Recovering document annotations for sentence-level bitext

Rachel Wicks[1,2], **Matt Post**[1–3], **Philipp Koehn**[1,2]
[1]Human Language Technology Center of Excellence, Johns Hopkins University
[2]Center of Language and Speech Processing, Johns Hopkins University
[3]Microsoft
rewicks@jhu.edu, mattpost@microsoft.com, phi@jhu.edu

Abstract

Data availability limits the scope of any given task. In machine translation, historical models were incapable of handling longer contexts, so the lack of document-level datasets was less noticeable. Now, despite the emergence of long-sequence methods, we remain within a sentence-level paradigm and without data to adequately approach context-aware machine translation. Most large-scale datasets have been processed through a pipeline that discards document-level metadata. In this work, we reconstruct document-level information for three (ParaCrawl, News Commentary, and Europarl) large datasets in German, French, Spanish, Italian, Polish, and Portuguese (paired with English). We then introduce a document-level filtering technique as an alternative to traditional bitext filtering. We present this filtering with analysis to show that this method prefers context-consistent translations rather than those that may have been sentence-level machine translated. Last we train models on these longer contexts and demonstrate improvement in document-level translation without degradation of sentence-level translation. We release our dataset, PARADOCS, and resulting models as a resource to the community.[1]

1 Introduction

Since the early days of statistical methods, machine translation has been centered within a sentence-level paradigm. N-gram based approaches, which typically obey sentence-boundaries, were the predominant machine translation method and did not effectively use the wealth of information contained across contexts (Mariño et al., 2006). Later, newer neural techniques, such as Transformers (Vaswani et al., 2017b), became popular and have been shown to be effective at handling longer sequences (Beltagy et al., 2020; Sun et al., 2022; Post and Junczys-Dowmunt, 2023). During this time, researchers have periodically considered context-aware machine translation in training methodologies (Voita et al., 2018), evaluation sets and metrics (Jiang et al., 2022; Vernikos et al., 2022; Müller et al., 2018), and its unique ability to address discourse phenomena which are otherwise impossible to correctly translate without context (Voita et al., 2019; Bawden et al., 2018). To present, little work has been done to address the most obvious hurdle: a lack of document-level training data.

Many cornerstone datasets were created by finding known sources of professionally translated documents (Koehn, 2005; Kocmi et al., 2023; Bañón et al., 2020). These parallel documents were then sentence-segmented and aligned with a document-based alignment technique. When more data was demanded, these datasets proved insufficient, so data curators moved towards *global mining*—treating web crawls as a bag of sentences and searching for the most similar sentence in the target language (Schwenk et al., 2021b,a; El-Kishky et al., 2020a). This removes document order and makes document-reconstruction impossible.

Some datasets, such as ParaCrawl, exist as a sentence-level resource but have been constructed in a way more amenable to document-reconstruction. This work confronts this resource gap by providing document-level annotations for News Commentary, Europarl, and the unfiltered ParaCrawl (Bañón et al., 2020) data. Our contributions are:

1. the reconstruction of documents from three large, popular datasets, illustrated in Figure 1;
2. the implementation of a document-level filtering technique as an alternative to traditional bitext filtering which destroys document-level metadata;
3. analysis that shows this filtering prioritizes context-consistent translations;

[1]https://huggingface.co/datasets/jhu-clsp/paradocs

Findings of the Association for Computational Linguistics: ACL 2024, pages 9876–9890

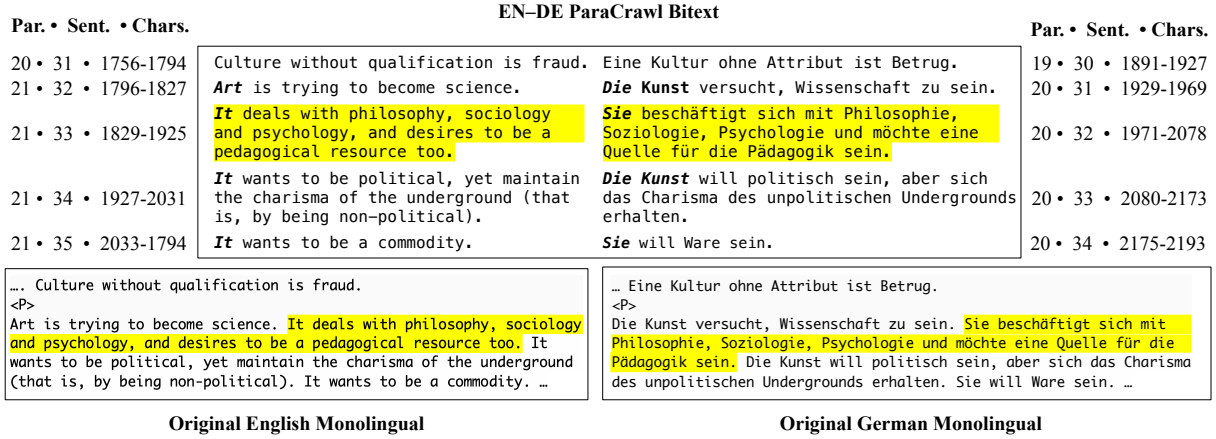

Figure 1: An example from ParaCrawl. The existing bitext has no contextual information. A model is trained to produce "Sie" (a feminine pronoun) from "It" without appropriate context. We restore this information by finding text in the corresponding monolingual dumps, and add document, paragraph, sentence, and character offset metadata.

4. results showing that models trained on this data are better at translating document-level phenomena without degrading sentence-level performance;

5. the public release of this data, PARADOCS, and these models.

2 Related Works

Three areas of research are relevant to this work: sentence-level bitext mining, document-level bitext mining, and context-aware machine translation.

Sentence-level Mining is the current default for most of the largest parallel datasets. In the most recent WMT translation task (Kocmi et al., 2023), parallel training data (i.e., ParaCrawl or WikiMatrix (Bañón et al., 2020; Schwenk et al., 2021a)) tends to be sentence-level. Similarly, of the top ten corpora on OPUS (Tiedemann, 2009; Schwenk et al., 2021b; El-Kishky et al., 2020b, 2021), which make up over 93% of their entire collection, eight[2] are sentence-level (Schwenk et al., 2021b; Fan et al., 2021) and comprises over 72% of the data, with the exceptions being OpenSubtitles (Lison and Tiedemann, 2016) and DGT[3] which have rather specific domains.[4] Many of these datasets are constructed via the global-mining technique, which indexes all sentences using some semantic-based hashing using LASER (Schwenk and Douze, 2017) and aligns based on similarity matches irrespective of document-boundaries.

Document-level Mining is representative of early datasets where document-alignment was already assumed and sentence-alignment could be done via simple features such as sentence lengths (Gale and Church, 1993). This technique was used to create original datasets such as Europarl (Koehn, 2005). Other techniques are constrained by document order when aligning sentences (Varga et al., 2005; Sennrich and Volk, 2011; Thompson and Koehn, 2019). These have been used to create more recent corpora such as News Commentary (Kocmi et al., 2023). ParaCrawl (Bañón et al., 2020) initially aligns documents; however, the preprocessing before the final release destroys any of this preserved document-structure. CCAligned (El-Kishky et al., 2020b) has a middle-ground methodology which first aligns documents determined by both a combination of URL pairs and document similarity, followed by a LASER-based alignment. CCAligned releases data in rough document-order; however there are no annotations for document boundaries or other context labeling.

Recent work has also explored using ParaCrawl as a source for document-level data due to the original alignment being constrained by documents (Al Ghussin et al., 2023). To extract documents, they extract paragraphs by using a subset of the data for which WMT released aligned document-level data in German–English as a document-alignment

[2]NLLB, CCMatrix, MultiCCAligned, ParaCrawl, XLEnt, MultiParaCrawl, LinguaTools-WikiTitles, CCAligned

[3]https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

[4]These statistics current as of February 5, 2024, according to their website: https://opus.nlpl.eu/

benchmark task. This work diverges from this idea in considering the full ParaCrawl release which is a significantly larger portion of data and also expands the language set.

Context-Aware Machine Translation concerns itself with any form of integrating additional context to translate. There are a handful of well-studied discourse phenomena, such as gendered-pronoun translation, which are impossible to translate without incorporating this information (Wicks and Post, 2023; Lopes et al., 2020; Voita et al., 2019; Müller et al., 2018; Bawden et al., 2018). Further, Transformers are entirely capable of incorporating longer contexts as input (Sun et al., 2022). Thus, the field of context-aware translation has either involved modifying the input framework to better incorporate the signal of context-features (Lopes et al., 2020; Tan et al., 2019; Miculicich et al., 2018), or studying alternative uses of training data (Post and Junczys-Dowmunt, 2023; Yu et al., 2020).

The recent introduction of Large Language Models (LLMs) has also had a swath of studies that show these LLMs are better-situated for document-level machine translation as they naturally have significantly large context-windows (Wang et al., 2023; Petrick et al., 2023; Robinson et al., 2023; Kocmi et al., 2023). LLMs provide a desirable path towards document-level translation; however with their novelty and training pipeline opacity, it is still unknown what types of data elicits this performance in order to replicate on different sets of languages or smaller scales.

3 Document Annotation and Reconstruction

Released bitext files rarely have document annotations necessary to create contextual inputs for context-aware training. In the case of Europarl, News Commentary, and ParaCrawl, document ids are released which point to the original monolingual files. We use these monolingual files to reconstruct the original parallel document structure.

We construct annotations for each original aligned segment in the bitext. Each segment has an individual source and target segment which receive separate annotations dependent on their original document. Following the reconstruction process described in Section 3.1, we define a pipeline (Section 3.2) to extract contexts from this data. Finally, we describe a novel document filtering method (Section 3.3) that can be used to filter out low-quality *documents* instead of only low quality *sentences*. With each progression of these steps, one can whittle the dataset smaller while improving the relative quality of documents.

All filtered splits, and the original unfiltered data, are publicly released to facilitate further research into context-aware machine translation, and document-based data selection and filtering methods. The sizes of each dataset with the increased filtering is displayed in Table 1. We explore reconstruction as an alternative to re-alignment of the original documents since it is less computationally intensive and re-uses existing annotations from published datasets. Future work can circumvent the need for reconstruction by preserving document metadata while data mining.

3.1 Reconstruction

If the end goal is to extract "context", it is only necessary to determine whether two source-target pairs are consecutive in the respective source and target documents. Given an ordered list of segments[5] from the unfiltered bitext, and the original web-crawled document, we can align each segment to the original document via exact string matching. Illustrated in Figure 1, this determines the index spans for each segment. These indices identify where precisely the segment occurs in the original monolingual document. This is relevant to identify the correct context as some sentences may have been left unaligned during the sentence-alignment process. We annotate the spans for all source and target segments in our data. For each segment in the RAW bitext, we list (with respect to the monolingual document):

- the paragraph index;
- the sentence index, determined by applying a Moses Sentence Splitter (Koehn et al., 2007);
- the starting character index, after normalizing whitespace;
- the ending character index, after normalizing whitespace;
- the probability of the language id according to NLLB's fasttext LID model;
- the duplication count: the number of times this segment was repeated in ParaCrawl.[6]

[5]We use "segment" to refer to a given source or target which roughly equates to one sentence but may vary.

[6]Only ParaCrawl has significant boilerplate text so we

		RAW	DOCS	LOOSE$_{75\%}$	MEDIUM$_{50\%}$	STRICT$_{25\%}$	SENTS
DE	# segs.	6.16B	161M	127M	87.9M	41.9M	257M
	# docs.	-	45.4M	34.1M	22.7M	11.4M	-
FR	# segs.	3.51B	115M	90.5M	62.4M	39.9M	231M
	# docs.	-	32.3M	24.2M	16.1M	8.07M	-
ES	# segs.	5.69B	137M	110M	76.7M	36.9M	189M
	# docs.	-	37.9M	28.5M	19.0M	9.5M	-
IT	# segs.	2.88B	40.4M	31.8M	22.0M	10.M	112M
	# docs.	-	12.8M	9.58M	6.38M	3.19M	-
PL	# segs.	1.09B	23.2M	17.9M	12.2M	5.85M	34.2M
	# docs.	-	7.55M	5.66M	3.77M	1.89M	-
PT	# segs.	2.49B	52.1M	41.1M	28.6M	13.8M	91.1M
	# docs.	-	15.0M	11.2M	7.5M	3.7M	-

Table 1: Data sizes. RAW is the portion of data we attempted to align to an original monolingual document. DOCS is the portion of data that meets a minimal document criteria (at least two consecutive segments; > 0.5 langid; < 100 duplications). LOOSE$_{75\%}$, MEDIUM$_{50\%}$, and STRICT$_{25\%}$ is the top 75%, 50%, and 25% (respectively) of documents scored using SLIDE-based filtering. # segs is the number of segments (roughly sentences) while # docs. is the number of distinct *sub-documents* created as described in Section 3.2.

Consecutive sentences are defined such that the starting character index is two more than the ending index of the previous segment. For instance, in Figure 1, "Art is trying to become a science." has a period that ends at index 1827 and "It" of the following sentence starts at 1829. This indicates there is a single whitespace token separating these two segments at index 1828 and they were originally consecutive; thus, they belong to the same context.

In Table 1, we report the size of the data we reconstruct by RAW. This data precludes approximately one-third of the ParaCrawl data for which the monolingual source is CommonCrawl. The CommonCrawl monolingual data is only accessible by querying servers which make recovering metadata annotations computationally difficult with respect to network issues and latency. We leave this section of ParaCrawl unannotated.

3.2 Context Extraction

The benefit of sentence-level bitext creation is that high-quality sentence-pairs are not discarded simply because the remaining document is poorly aligned. This allows for much higher rates of alignment, and a larger overall dataset. Unfiltered ParaCrawl documents (defined by a URL pair) are often poorly aligned, or have extraneous portions of documents in one or both languages. We still want to keep a *paragraph* of context even if the boilerplate text is of low quality. To achieve this, we define *document-breaking criteria*. As we iterate over the annotated data, we accumulate context within document boundaries. When a given segment fails to meet some criteria, we break the preceding and proceeding contexts into *sub-documents*. Consider an article which may have been loosely translated. Some paragraphs may be literally translated while others may have been paraphrased or given additional context to make it more understandable to the audience. In a given paragraph, when one segment is *unaligned*, we break the paragraph into two *sub-documents*: those preceding the unaligned segment and those proceeding it These sub-documents can be still be linked by their parent document, but can be treated as independent for the sake of context extraction. In our experiments, we break on three conditions:

1. a segment is unaligned (minimum sub-document length is two)
2. the language id probability is less than 0.5 (as predicted by NLLB's fasttext)
3. the duplication count is more than 100—this divides documents on boilerplate texts which have high frequency and do not contribute much meaningful contextual information.

leave News Commentary and Europarl unannotated.

The resulting dataset is titled DOCS and statistics can be seen in Table 1.

3.3 Document Filtering

A key issue of web-crawled bitext is that much of it was machine translated (Thompson et al., 2024). Sentence pairs that are *well* translated will still circumvent sentence-level filtering methods such as LASER. When these pairs sneak through, any resulting model will be nothing more than a mixture of distilled MT models from throughout the history of the internet. As machine translated sentences on the internet are likely to have been translated at the sentence-level, this makes any resulting translation particularly prone to making errors when when translating context-phenomena (Müller et al., 2018; Bawden et al., 2018). Thus, it is necessary to filter out these errors before training.

Peter et al. (2023) showed that quality estimators (QEs), which notably do not require a reference, are capable of distinguishing fine-grained quality differences necessary for filtering. Additionally, SLIDE (Raunak et al., 2023) showed that these same quality estimators can discriminate between context-consistent and context-inconsistent translations—as one might see when translating each sentence individually. We combine these two ideas to propose a document-filtering methodology.

SLIDE works by creating a series of context-chunks by sliding a window across the document. Each window is evaluated by a QE system and the scores are averaged in order to create a document-level score. Raunak et al. (2023) evaluates a combination of window and stride sizes. They find some effectiveness starting at a minimum window size of three. We use a window of three and a stride of one for our scoring. We also chose to use the CometKiwi QE model (Rei et al., 2022) as it is consistently a high performing model in these works.

The initial dataset is large, and much of it is low quality. We rank documents with this scoring technique, and experiment with three different filtering cutoffs, top 75%, 50%, and 25%, scored at the *sub-document* level (as described in Section 3.2). They are described as LOOSE$_{75\%}$, MEDIUM$_{50\%}$, and STRICT$_{25\%}$, respectively, in Table 1.

4 Source Data

To produce our dataset, we select source data from three large, publicly-available datasets that initially used document-level alignment.

ParaCrawl (Bañón et al., 2020) prioritizes a high-recall alignment so the original unfiltered data is orders of magnitude larger than the official release, but has an inferior quality. For instance, ParaCrawl v9 `en-de` is approximately 278M lines of cleaned sentences in arbitrary order, but an additional "`RAW`" file is available that contains nearly 10B of uncleaned sequential sentence alignments.[7] The official release has undergone the aforementioned sentence-level filtering (deduplication, similarity score filtering, etc) that removes the original context. Fortunately, ParaCrawl releases an original unfiltered (`RAW`) data version. This equates to each source–target pair, a document id, and a pointer to the original monolingual source.[8]

News Commentary (Kocmi et al., 2023) is a smaller, albeit cleaner newswire dataset released annually for WMT training data.[9] The released versions *do* maintain document order, but documents are not labeled.

Europarl (Koehn, 2005) was produced from the *Proceedings of the European Parliament* which is obligatorily translated into a handful of European languages. Europarl is typically n-way parallel which makes it ideal for machine translation despite the specific domain. The most recent version (v10) releases document ids; however, it is not available in all languages. We produce the alignment ourselves with the accompanying tools.

5 Experimental Design

In order to show the viability of constructing a dataset in such a fashion, we need to show two things: (1) a contextual model is at least as good at sentence-level translation quality as a sentence-level model and (2) a contextual model outperforms sentence-level models when considering context-based phenomena.

5.1 Baselines

We consider two types of sentence-level models for baselines. For each data filtering level, we additionally train sentence-level models without concatenation. This allows for comparison on models

[7] Most of these segments are justifiably discarded during preprocessing due to low quality which is why the v9 release is substantially smaller.

[8] The monolingual sources for two-thirds of ParaCrawl was released at `https://paracrawl.eu/moredata`.

[9] With special thanks to Barry Haddow who was kind enough to deliver us copies of the intermediate processing steps so we could recreate these annotations

		SENT.	DOCS			LOOSE$_{75\%}$			MEDIUM$_{50\%}$			STRICT$_{25\%}$		
Training Type:		snt.	snt.	context		snt.	context		snt.	context		snt.	context	
Inference Type:		snt.	snt.	snt.	ctx.	snt.	snt.	ctx.	snt.	snt.	ctx.	snt.	snt.	ctx.
EN-DE	W23	39.8	40.2	39.1	40.5	40.6	40.6	42.1	40.6	40.6	**42.3**	40.9	40.0	41.9
	FLO.	39.7	39.8	38.8	40.5	40.6	40.3	**41.1**	40.2	40.2	41.0	40.7	40.5	40.7
EN-FR	W15	41.5	41.6	41.8	42.3	41.9	41.8	42.5	41.9	41.9	**42.9**	41.9	41.5	42.1
	FLO.	51.9	51.6	51.5	52.5	51.7	51.3	52.1	52.2	52.4	**52.6**	52.1	51.0	52.0
EN-ES	W13	36.0	36.3	36.2	36.2	36.1	36.1	36.1	**36.6**	36.2	36.4	36.3	36.2	36.2
	FLO.	27.5	27.8	27.9	28.1	27.9	28.0	**28.4**	28.0	27.8	28.0	28.0	28.2	28.3
EN-IT	W09	33.2	33.3	33.3	33.2	33.6	33.4	33.6	33.7	33.4	**33.8**	33.2	32.8	33.3
	FLO.	29.8	29.1	29.2	29.1	29.8	29.8	29.8	**30.3**	29.7	29.8	30.0	29.9	29.7
EN-PL	W20	25.6	25.3	24.9	24.5	**26.0**	25.2	**26.0**	25.9	25.5	**26.0**	25.9	25.0	25.0
	FLO.	**22.7**	22.1	21.8	21.6	22.5	22.0	21.8	22.4	22.2	22.0	22.3	21.7	21.8
EN-PT	FLO.	51.1	50.2	49.8	50.1	50.4	50.1	**51.4**	51.2	49.9	50.5	50.7	49.4	50.4

Table 2: BLEU scores on evaluation sets. The top row indicates the training data and its filtering level. SENTS is all of our data filtered through a bitext-filtering pipeline where as DOCS, LOOSE$_{75\%}$, MEDIUM$_{50\%}$, and STRICT$_{25\%}$ only include the top 100%, 75%, 50%, and 25% of documents scored under a SLIDE–CometKiwi filtering metric (Section 3.3). We indicate whether sentences were concatenated (contextual) or isolated (sentences) during training. We similarly indicate inference input.

trained on the same quantity and distribution of data. Additionally, we produce a *new* dataset that is filtered at the *sentence level* instead of the document level. Using the entirety of the original data, we filter using traditional bitext filtering pipeline. After deduplication, we remove sentences with: empty lines, more than fifty-percent punctuation, irregular frequencies of characters based on language histograms (Fan et al., 2020), uneven length ratios (greater than 1.5) of source-target, less than 0.5 of target language according to NLLB's fasttext and langid.py, and those below a 0.85 LASER score (Schwenk and Douze, 2017). The resulting dataset is described as SENTENCES in Table 1.

5.2 Training

All models trained in this paper are trained with the Marian NMT (Junczys-Dowmunt et al., 2018) toolkit. We train context-aware models with a simple concatenation strategy (Tiedemann and Scherrer, 2017). We also train on both contexts and sentences during training via SOTASTREAM which mixes two data streams during training (Post et al., 2023). The first stream which samples from the PARADOCS data, concatenates documents up to ten sentences or 256 tokens (whichever is lesser) and inserts an `<eos>` token to separate segments. The second stream pulls from a supplementary dataset composed of preprocessed sentence-level bitext. These streams are mixed with a 1:1 ratio. These models are trained with typical next-token prediction and no further alterations.

The models are Transformers (Vaswani et al., 2017a) with 12 encoder layers and 6 decoder layers. We use a feed-forward dimension of 16,384 and an embedding size of 1024. We train a single `sentencepiece` vocabulary (Kudo and Richardson, 2018) for each language pair trained on the supplementary data with a shared vocabulary size of 64k. All model iterations uses the corresponding vocabulary. The effective batch size is 500k tokens and one logical epoch is 1B tokens. We train for 10 logical epochs for a total of 10B tokens/20k updates. For evaluation, we use the model with the lowest cross-entropy loss per token on the FLORES200 dev set (NLLB Team, 2022; Goyal et al., 2021; Guzmán et al., 2019).

Supplementary Data is used in addition to the our contextual data. This ensures that the model can translate both sentences and documents, and is not burdened by a lack of language coverage from high quality sentence pairs mined via other techniques (i.e., global mining that produces high quality datasets such as CCMatrix). This supplementary data is controlled across experiments. We filter this bitext in the same way we filter the SENTS baseline described in Section 5.1.

5.3 Evaluation

We evaluate two-fold: sentence-level translation quality and the ability to address context-dependent phenomena. The former can be addressed by regu-

		SENT.	DOCS			LOOSE$_{75\%}$			MEDIUM$_{50\%}$			STRICT$_{25\%}$		
Training Type:		snt.	snt.	context		snt.	context		snt.	context		snt.	context	
Inference Type:		snt.	snt.	snt.	ctx.	snt.	snt.	ctx.	snt.	snt.	ctx.	snt.	snt.	ctx.
EN-DE	Gen.	45.5	44.6	43.3	57.7	44.8	45.4	58.1	45.8	44.0	58.5	45.5	40.8	**60.4**
	For.	40.4	42.5	41.5	42.2	43.0	41.4	43.6	41.8	41.5	43.3	42.1	41.5	**44.0**
	Aux.	4.4	4.7	3.6	7.5	4.9	4.8	8.3	3.9	3.6	7.2	4.6	4.4	**9.6**
EN-FR	Gen.	39.5	40.4	40.1	48.4	40.1	39.7	48.7	39.7	39.7	**49.7**	39.6	39.6	**49.7**
	For.	39.7	38.6	37.9	38.9	38.9	39.8	**42.7**	39.9	39.5	42.0	38.8	39.0	39.8
	Aux.	0.9	0.9	0.9	7.7	0.8	0.9	10.6	0.9	0.8	10.3	1.2	0.9	**11.6**
EN-ES	Gen.	39.3	37.8	37.5	39.3	36.9	38.2	37.3	38.4	37.8	35.7	37.9	38.2	**42.9**
	For.	**35.2**	34.5	34.2	32.6	34.3	34.5	31.4	34.9	34.5	30.2	34.7	33.8	16.3
	Aux.	0.9	1.0	0.8	9.6	1.0	1.0	10.0	1.0	0.9	10.9	1.6	0.9	**12.5**
EN-IT	Gen.	53.0	52.6	51.5	58.0	53.2	53.2	59.2	54.7	53.4	**61.2**	53.7	52.7	60.2
	For.	35.2	34.8	34.8	33.6	**36.0**	35.9	34.3	35.9	35.8	35.7	35.7	35.1	34.6
	Aux.	1.5	1.7	1.4	5.7	1.8	1.7	5.8	1.8	1.8	7.0	1.8	1.7	**9.9**
EN-PL	Gen.	30.9	31.3	31.2	36.8	32.0	31.3	37.1	32.6	31.4	37.6	31.8	31.0	**38.5**
	For.	30.3	30.6	30.9	31.6	31.4	30.7	30.5	31.4	31.4	**32.8**	30.8	29.9	31.9
	Aux.	4.2	4.7	4.7	15.5	6.4	5.2	**18.3**	5.3	6.6	17.0	6.2	4.4	12.3
	Inf.	33.3	37.6	37.0	40.5	37.0	37.5	40.9	39.2	37.1	**41.3**	38.3	36.6	39.9
EN-PT	Gen.	36.3	37.3	37.5	42.4	37.1	37.8	43.8	38.2	39.2	46.3	39.5	37.7	**47.0**
	For.	21.1	22.7	**22.5**	20.8	22.2	22.3	21.1	21.7	22.2	20.7	21.9	22.1	20.5
	Aux.	1.4	2.3	1.5	18.5	3.8	1.6	18.9	3.7	1.6	**24.7**	4.6	1.4	23.3

Table 3: CTXPRO scores on evaluation sets. The top row indicates the training data and its filtering level. SENTS is all of our data filtered through a bitext-filtering pipeline where as DOCS, LOOSE$_{75\%}$, MEDIUM$_{50\%}$, and STRICT$_{25\%}$ only include the top 100%, 75%, 50%, and 25% of documents scored under a SLIDE–CometKiwi filtering metric (Section 3.3). We indicate whether sentences were concatenated (contextual) or isolated (sentences) during training. We similarly indicate inference input.

lar machine translation test sets. We choose to use FLORES200 test sets (NLLB Team, 2022; Goyal et al., 2021; Guzmán et al., 2019) as they are available in all of the languages in consideration. We additionally evaluate on the most recent WMT test set available for that language pair. For these evaluation sets, we report BLEU (Papineni et al., 2002) scored using SACREBLEU (Post, 2018). We additionally report COMET (Rei et al., 2020) in Appendix A but acknowledge that those scores may be inflated from the COMET-based filtering method.

In order to evaluate the translation of context-dependent phenomena, we turn to CTXPRO (Wicks and Post, 2023) which releases evaluation sets in these languages to evaluate three phenomena: gender, formality, and auxiliary verbs. Briefly, these handle the translations of English 'it' to gendered languages; the translations of English 'you' to languages with a T-V distinction; and the translation of English auxiliary verbs in elided sentences such as 'I do.' These are impossible to consistently translate correctly without context. We report accuracy scores as calculated by the CTXPRO PyPI package.

To translate, we employ two inference designs. First, for all models we translate using single sentences as input. Second, for the contextual models, we additionally evaluate them under contextual inference. For each input, we concatenate up to ten sentences or 256 tokens (whichever is lesser) of preceding context. This parallels the contextual data during training. The last segment is then split (determined by <eos>) and used as the predicted translation.

6 Results

In general, we find that context-aware models trained with the PARADOCS data outperform any model trained without contextual inputs across evaluation metrics. Further, we find that the document-level filtering method is able to improve performance even as data quantity diminishes—indicating a higher quality of data selection. In Table 2, we display BLEU scores as an evaluation of general translation ability and subsequently in Table 3, we display CTXPRO accuracy scores as an evaluation of the ability to translate context-dependent phenomena. We additionally perform paired bootstrap resampling statistical tests on the evaluation sets to understand whether these differences are meaningful ($p < 0.05$) and comment

where appropriate (Koehn, 2004).

6.1 Effects of Document-Level Filtering

In Table 2, we show BLEU performance of all models. We compare the contextual models with the sentential baselines. Within each inference setting, performance steadily, and consistently improves between the models trained on the noisiest data (DOCS) and the cleanest ($\text{STRICT}_{25\%}$). This is true even when *not* using the contextual information available as shown in the sentence-level inference setting of the contextual models. This is reinforced by the same trend in CTXPRO scores displayed in Table 3. We note for BLEU scores, there is a small plateau or deterioration between the $\text{MEDIUM}_{50\%}$ and $\text{STRICT}_{25\%}$ data filtering stages and suspect this is a critical trade off point between having high-quantity, high-quality sentence translations (that are *poorer*-quality contextual translations) and having low-quantity, high-quality contextual translations. This is supported by the fact that we do not see this trend across the CTXPRO scores.

6.2 Sentence-Level Translation Performance

We also find that by training a model with context-awareness, the model does not lose the ability to translate stand-alone sentences and in many cases, marginally improves. This is evident by comparing the the models trained under each paradigm (Sentence, and Context Training), when only given sentences during *inference*. In Table 2, this is evident as we see little-to-no difference between the models trained with and without context when they translate individual sentences. In some cases, we see marginal improvement from the model trained with context. This indicates that training a machine translation within a context-aware paradigm is *no worse than* training one without.

6.3 Context-Awareness Boosts BLEU Performance

Further, we identify benefit from leveraging additional context to translate, even when translating datasets which are not dense with context-dependent phenomena, represented by WMT test sets and Flores200. Contextual models that are given preceding context during inference get consistent small gains in BLEU compared to their analogous sentence-level models in Table 2. In the larger language pairs (e.g., German and French), we additionally found that the top performing models were statistically outperforming *all* of the sentence-level models however the statistical differences degraded with the smaller language pairs.

6.4 Translation of Context-Based Phenomena

Finally, we show that these context-trained models are *more effective* at translating context-dependent phenomena than their sentence-level counterparts. When considering both Gender and Auxiliaries, the improvement is evident. We do find the translation of Auxiliaries is still unfortunately low, but may be due to a relatively *low* rare number of occurrences in the dataset. We also see that in the case of formality, performance is marginally better, though relatively unaffected. We hypothesize this is due to the particularities of formality translation, rather than exemplification of the dataset, though acknowledge that alternatives may be necessary to target this particular phenomena. When investigating the statistical differences, we note specifically that in all cases, the context-aware models are statistically superior than their sentence-level counterparts specifically for the auxiliary class. This is particularly noteworthy as the auxiliary task is the hardest task: the potential correct answer is open to the set of all verbs rather than a simple one-of-three-genders problem as we see with formality. We also note there is less statistical support for the formality class. This speaks to the difficulty of this problem and may instead indicate that data is not the simplest solution to approach formality translation.

7 Analysis of Document Filtering

Document-level filtering is rare, and to our knowledge, has not been studied. We demonstrate that our SLIDE–CometKiwi filtering improves contextual performance, but this may be an indication of better sentence-level translations and not better document-level translations. For a document-level filtering method to hold up against web-crawled data, it should not only filter out poor quality translations, but should also filter out context-inconsistent translations. By context-inconsistent, we refer to translations that were produced at the sentence-level. This is roughly synonymous with machine translated text in a web-crawl setting; however, there is no robust machine translation detection methodology to the best of our knowledge.

We instead turn towards a proxy measure to determine if context-inconsistent translations are filtered out. After scoring each document as de-

scribed in Section 3.3, we can rank documents sequentially based off these SLIDE-CometKiwi scores. We should find more context-consistent translations in the upper percentiles of this distribution and fewer at the lower-end.

	QUARTILES			
	1st	2nd	3rd	4th
Inter. Fem.	46%	30%	17%	6.4%
Inter. Masc.	33%	41%	20%	5.8%
Inter. Neut.	30%	36%	24%	10%
Intra. Fem.	54%	29%	13%	3.5%
Intra. Masc.	39%	39%	18%	4.5%
Intra. Neut	35%	35%	21%	9.4%

Table 4: For each category of gendered pronoun, what percent of all examples occur in each quartile defined by SLIDE–CometKiwi scores. Intersential indicates the antecedent occured in the same sentence while intrasential indicates it occured in a different sentence.

The `ctxpro` toolkit, which generated these evaluation sets, also identifies when there are ambiguous uses of gendered pronouns. As the toolkit is intended to be high precision, low recall, it is extremely likely to not identify ambiguity in pronouns when the translations are incorrect or inconsistent. Thus, we can use the `ctxpro` toolkit as a proxy measure for context-consistent translations.

We identify all examples of gendered-pronouns in our data according to the `ctxpro` toolkit, and investigate where SLIDE-CometKiwi ranks them in relation to the remaining documents. We show percentages in Table 4. We discriminate between intersentential—where the antecedent is in the same sentence—and intrasentential—where the antecedent is in a previous sentence. The former is quite easy to correctly translate for a sentence-level model while the latter is impossible. We further distinguish by the gender of the pronoun—where neuter is the majority class.

As shown in Table 4, most of the identified examples occur in the 1st or 2nd quartile. Further, we see that this is *especially* true for the feminine pronouns (a quintessential minority class for pronouns). Conversely, neuter pronouns have a more uniform distribution across quartiles. We hypothesize this indicates there is more machine translated texts in the lower quartiles—as they would have incidentally correctly translated neuter pronouns—and less in the top. More specifically, the *most challenging* translation for machines, ambiguous pronouns with an intrasentential feminine antecedent, are mostly ranked in the top quartile. We assume this also means the top quartile was translated by humans.

8 Conclusions

Research in contextual machine translation is hampered by the lack of document annotations on parallel data. We augment three large popular MT datasets (ParaCrawl, News Commentary, and Europarl) with this information, creating a document-level dataset, PARADOCS. We introduce a document-level filtering method to apply to this data in lieu of traditional context-destroying sentence filtering methods. Simple, context-aware machine translation models trained on this data have shown to be better at machine translation in both general performance as measured by WMT test sets as well as targeted performance—measured by the ability to correctly translate discourse phenomena. We release the data as a resource to the community.[10]

9 Limitations

As mentioned in this work, document reconstruction is particularly constrained by the original dataset processing. Many languages have only been processed with the global mining technique (i.e., CCMatrix) and ParaCrawl notably only supports European languages. This works also assumes there exists many well-translated parallel documents in web-crawled corpora. Not only is this not true for many language pairs, but there is recent evidence to suggest that the more multi-way parallel data is, the more likely it was machine translated (Thompson et al., 2024).

This also significantly constrains the amount of data. In our smallest setting, English–German (an extraordinarily high-resource language pair) is still limited to 42M lines. If we were to extend this to lower-resource languages, we would be limited to perhaps a few thousand lines which are unlikely to make any meaningful difference in performance.

10 Acknowledgements

We would like to thank the reviewers for their discussion and feedback. We additionally would like to thank Barry Haddow, Marc Marone, Neha Verma, and Joe McKnight for their resources, technical discussions, and support.

[10] https://huggingface.co/datasets/jhu-clsp/paradocs

References

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1304–1310, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020a. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco GuzmÃ¡n, and Philipp Koehn. 2020b. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

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. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 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).

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.

José Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

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.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

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.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2023. Document-level language models for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 375–391, Singapore. 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.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Slide: Reference-free evaluation for machine translation using a sliding document window.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte

Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. 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 and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. 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 and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

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.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level 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 1576–1585, Hong Kong, China. 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.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

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

Dániel Varga Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP*, pages 590–596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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. 2019. 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.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. 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.

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.

A Appendix

Dataset	# Lines
OPUS-ccaligned-v1-deu-eng	9.2M
OPUS-ccmatrix-v1-deu-eng	167M
OPUS-news_commentary-v16-deu-eng	222k
OPUS-paracrawl-v9-deu-eng	140M
OPUS-wikimatrix-v1-deu-eng	1.1M
OPUS-wmt_news-v2019-deu-eng	35k
total	317M

Table 5: English–German supplementary sentence-level bitext. Names based on the published `mtdata` name. https://github.com/thammegowda/mtdata. Flores200 dev and devtest specifically removed from this data before training.

Dataset	# Lines
OPUS-ccaligned-v1-eng-spa	8.6M
OPUS-ccmatrix-v1-eng-spa	345M
OPUS-elrc*	1.5M
OPUS-europarl-v8-eng-spa	1.7M
OPUS-multiccaligned-v1-eng-spa	30M
OPUS-multiparacrawl-v7.1-eng-spa	54M
OPUS-news_commentary-v16-eng-spa	38k
OPUS-paracrawl-v9-eng-spa	154M
OPUS-ted2020-v1-eng-spa	299k
OPUS-wmt_news-v2019-eng-spa	11k
OPUS-wikimatrix-v1-eng-spa	2.6M
OPUS-wikipedia-v1.0-eng-spa	1.3M
OPUS-wikimedia-v20210402-eng-spa	871k
total	600M

Table 6: English–Spanish supplementary sentence-level bitext. Names based on the published `mtdata` name. https://github.com/thammegowda/mtdata. Flores200 dev and devtest specifically removed from this data before training.

Dataset	# Lines
OPUS-ccaligned-v1-eng-fra	9.4M
OPUS-ccmatrix-v1-eng-fra	263M
OPUS-elrc*	2.9M
OPUS-europarl-v8-eng-fra	1.8M
OPUS-news_commentary-v14-eng-fra	118k
OPUS-paracrawl-v9-eng-fra	137M
OPUS-ted2020-v1-eng-fra	276k
OPUS-wikimatrix-v1-eng-fra	2.2M
OPUS-wikimedia-v20210402-eng-fra	657k
OPUS-wikipedia-v1	417k
OPUS-wmt_news-v2019-eng-fra	19k
total	418M

Table 7: English–French supplementary sentence-level bitext. Names based on the published `mtdata` name. https://github.com/thammegowda/mtdata. Flores200 dev and devtest specifically removed from this data before training.

Dataset	# Lines
OPUS-ccaligned-v1-eng-ita	8.9M
OPUS-ccmatrix-v1-eng-ita	121k
OPUS-elrc*	1.0M
OPUS-news_commentary-v16-eng-ita	67k
OPUS-multiccaligned-v1-eng-ita	20.4M
OPUS-paracrawl-v9-eng-ita	6.1M
OPUS-wmt_news-v2019-eng-ita	2.4k
OPUS-wikimatrix-v1-eng-ita	1.7M
OPUS-wikipedia-v1.0-eng-ita	307k
OPUS-wikimedia-v20210402-eng-ita	232k
total	215M

Table 8: English–Italian supplementary sentence-level bitext. Names based on the published `mtdata` name. https://github.com/thammegowda/mtdata. Flores200 dev and devtest specifically removed from this data before training.

Dataset	# Lines
OPUS-ccaligned-v1-eng-pol	6.1M
OPUS-ccmatrix-v1-eng-pol	50M
OPUS-elrc*	1.3M
OPUS-europarl-v8-eng-pol	520k
OPUS-multiccaligned-v1-eng-pol	8.1M
OPUS-multiparacrawl-v7.1-eng-pol	8.7M
OPUS-paracrawl-v9-eng-pol	21.9M
OPUS-ted2020-v1-eng-pol	107k
OPUS-wikimatrix-v1-eng-pol	370k
OPUS-wikipedia-v1.0-eng-pol	100k
OPUS-wikimedia-v20210402-eng-pol	31k
OPUS-elra*	122k
OPUS-kde4-v2-eng-pol	37k
OPUS-dgt-v2019-eng-pol	2.1M
total	102M

Table 9: English–Polish supplementary sentence-level bitext. Names based on the published `mtdata` name. https://github.com/thammegowda/mtdata. Flores200 dev and devtest specifically removed from this data before training.

Dataset	# Lines
OPUS-ccaligned-v1-eng-por	7.3M
OPUS-ccmatrix-v1-eng-por	147M
OPUS-elrc*	1.4M
OPUS-europarl-v8-eng-por	1.7M
OPUS-multiccaligned-v1-eng-por	13M
OPUS-multiparacrawl-v7.1-eng-por	22M
OPUS-news_commentary-v16-eng-por	48k
OPUS-paracrawl-v9-eng-por	54M
OPUS-ted2020-v1-eng-por	227k
OPUS-wikimatrix-v1-eng-por	2.0M
OPUS-wikipedia-v1.0-eng-por	1.0M
OPUS-wikimedia-v20210402-eng-por	363k
total	253M

Table 10: English–Portuguese supplementary sentence-level bitext. Names based on the published `mtdata` name. https://github.com/thammegowda/mtdata. Flores200 dev and devtest specifically removed from this data before training.

		SENT.	DOCS			LOOSE$_{75\%}$			MEDIUM$_{50\%}$			STRICT$_{25\%}$		
Training Type:		snt.	snt.	context		snt.	context		snt.	context		snt.	context	
Inference Type:		snt.	snt.	snt.	ctx.	snt.	snt.	ctx.	snt.	snt.	ctx.	snt.	snt.	ctx.
EN-DE	W23	79.9	80.4	81.4	81.5	81.7	80.1	81.0	81.5	81.1	80.5	81.7	**82.2**	81.8
	FLO.	87.2	87.2	87.8	87.7	**88.1**	86.9	87.5	88.0	87.7	87.7	87.9	87.9	87.8
EN-ES	W13	86.1	86.2	86.0	86.1	86.1	86.1	86.1	**86.3**	**86.3**	86.1	**86.3**	86.2	86.1
	FLO.	86.0	86.0	86.0	86.0	86.2	86.1	86.2	86.3	86.3	86.1	**86.5**	86.3	86.2
EN-FR	W15	83.5	83.8	84.0	84.2	84.1	83.6	84.1	**84.3**	83.9	83.1	83.4	84.1	83.5
	FLO.	88.0	87.7	88.0	88.2	88.4	87.9	88.2	**88.5**	88.1	88.0	88.0	88.1	88.0
EN-IT	W09	86.6	86.5	86.5	86.5	86.9	86.7	86.9	**87.1**	86.9	86.9	86.9	86.6	86.6
	FLO.	87.7	87.4	87.3	87.0	87.8	87.8	87.6	**88.2**	88.0	87.9	88.0	87.8	87.6
EN-PL	W20	86.7	86.3	85.5	83.8	**87.3**	86.4	86.5	87.1	86.8	86.9	86.9	86.3	86.3
	FLO.	88.0	87.4	87.0	86.9	88.1	87.8	87.8	**88.4**	88.1	88.0	88.2	87.8	87.6
EN-PT	FLO.	89.6	89.1	89.0	89.1	89.3	89.4	89.7	**89.8**	89.5	89.7	**89.8**	89.4	89.4

Table 11: COMET scores (x100) on evaluation sets. The top row indicates the training data and its filtering level. SENTS is all of our data filtered through a bitext-filtering pipeline where as DOCS, LOOSE$_{75\%}$, MEDIUM$_{50\%}$, and STRICT$_{25\%}$ only include the top 100%, 75%, 50%, and 25% of documents scored under a SLIDE–CometKiwi filtering metric (Section 3.3). We indicate whether sentences where concatenated (contextual) or isolated (sentences) during training. We similarly indicate inference input.