Document-Level Machine Translation with Large-Scale Public Parallel Corpora

Anonymous ACL submission

Abstract

Despite the fact that document-level machine translation has inherent advantages over sentence-level machine translation due to additional information available to a model from document context, most translation systems continue to operate at a sentence level. This is primarily due to the severe lack of publicly available large-scale parallel corpora at the document level. We release a large-scale open parallel corpus with document context extracted from ParaCrawl in five language pairs, along with code to compile document-level datasets for any language pair supported by ParaCrawl. We train context-aware models on these datasets and find improvements in terms of overall translation quality and targeted document-level phenomena. We also analyse how much long-range information is useful to model some of these discourse phenomena and find models are able to utilise context from several preceding sentences.

1 Introduction

Machine translation has traditionally been framed as a problem of translating source text to target text one sentence at a time. However, depending on the languages and the content of the text being translated, it is often the case that a sentence is impossible to translate well in isolation – that further contextual information is required. Maruf et al. (2021) summarise several discourse phenomena that are impossible for sentence-level machine translation systems to deal with – including anaphoric pronouns, lexical cohesion, deixis, and ellipsis. There are also features like grammatical gender, number, style, and formality, that can sometimes not be determined from the individual sentence but are dependent on surrounding context. Therefore, it has been clear for many decades (Bar-Hillel, 1960) that a machine translation system cannot translate some sentences without the ability to capture other linguistic cues from context. Läubli et al. (2018) also showed that while sentence-level neural machine translation can appear high-quality out of context, human evaluators had a much stronger preference for human translation when evaluating translation at the document level.

As a result, there have been many efforts to incorporate document context into neural machine translation. What almost all of these methods have in common is that they require parallel training data with document context.

ParaCrawl (Bañón et al., 2020) produced large-scale parallel corpora and the released data includes information about the URLs from which the sentences were extracted, but the released corpora were only sentence-level. We use raw webpage text publicly available from ParaCrawl along with the officially released sentence-level corpora to assemble large-scale document-level parallel corpora for several language pairs. We release our code to generate the document-level datasets[1] as well as the datasets in five selected language pairs[2].

We then validate the usefulness of our datasets by training context-aware translation models for all of these language pairs, and find that models that are aware of target context perform better than sentence-level baselines, often in terms of overall translation quality, but more significantly when evaluated with respect to targeted discourse phenomena. We experiment with varying the amount of preceding context available to these context-aware models at test time, and find that while information from the immediately previous sentence is most useful – as is intuitively obvious, longer-range context up to a certain extent can also help models translate phenomena like anaphoric pronouns more accurately.

[1] Link redacted for anonymity

[2] More language pairs may be released soon; it requires a resource-intensive but simple process of running our code on any language pair released by ParaCrawl.

2 Related Work

Many attempts to utilise document context in machine translation have introduced specialised architectures to encode context (Tu et al., 2018; Jean et al., 2017; Kuang et al., 2017; Maruf and Haffari, 2018; Voita et al., 2018; Miculicich et al., 2018) alongside source sentences.

Some methods like Junczys-Dowmunt (2019); Sun et al. (2022); Post and Junczys-Dowmunt (2023) eliminate sentence boundaries altogether and use a standard transformer architecture to translate an entire chunk of text as a single sequence. However, unless they retain some sentence markers like in Junczys-Dowmunt (2019) or Tiedemann and Scherrer (2017), these models can be difficult to evaluate with our existing evaluation paradigms and metrics due to the dependence on sentence-level test sets. In those cases, we often have to rely on sentence-splitting heuristics and alignment methods just to be able to compute a sentence-level metric on the model outputs. As a result, our work chooses a method to translate with one sentence at a time as input, but with document context provided as a separate additional input. This allows the model to benefit from context information while still being simple to evaluate with existing metrics and test sets.

There are few existing parallel corpora of significant size that retain document metadata – examples are Europarl (Koehn, 2005) which had only around 2M sentences for the largest language pairs; CzEng (Kocmi et al., 2020) containing 61M sentence pairs with document annotation along with more than 100M synthetic sentence pairs, but only for eng↔ces; OpenSubtitles (Lison and Tiedemann, 2016); and NewsCrawl, where the latter two datasets are relatively large corpora for several language pairs but restricted in domain.

The CCAligned corpus (El-Kishky et al., 2020) includes hundreds of millions of comparable document pairs across many languages, from which sentence-level datasets were extracted and released. A dataset with sentence pairs and corresponding document contexts was not created; however, it should be possible to extract a similar dataset as the one we present here from the available data released by CCAligned which includes documents, URLs, and sentences.

Closest to our data, Al Ghussin et al. (2023) used publicly available parallel document metadata from ParaCrawl[3] to extract aligned paragraphs of text, and used these paragraphs as a proxy for documents. Even though their extracted datasets were at a relatively small scale due their use of only a subset of ParaCrawl data and strict filtering, they observed clear improvements in targeted evaluations of document-level translation phenomena.

Post and Junczys-Dowmunt (2023) showed that using only monolingual documents and back-translating (Sennrich et al., 2016) them sentence by sentence to produce synthetic document pairs can surprisingly produce better results than using actual document pairs to train a document-level model. Their results, however, were mostly on unreleased private data, and their comparison could not be reproduced on public data precisely because of the absence of public datasets of adequate size.

Another recent orthogonal approach is to use large language models' inherent ability to model long context to perform document-level translation (Wang et al., 2023; Zhang et al., 2023; Karpinska and Iyyer, 2023) with no or very few parallel training examples. While this paradigm is gaining popularity, it is yet to be comprehensively explored, and the need remains to have large-scale datasets of parallel text with document context.

3 Dataset

At the time of its release, ParaCrawl (Bañón et al., 2020) was the largest publicly available sentence-level parallel corpora for most of the languages it supported. The ParaCrawl corpus mining process included steps to match documents that were estimated to be translations of each other, from which sentences were extracted and aligned, but unfortunately, the released corpora did not preserve document context or structure, and only contained isolated sentence pairs along with the source URLs they were originally extracted from.

However, separately, a lot of the raw text crawled from the web was also released[4] as language-classified base64-encoded text with their corresponding URLs. Therefore, we were able to match the webpage contents to their URLs in the sentence-level parallel corpora to recover the corresponding documents.

To build document-level parallel datasets from these sources of data, we chose five language pairs – Czech (ces), Polish (pol), German (deu), French

[3] https://www.statmt.org/paracrawl-benchmarks/
[4] https://paracrawl.eu/moredata

Lang.	Sentences	Source	Target	Both
deu	278.3	105.6	110.3	92.1
fra	216.6	83.5	86.3	72.2
ces	50.6	18.7	21.0	16.3
pol	40.1	16.8	18.4	14.9
rus	5.4	3.1	2.8	2.4

Table 1: Sizes of our document-level datasets in millions of lines. "Sentences" is the size of the original ParaCrawl sentence-level datasets. "Source/Target" denotes the subset of sentence pairs where there is at least one source/target context – eng is always considered the source language in this case. "Both" denotes the subset of sentence pairs with at least one source context and one target context. Note: the eng-rus dataset is significantly smaller because it was not part of the ParaCrawl main release, but a smaller "bonus" release.

(fra), and Russian (rus), all paired with English (eng) – and used the following method:

1. Extract the source URLs and corresponding sentences from the TMX files from ParaCrawl release 9[5] (or the bonus release in the case of eng-rus). Each sentence is usually associated with many different source URLs, and we keep all of them.
2. Match the extracted URLs with the URLs from all the raw text data and get the corresponding base64-encoded webpage/document, if available.
3. Decode the base64 documents and try to match the original sentence. If the sentence is not found in the document, discard the document. Otherwise, keep the 512 tokens preceding the sentence (where a token is anything separated by a space), replace line breaks with a special `<docline>` token, and store it as the document context. Since some very common sentences correspond to huge numbers of source URLs, we keep a maximum of 1000 unique contexts per sentence separated by a delimiter `|||` in the final dataset.
4. Finally, we compile three different files per language pair – a dataset with all sentence pairs where we have one or more source contexts, one with all sentence pairs with target contexts, and a third dataset with both contexts.

Even though the TMX files have source URLs for all released sentences, this process was lossy due to a few different reasons:

- ParaCrawl was compiled from a number of separate crawls or "collections", and there were inconsistencies in how URLs were formatted in intermediate steps. We employed some basic heuristics to match as many URLs as possible, such as removing `http://` and `https://` and trailing slashes before matching, but there is a still a chance some URLs were missed in this process.
- Data from CommonCrawl was not duplicated in the released raw text from ParaCrawl. To avoid re-downloading huge amounts of data, any URLs that were present in CommonCrawl but not in the other collections are missing from our dataset.
- There were many instances where the original sentence could not be found in the contents of a webpage corresponding to its source URL. This is most likely due to the same URLs being crawled at different times and finding dynamic or possibly entirely changed content.

The sizes of our extracted datasets are shown in Table 1. They will be made publicly available at [redacted link for anonymity].

4 Document-level Translation Models

To evaluate the usefulness of our datasets, we train document-level translation models using only our datasets. Even though higher-quality document-level training data exists at a smaller scale, we choose to train our models only on data from ParaCrawl in order to accurately evaluate the quality and utility of our data and to make a fair comparison with our sentence-level baselines.

At training time, for each input example, we first sample one context out of up to 1000 that are present in the document-level dataset. To ensure that the models are capable of using variable-length context at test time, we uniformly sample a context length l from $\{1, \ldots, 256\}$ for each training example. We then retain at least l tokens from the preceding context, possibly exceeding the limit to avoid mid-sentence splits. We then use the source sentence as the main model input, the sampled context as a second input (see detailed model architecture in Section 4.1), and the target sentence as the target model output.

We train separate models for each language pair using either source or target context information. While the model is always provided ground-truth

[5] https://paracrawl.eu/releases

context at training time, this is not always available in the case of target context at test time, so we test using both ground-truth target context and real predicted output context. However, we note that one of the most common use cases of machine translation is in the context of computer-assisted translation (CAT) tools, where translators can see preceding context but typically machine translate and post-edit one sentence at a time, as a result of which gold-standard target context is available for each sentence.

4.1 Model Architecture and Training

We use the dual-encoder transformer architecture from Junczys-Dowmunt and Grundkiewicz (2018) but without tied parameters between the two encoders, implemented in the Marian framework (Junczys-Dowmunt et al., 2018). In other words, we modify a standard transformer encoder-decoder model (Vaswani et al., 2017), which takes the source sentence as input and produces the target sentence as output, to add a second encoder which takes additional source/target context as input. This is similar in spirit to Zhang et al. (2018), but we do not incorporate the context encoding into the source encoding, instead directly feeding both encodings to the decoder. As shown in Figure 1 of Junczys-Dowmunt and Grundkiewicz (2018), the decoder has two stacked cross-attention sub-layers to attend to the two encoder contexts. We use default transformer-big hyperparameters. This choice of architecture is less complex than the specialised architectures described in Section 2, but still allows for separating the current sentence and context inputs, giving us greater control over evaluation and interpretability compared to models which translate an entire large chunk of text at a time. Moreover, the addition of the second encoder automatically accounts for the need for extra model capacity to encode the document context.

We train context-aware models for the following language pairs: eng→deu, eng→fra, eng→rus, eng→ces, and pol→eng.

We train our models with dynamic batch size to make optimal use of GPU memory. We train all models on 4 or 8 Nvidia A100 or 3090 GPUs, using gradient accumulation to ensure that the average effective batch sizes are approximately equivalent in each case. We validate every 50 million target tokens for eng→rus and every 500 million target tokens for all other language pairs, early stopping when cross-entropy calculated on the validation set does not improve for 10 consecutive validations.

4.2 Test Data for Evaluation

To assess the translation quality of our models, we perform two kinds of evaluation – general machine translation quality metrics, and using contrastive evaluation to measure the accuracy of the models on targeted discourse phenomena.

4.2.1 General Translation Quality Metrics

We compute standard sentence-level quality metrics – BLEU (Papineni et al., 2002) using the sacreBLEU implementation[6] (Post, 2018) and COMET[7] (Rei et al., 2022) – on the following WMT test sets, all of which were released with document metadata: WMT22 eng→deu and eng→ces (Kocmi et al., 2022), WMT23 eng→rus (Kocmi et al., 2023), WMT20 pol→eng (Barrault et al., 2020), and WMT15 eng→fra (Bojar et al., 2015).

4.2.2 Contrastive Evaluation

Contrastive test sets consist of input text and translations which appear correct at the sentence level, but may be wrong given more context. Models are evaluated by their ability to assign higher probability to the sentences that are correct in context. We evaluate our models on a few different contrastive test sets which measure the following types of discourse phenomena for specific language pairs.

- **Anaphoric pronouns**: The ContraPro test sets for eng→deu (Müller et al., 2018) and eng→fra (Lopes et al., 2020) translation evaluate the accuracy of pronoun translation where the source English sentence does not contain enough information to determine the correct pronoun in the target language and context is required to generate the correct translated pronoun. One part of the DiscEvalMT test set (Bawden et al., 2018) also evaluates anaphoric pronoun translation in eng→fra.
- **Deixis and ellipsis**: Good Translation Wrong in Context (GTWiC) (Voita et al., 2019) is a collection of contrastive test sets to evaluate a number of discourse phenomena in eng→rus translation, among which are deictic expressions, verb phrase ellipses, and correct inflection of nouns which depend on elided verbs.
- **Lexical choice**: The DiscEvalMT contrastive test set from Bawden et al. (2018) tests lexical choice in eng→fra translation where the

[6] `BLEU|#:1|c:mixed|e:no|tok:13a|s:exp|v:2.4.0`
[7] Specifically wmt22-comet-da

Model	BLEU / COMET				
	eng→deu	eng→fra	eng→ces	eng→rus	pol→eng
Sentence-level	35.2 / 85.4	40.5 / 83.1	36.8 / 88.4	**22.8 / 75.4**	**33.5 / 83.7**
Subset - source	35.0 / 85.5	–	36.3 / 87.5	22.0 / 74.8	32.6 / 83.3
Subset - target	34.3 / 85.3	40.6 / 83.1	35.9 / 87.8	22.0 / 75.3	32.4 / 83.2
Source context	34.9 / 85.0	–	36.6 / 88.1	19.4 / 72.4	32.4 / 83.0
Gold target context	**37.4 / 85.9**	**42.6 / 83.2**	**37.3 / 88.5**	21.9 / **75.4**	32.8 / 83.3
Predicted target context	34.7 / 85.4	40.5 / 82.8	35.4 / 87.1	21.5 / 74.9	32.8 / 83.4

Table 2: Overall sentence-level BLEU/COMET scores on test sets for models in different configurations. Bold text highlights the highest score for each language pair. "Subset - source" and "Subset - target" are sentence-level baselines trained on the subsets of sentences that have source or target contexts respectively, i.e. the same number of training examples as the corresponding context-aware models. "Gold target context" and "Predicted target context" are the same model which encodes target-side context, but the latter uses the predictions from previous lines in the same document as its context instead of the ground-truth context.

choice of certain words/phrases in translation is ambiguous without context information. It also tests lexical cohesion, i.e. the repetition of translated entities when the entity is repeated on the source side. An eng→ces extension of the lexical cohesion subset was created by Jon (2019). GTWiC also has a similar subset to test lexical cohesion in eng→rus.

5 Results and Analysis

We train sentence-level and document-level models in a few different configurations for comparison. Our baseline is a standard sentence-level transformer-big model trained on all of the ParaCrawl parallel data for a given language pair. We also train sentence-level baselines trained on the subsets of sentence pairs for which source or target contexts could be extracted, thus ensuring a fair comparison in terms of the number and content of training examples. Finally, for each language pair, we train two different document-level models: one which is aware of source context and one which uses target context. We further test the target context model in two different scenarios: using original ground-truth context for each test example and using the actual model output from previous sentences within a document as target context.

5.1 Effect on Overall Translation Quality

One of the ways we evaluate our document-translation models is simply in terms of overall sentence-level translation quality metrics. The results are summarised in Table 2.

We find that while source context does not seem to benefit overall translation quality, or at least not in a way that is reflected in these metrics, using the ground-truth target context generally improves translation quality over the baseline using the same number of training examples, i.e. "Subset - target" compared to "Gold target context" in Table 2. This improvement is not observed for eng→rus, probably due to the small number of training examples in the subset of sentences with target context not being enough to train a high-quality context-aware model. The sentence-level baseline using the full set of ParaCrawl sentence pairs ("Sentence-level") out-performs the context-aware models for the smaller language pairs, benefitting from having a much larger number of training examples.

A machine translation system in an environment where translated output is post-edited sentence by sentence, such as in CAT tools, has access to ground-truth target context for every line, and can thus benefit from the improved translation quality of the context-aware model. However, a fully automated document-level translation pipeline does not have this information available. To reproduce this scenario, we also try using actual model predictions from previous sentences within a document as target context ("Predicted target context" in Table 2), and find that this does not yield the same improvement as using ground-truth context. This is a manifestation of exposure bias (Ranzato et al., 2016) due to the models only being trained on ground-truth contexts and not being robust enough to accurately use the relatively noisy predicted context, and could possibly be remedied by scheduled sampling (Bengio et al., 2015) to expose some gen-

erated context to the model during training.

5.2 Accuracy on Contrastive Test Sets

We also perform evaluations on selected contrastive test sets for some language pairs, as mentioned in Section 4.2.2. Each example in a contrastive test set has a source sentence and source/target context along with two or more possible outputs, one of which is correct. We use our models to score all the possible outputs and say the model gets an example right if it assigns higher probability to the correct output than to the other options. We then calculate accuracy over the entire test set. Since each contrastive test set is different and designed for specific languages, we discuss them separately in this section.

5.2.1 ContraPro (eng→deu and eng→fra)

We use ContraPro (Müller et al., 2018) to evaluate the accuracy of pronoun translation for our eng→deu models. This test set contains examples of sentence pairs where the source English sentence contains the pronoun *it* which needs to be translated into one of *es*, *sie*, or *er* in the target German. The pairs are designed so that provided context information is required to determine the correct translation of the pronoun. Each example has two translations which are both apparently correct at the sentence level but one of them uses the wrong pronoun in context. Our models are not required to generate translations, only to score each alternative output given the source sentence and ground-truth target context. If the model is able to take the context into account, it should assign higher probability to the option with the correct pronoun.

A similar test set created by Lopes et al. (2020) evaluates the same phenomenon in eng→fra translation, where the English *it* is translated to *il* or *elle* in French and *they* is translated to *ils* or *elles*.

We find that while our eng→deu sentence-level model has an accuracy of 0.507, i.e. approximately random chance, the model with target context scores 0.784 when provided 5 sentences of preceding context. This shows that the model learns to accurately disambiguate the correct choice of pronouns using the context information.

For eng→fra, we find that the sentence-level model already achieves a reasonably high accuracy of 0.815, due to the fact that the antecedents of the pronouns are located within the same sentence in approximately 43% of the examples in this test set. Using our target context-aware model, we still see an increase in accuracy to 0.824 given a single sentence of preceding context, but no further improvement with longer context.

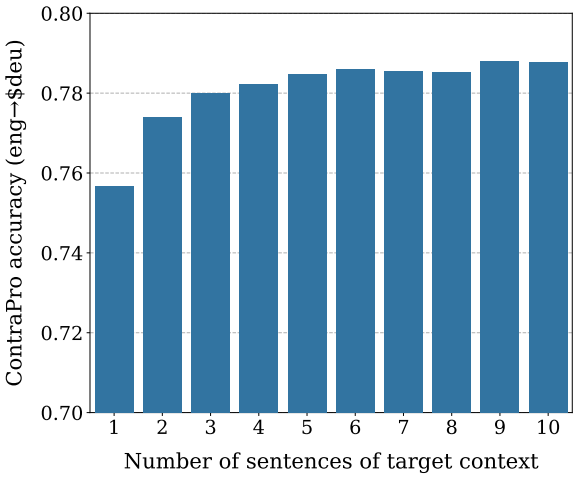

Figure 1: Effect of varying the number of lines of target context on ContraPro eng→deu pronoun translation accuracy. The baseline accuracy achieved by the sentence-level model is 0.507. Accuracy increases steadily with up to 3 or 4 sentences of target context with only marginal gains beyond that.

Effect of Context Size Figure 1 shows the effect of the amount of context that is exposed to the eng→deu model on its ability to accurately translate the anaphoric pronouns in ContraPro. We find that a single sentence of target context is enough to significantly increase the accuracy of pronoun translation beyond a sentence-level baseline's 0.507, and context longer than 3 sentences does not make much of a further difference. This makes intuitive sense, since the antecedent of a pronoun is most often in the immediately preceding sentence and only very rarely more than 2 or 3 sentences away.

5.2.2 Good Translation Wrong in Context (eng→rus)

Voita et al. (2019) introduced a collection of test sets for contrastive evaluation of a range of discourse phenomena in eng→rus translation. The test set is thus divided into 4 subsets:

- **Deixis:** These examples are related to gender and formality marking in Russian that are absent in the source English. The model needs to use context to translate these deictic words or phrases correctly.
- **Ellipsis:** These examples have elliptical constructions in the English text that cannot be elided in Russian, so the translation needs to

Model / Context Length	GTWiC Accuracy			
	Deixis	Ellipsis		LC
		Infl.	VP	
Sentence-level	0.5	**0.5**	0.058	0.458
Trg context/1	0.586	**0.5**	0.07	0.468
Trg context/2	0.654	0.494	0.07	**0.472**
Trg context/3	**0.692**	**0.5**	**0.074**	**0.472**

Table 3: Accuracy of our target context-aware model on the GTWiC test sets with varying number of sentences of target context. "LC" denotes lexical cohesion. While performance on deictic expressions improves steadily with more context, lexical cohesion only improves very marginally, and verb phrase (VP) ellipsis accuracy remains very low.

Model	eng→fra	eng→ces
Lexical Choice:		
Sentence-level	0.5	0.5
Target context	**0.525**	**0.533**
Anaphora:		
Sentence-level	0.5	–
Target context	**0.545**	–

Table 4: Accuracy of target context-aware models compared to sentence-level models on the DiscEvalMT test sets in eng→fra and eng→ces. Context-aware models achieve higher accuracy in each case.

expand the ellipsis. There are two kinds of ellipsis-related errors targeted here: where the target text has wrong morphological inflection due to missing information from the source ellipsis, and where the wrong verb is generated for a verb phrase ellipsis.

- **Lexical cohesion:** These examples evaluate the ability of the model to ensure that named entities that are repeated in the source are translated consistently in the output. Models need to be aware of preceding context to translate cross-sentential repetitions consistently.

Unlike ContraPro, contrastive examples in GTWiC have several incorrect translations and one correct translation for each given source sentence and context. A test example is considered correct if our translation model scores the correct translation higher than all of the incorrect translations. We report accuracies separately for each GTWiC test set evaluating different phenomena.

The performance of our target context-aware model on the GTWiC test sets is reported in Table 3. We observe that the context-aware model is significantly more capable of translating deictic expressions accurately. However, we find that they do not perform well on the ellipsis test sets, with verb phrase ellipsis accuracy surprisingly being worse than chance, and improvements on lexical cohesion are also marginal. This is possibly because the models need *both* source and target context to be able to model these phenomena accurately. For example, it is difficult for the model to be aware that an entity should be repeated if is not aware that both the preceding source and target contexts had occurrences of the same entity.

A maximum of 3 context sentences is available per example in GTWiC, and we once again find that having more target context can be useful for the model to translate deictic expressions correctly, but the benefits diminish as the model usually gets adequate context from the last one or two sentences.

5.2.3 DiscEvalMT (eng→fra and eng→ces)

The DiscEvalMT eng→fra contrastive test sets evaluate two document-level translation phenomena:

- **Anaphora:** Similar to ContraPro, this test set also contains examples where correctly generating a pronoun in the target French requires the model to use context information about the antecedent.
- **Lexical choice:** Some of these examples contain an ambiguous word in the source and the model needs to be able to use context information to disambiguate the correct sense of the word and translate it correctly. The rest of the examples test models' ability to consistently translate a repeated word or phrase.

The eng→ces extension of DiscEvalMT only includes the lexical cohesion test set.

These test sets also have two alternative translations for each input sentence and context, and our models are expected to score the correct option in context higher than the other option.

We can see in Table 4 that while our document-level models do not score very highly on this benchmark, they still out-perform the sentence-level baseline. Similar to GTWiC, for the lexical cohesion test sets, we believe that the models would perform better if they were able to access both source and target contexts, since they are otherwise unaware of the repeated word or phrase.

6 Conclusion

We release large-scale document-level parallel corpora in five language pairs extracted from the ParaCrawl datasets in an effort to mitigate the dearth of publicly available machine translation training data with document context. We also open-source code to enable the community to compile such datasets in more language pairs. Due to both the ParaCrawl pipeline and our code being open-source, it is theoretically possible to create document-level datasets for any supported language by crawling the web, and not just those already released by ParaCrawl.

While we treat any preceding text at the same URL as "context", it is often the case that these are completely unrelated to a given sentence, such as UI elements, boilerplate text, or entirely unrelated content on the same webpage. Future work should also explore filtering these datasets to retain only genuine contextual information, which is likely to be much more useful to the model, although even content that is not strictly document context may help guide translation through indirect domain or style cues.

Our document-level translation experiments show that models aware of target context improve in terms of overall translation quality as well as in terms of evaluation of some targeted discourse phenomena compared to a sentence-level baseline. We show that machine translation can benefit from multiple sentences of preceding context to accurately translate discourse phenomena like anaphoric pronouns, although very long-range context is rarely useful.

7 Limitations

Relevance of context Our work assumes that any extracted text preceding a given sentence on a webpage is relevant "document context" for that sentence. However, it is likely in many cases that the extracted context is unrelated to the sentence, since most webpages are not formatted as a coherent "document". As a result, the dataset often includes irrelevant context like lists of products, UI elements, or video titles extracted from webpages which will not be directly helpful to document-level translation models.

Availability of both contexts Our models are all trained with either source or target context being available to the models, but for some document-level phenomena like lexical consistency of repeated named entities, it is probably necessary for the model to be aware of both source and target context. Our datasets make it possible for future work to extract training data with both contexts and train such models.

Model quality Our models are trained only on noisy ParaCrawl data and tested on high-quality WMT data. While there are much smaller but relatively high-quality document-level training datasets available, all our experiments were conducted only on ParaCrawl data to test the quality of the datasets without being influenced by other data. As a result, these models are not necessarily the strongest possible translation models, but they are useful to fairly and clearly compare document-level machine translation against sentence-level models.

Language coverage ParaCrawl was focused on European Union languages with only a few "bonus" releases for other languages. Moreover, most of the corpora were for English-centric language pairs. Due to the high computational requirements to extract these corpora, our work further chose only a subset of these languages, resulting in corpora for only a few European languages, some of them closely related. Given the availability of raw data and tools to extract such corpora for many more languages from all over the world, we hope the community is encouraged to build such resources for a much larger variety of language pairs. Document-level translation phenomena also vary widely by source and target language, so such experiments for more languages is left for future work.

8 Ethical Considerations

Harmful content The main released corpora from ParaCrawl were filtered to remove sensitive content, particularly pornography. Due to pornographic websites typically containing large amounts of machine translated text, this filtering also improved the quality of the resulting corpora. However, when we match sentences with their source URLs, it often happens that an innocuous sentence was extracted from a webpage with harmful content, and this content is present in our document contexts. We may release filtered versions of these corpora in the future, pending further work to filter harmful content at the document level.

Eurocentricity The Eurocentric nature of our work is remarked upon in Section 7. Due to most

large-scale publicly available parallel corpora being English-centric, almost all machine translation research remains English-centric, at the cost of the majority of the world's language users. This limits both the generalisability and usefulness of a lot of research. We hope the release of more data and tools helps the community expand these efforts to more languages.

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.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. In *Advances in computers*, volume 1, pages 91–163.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, 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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

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

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. 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.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Josef Jon. 2019. discourse-test-set. `https://github.com/cepin19/discourse-test-set`.

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

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. 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.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. 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.

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.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

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

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Cache-based document-level neural machine translation.

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).

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.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

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.

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

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.

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

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

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

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.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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.