We begin by cataloguing those that we make use of in this paper (§ 5.2). We then describe a generative extension that makes better use of these contrastive test sets in (§ 3.2).

3.1 Contrastive test sets

The dominant paradigm for evaluation of long-tail document phenomena has been so-called *contrastive evaluation*, 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. Table 3 contains examples of each test set.

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 (Lison and Tiedemann, 2016) with two variants that are identical except that the correct pronoun has been replaced with each of the two incorrect ones.

ContraPro (EN-FR) In the course of evaluating a number of metrics for document MT, Lopes et al. (2020) introduced an extension of the EN-DE ContraPro for EN-FR. Its examples are also drawn from OpenSubtitles, but since French has only two pronouns, there is only one contrastive

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

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 3: Examples from contrastive test sets.

pair per found instance (contrastive pronouns retain the grammatical number of their counterpart).

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.

3.2 Testing generative ability

The challenge sets above test whether a model can discriminate between good and bad examples. As we will show, many document models perform extremely well on these tasks (Table 5), but produce the wrong pronoun when asked to translate the source. 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, we can transform them into generative test sets. We simply translate the source side, in context, and then determine whether the correct pronoun is present in the output. We then compute accuracy over the test set. 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.

4 Experimental setup

All of our models are trained from the parallel ($\mathcal{P}$) and back-translated monolingual ($\mathcal{B}$) data pools. 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[4] on the parallel data.

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 between these two pools uniformly. 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 samples a maximum token length, and concatenates sentences on both sides until this length is reached on the source side, or the document is exhausted. Concatenated sentences are joined with a special ⟨SEP⟩ token, which facilitates sentence alignment at inference time. 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.

Models All of our models are transformers trained with Marian (Junczys-Dowmunt et al., 2018a,b). For each language, we build a single joint unigram subword model (Kudo, 2018) of size 32k. Our experiments with different model capacities (Appendix A) 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 then vary based on whether multi-sentence samples are sourced from the backtranslated data, the parallel data, both, or neither. We compare the following models, 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⋆($\mathcal{P}$,$\mathcal{B}$). A deficient setting that takes the sentence-level baseline and tests it with document-context inputs.[5]
- DOC($\mathcal{P}_d$,$\mathcal{B}_d$). A contextual system, with documents from parallel and back-translated data.

[4] We define a virtual epoch as updates from one billion target-side tokens.

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

- 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.

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 then 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.

Evaluation In addition to the contrastive and generative contextual test suites described in Section 3, we compute COMET[6] (Rei et al., 2020) and BLEU (Papineni et al., 2002) scores, the latter using sacrebleu (Post, 2018),[7] on a WMT test set.[8]

5 Results

5.1 Sentence-level metrics

We begin by establishing baseline scores on standard corpus-level metrics when translating with each model at the sentence level. In addition to a commercial baseline (Microsoft, accessed via API), we present results when translating at both the sentence and document levels. Table 4 contains results for all models translating the test corpora in two modes: without context (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 observe:

- State-of-the-art performance for all models when translating at the sentence level, without context;
- 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($\mathcal{P}$,$\mathcal{B}_d$);
- No consistent improvement in these metrics when adding context at inference time to the DOC($\mathcal{P}$,$\mathcal{B}_d$) system (or other doc systems)

It seems that training with extended context improves the systems' ability to translate, even without context, but on these test sets, there are not widespread gains from employing context at inference time. Section 5.4 suggests this may be because discourse phenomena are too rare in these largely news test sets.

5.2 Contrastive suites

Next, we turn to the document-level contrastive and generative metrics described in § 5.2–3.2.

For generative document metrics, we took special care with SENT⋆($\mathcal{P}$,$\mathcal{B}$). It was not trained with the separator token, so we do not use it when joining sentences for inference. This means that we cannot reliably identify the payload sentence, which complicates evaluation. We work around this by applying the Moses sentence splitter.[9] Spot-checking suggests this to be a reasonable heuristic.

Table 5 contains results for all three language pairs. 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 in the* DOC($\mathcal{P}$,$\mathcal{B}_d$) *system, often by a large margin*. This is especially true for ContraPro and GenPro for EN→DE and EN→FR. Additionally, the SENT⋆($\mathcal{P}$,$\mathcal{B}$) 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.

5.3 A closer look at GenPro

In this section we look closer at the difference between the DOC($\mathcal{P}_d$,$\mathcal{B}_d$) and DOC($\mathcal{P}$,$\mathcal{B}_d$) EN→DE systems in Table 5, which have similar ContraPro scores but divergent GenPro scores. Table 6 provides a breakdown in performance between the two systems by antecedent distance and pronoun type. The systems perform similarly intrasententially (a distance of 0), but are quite divergent when the prediction requires looking into the context. Interestingly, we see that the gains are due to DOC($\mathcal{P}$,$\mathcal{B}_d$)'s ability to correctly predict *sie* and *er*. DOC($\mathcal{P}_d$,$\mathcal{B}_d$) is actually better at predicting *es*. This suggests that it simply overpredicts *es*, the majority pronoun.

We note that GenPro may penalize a system that produces a correct sentence not containing the pro-

[6] COMET version 1.1.3 with model `wmt20-comet-da`; we multiply scores by 100 for readability.

[7] Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`

[8] WMT22 for en-de and en-ru (2,037 lines in 271 docs), and WMT15 for en-fr (1,500 lines in 76 docs).

[9] https://github.com/mediacloud/sentence-splitter

	model	EN→DE BLEU	EN→DE COMET	EN→FR BLEU	EN→FR COMET	EN→RU BLEU	EN→RU COMET
	Microsoft	37.3	62.0	40.8	67.6	33.1	67.3
sent-level	SENT($\mathcal{P}$,$\mathcal{B}$)	37.2	61.6	45.6	69.0	34.0	70.0
sent-level	DOC($\mathcal{P}_d$,$\mathcal{B}_d$)	37.5	62.0	45.1	70.0	34.1	70.4
sent-level	DOC($\mathcal{P}_d$,$\mathcal{B}$)	37.0	61.3	45.4	69.2	33.5	70.0
sent-level	DOC($\mathcal{P}$,$\mathcal{B}_d$)	37.2	**62.2**	44.5	69.8	34.3	**70.2**
doc-level	DOC($\mathcal{P}_d$,$\mathcal{B}_d$)	37.9	62.1	42.5	69.1	34.3	69.2
doc-level	DOC($\mathcal{P}_d$,$\mathcal{B}$)	37.5	62.1	43.6	67.6	33.6	68.5
doc-level	DOC($\mathcal{P}$,$\mathcal{B}_d$)	37.0	62.1	43.1	**70.1**	34.1	**70.6**

Table 4: Metric scores on WMT22/WMT15 test sets when translating as sentences (top block) and with document context (bottom block). Numbers within a column are comparable. A main comparison is SENT($\mathcal{P}$,$\mathcal{B}$)—the sentence level baseline translating sentences—with DOC($\mathcal{P}$,$\mathcal{B}_d$)—the best doc system, translating as documents.

model	EN→DE C/Pro	EN→DE G/Pro	EN→FR C/Pro	EN→FR G/Pro	EN→RU ell_{infl}	EN→RU G/ell_{infl}	EN→RU ell_{VP}	EN→RU G/ell_{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⋆($\mathcal{P}$,$\mathcal{B}$)	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 5: Document contrastive test suites and their generative variants. Contrastive scores (C/*) are over the entire dataset in order to compare with the literature, while generative scores 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⋆($\mathcal{P}$,$\mathcal{B}$) (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.

noun, or unfairly credit a system that happens to generate the pronoun by accident. We do not expect that this will systematically favor any one system, but it does mean that small differences may not be important. Spot-checking suggests to us that the large differences reported in Table 5 capture actual improvements.

5.4 Discourse-dense datasets

Table 4 show modest improvements when translating WMT news test sets as documents, with document systems. What is not clear is what the upper bound on performance is for document-level systems; in other words, how much unrealized gain is there that could have been addressed by contextual translation? This is difficult to answer because we don't know how many document-level phenomena there are in these test sets (in fact, we suspect there are relatively few).

As a means of assessing the question, we turn again to the ContraPro (EN→DE and EN→FR) datasets, which we know to be extremely rich in one particular kind of discourse phenomena: pronoun selection.[10] We take the 12k positive examples along with their references as a *dense* test set. We also create a second, *shifted* test set comprising the set of sentences that occur ten sentences after each sentence in ContraPro. This second test set is likely to be significantly less rich in document phenomena than ContraPro. We then compute COMET scores on these two test sets, translating their sentences with both SENT($\mathcal{P}$,$\mathcal{B}$) and DOC($\mathcal{P}$,$\mathcal{B}_d$).[11]

As we see in Table 7, the best document system (DOC($\mathcal{P}$,$\mathcal{B}_d$)) is *much* better than the sentence base-

[10] We confirm that OpenSubtitles is not in our training data.

[11] Documents have a maximum of 250 SPM tokens and 10 sentences.

	0		1+	
	all	BT	all	BT
all	74.3	73.2	47.7	70.5
es	86.2	81.2	94.3	88.8
sie	72.9	73.6	29.9	64.7
er	61.7	63.5	20.7	58.8
sie|er	67.5	68.7	25.2	61.8

Table 6: Breakdown of ContraGen pronoun prediction accuracy by antecedent distance between two document systems: one ("all") trained on docs from everywhere (DOC($\mathcal{P}_d$,$\mathcal{B}_d$)), and the other ("BT") trained on docs only from BT data (DOC($\mathcal{P}$,$\mathcal{B}_d$)). The former has significantly-lower extra-sentential generative capacity.

	EN→DE		EN→FR	
system	dense	shifted	dense	shifted
SENT($\mathcal{P}$,$\mathcal{B}$)	21.4	31.4	36.2	38.5
DOC($\mathcal{P}_d$,$\mathcal{B}_d$)	27.8	33.9	38.4	39.2
DOC($\mathcal{P}_d$,$\mathcal{B}$)	26.0	34.1	37.6	39.3
DOC($\mathcal{P}$,$\mathcal{B}_d$)	32.4	34.7	40.6	39.6
improvement	+11.1	+3.3	+4.4	+1.1

Table 7: COMET scores on the two OpenSubtitles datasets, the first (dense) discourse-dense, the second (shifted) less so. The gap between translating without and with context is much larger on the discourse-dense subset.

line on ContraPro (+11.1 COMET for EN→DE, +4.4 for EN→FR), suggesting that where document phenomena are rich, a document system's gains can be captured by standard metrics. The gap also exists on the shifted test set (+3.3 EN→DE, +1.1 EN→FR), but is much smaller. Notably, the other document systems are clustered close to DOC($\mathcal{P}$,$\mathcal{B}_d$) on the shifted test set, but trail in the middle on the dense test set. Finally, the gaps are much tighter for EN→FR than for EN→DE, which might suggest this dataset is less discourse-dense, or that the general task—with two pronouns, instead of three—is easier for that language.

Together, these facts suggest a challenge for the evaluation of document-level systems, which is the need to automatically identify sentences that require context to translate correctly.

6 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 data. 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.

	WMT22			
system	BLEU	COMET	C/Pro	G/Pro
SENT($\mathcal{P}$,$\mathcal{B}$)	35.8	60.6	56.7	23.9
DOC($\mathcal{P}$,$\mathcal{B}_d$)	35.8	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 4 and 5.

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 8. 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($\mathcal{P}$,$\mathcal{B}$).

7 Related Work

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

The transition to neural architectures was therefore a paradigm enabler for document translation. Much work, including that with transformers, has focused on separately encoding the con-

[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.

text from the current sentence, in attempts to concentrate the relevant portions of the history and decrease sequence length. 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). Another approach is post-processing approaches inspired by automatic-post-editing but using document-level language models (Voita et al., 2019a). Yu et al. (2020) use Bayes' rule to factor translation and language modeling, translating sentences independently but using a document-level target language model to rerank candidates. Sun et al. (2022) also proposed to use standard transformer models, using small architectures and no monolingual data, and with a "multi-resolutional" training approach that creates overlapping documents.

Standard datasets containing 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. Dobreva et al. (2020) incorporate finer-grained document structure using side constraints and the cache model of Kuang et al. (2018). The idea to draw document data only from monolingual sources has also been tried. 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". They found that it helped, but their models were small. Sugiyama and Yoshinaga (2019) also used target-side data for backtranslation, evaluating in small-data settings with BLEU and contrastive metrics. Our work differs by scaling this to very large web-crawled datasets, and by showing that parallel data, as a whole, may be harmful.

A center point of document-level research is on metrics. PROTEST (Guillou and Hardmeier, 2016) was similar in spirit to our ContraGen. They used hand-designed pronoun test cases and looked for the correct pronoun in the system output. Failure cases were referred to humans for analysis. Läubli et al. (2018) provided early evidence that document-level metrics would be helpful. There has also been recent work in building automatic metrics that make use of context. BlonDe (Jiang et al., 2022) was evaluated in Chinese–English and works by automatically identifying discourse-relevant phenomena in the output and comparing them 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 noted the problem with the disconnect between contrastive evaluation and generative ability for machine translation, but suggest using machine-generated minimal pairs that are closer to model distributions, and don't explore directly measuring generative ability. Fernandes et al. (2023) use translation models to identify sentences that then informed the development of rules to identify contextually-dependent sentences. Wicks and Post (2023) adopt a similar approach that identifies contextual sentences with hand-built rules.

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 shown that parallel data may actually not be a trustworthy source of document training samples, and that good systems can be built with documents sourced from back-translated monolingual data alone, where document annotations are easier to come by. Although we have not investigated *why* this is the case, a reasonable explanation is that it is due to contamination from low-quality sentence-level machine-translated (and potentially also human) translations. 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. Finally, we have shown that standard sentence-level metrics can distinguish between document- and sentence-level systems, so long as they are sufficiently dense in discourse phenomena.

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, 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.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Radina Dobreva, Jie Zhou, and Rachel Bawden. 2020. Document sub-structure in neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3657–3667, Marseille, France. European Language Resources Association.

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, 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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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, Singapore. Empirical Methods in Natural Language Processing.

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. 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.

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.

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 9: 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.

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.

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 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 9 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.