Do We Need Source Context for Document-level Neural Machine Translation?

Anonymous EACL submission

Abstract

Standard context-aware neural machine translation (NMT) typically relies on parallel document-level data, exploiting both source and target contexts. In this work, we investigate whether source context data could actually be dispensed altogether within a standard concatenation-based approach to context-aware NMT, thus supporting further use of monolingual data without the need for a specific NMT architecture. We propose a simple approach based on prepending context sentences of the target language to both the source sentence to be translated and the target reference sentence. We show that this method can lead to significant improvements over a strong baseline on discourse-level phenomena that depend on target language information, while achieving parity for phenomena where the relevant information is present in both source and target languages. Additionally, we show that target monolingual data can be better exploited via back-translation under this approach, and that the use of machine-translated target context did not significantly impact translation quality overall. We experimented in two language pairs, English-Russian and Basque-Spanish, for which challenge test sets are available on multiple contextual phenomena.

1 Introduction

Significant progress has been achieved in Machine Translation within the Neural Machine Translation (NMT) paradigm (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). For the most part though, most NMT models translate sentences in isolation, preventing the adequate translation for document-level phenomena such as cohesion, discourse coherence or intersentential anaphora resolution (Bawden et al., 2018; Läubli et al., 2018; Voita et al., 2019b; Lopes et al., 2020; Post and Junczys-Dowmunt, 2023).

Several research paths have been explored to design context-aware NMT models that can exploit the available context to provide more accurate translation. Among the main approaches are input augmentation via concatenation of context sentences (Tiedemann and Scherrer, 2017), alternative NMT architectures (Jean et al., 2017; Zhang et al., 2018; Voita et al., 2019b; Li et al., 2020; Bao et al., 2021), or, more recently, pretrained large language models (Wu et al., 2022; Wang et al., 2023). Among these approaches, simple concatenation of context sentences, as initially proposed by Tiedemann and Scherrer (2017), remains a solid baseline typically used in practice with varying amounts of source-target context pairs (Agrawal et al., 2018; Junczys-Dowmunt, 2019; Majumder et al., 2022; Post and Junczys-Dowmunt, 2023; Sun et al., 2022).

Context-aware models typically rely on parallel document-level data, a scarce resource overall despite recent efforts to provide this type of resource (Barrault et al., 2019; Voita et al., 2019b; Gete et al., 2022). To the exception of approaches such as the monolingual repair framework of Voita et al. (2019a), context data in the source language is generally used to model context-awareness. However, most, if not all, discourse-level phenomena feature information that is either present mainly in the target language (e.g., lexical cohesion, deixis) or in both the source and target languages (e.g., gender selection, ellipsis). Considering this, we aimed to further explore the use of target language data in isolation, dispensing with source data altogether, within a standard NMT architecture to avoid the need for additional architectural complexity, processes and resources.

Our approach consists in simply prepending, at training time, context sentences from the target language to both the source sentence to be translated and the reference translation, discarding source context data altogether. The underlying intuition is that target-side context information would still help model contextual phenomena at the decoder level, whereas, on the encoder side, it will be either ignored and copied, as foreign data, or associated with source information to further model context. At the same time, this approach supports the use of a standard architecture and approach to context-aware NMT. We show that this use of target context data on both sides of the training pairs can provide significant improvements over the use of source data in combination or in isolation, for discourse-level phenomena that depend on target-language information, while achieving parity for phenomena where the relevant contextual information is present in both the source and target languages.

We establish our results on two language pairs, English-Russian and Basque-Spanish, for which contrastive test sets are publicly available on a range of phenomena that depend on either only the target language or both the source and target languages. In addition to accuracy results on specific phenomena, we compare overall translation quality on parallel test sets as well. We also measure the impact of using reference vs. machine-translated output as context at inference time, with only minor loss in our experiments. Finally, we evaluate the use of back-translated data, with similar comparative gains as when using parallel document-level data. Overall, our experimental results indicate that using only target context data within a standard NMT architecture can be a promising alternative for context-aware machine translation.

2 Related Work

An increasing number of studies centred on context-aware NMT approaches have demonstrated that significant improvements can be achieved over non-contextual baselines, for typical discourse-level linguistic phenomena (Li et al., 2020; Ma et al., 2020; Lopes et al., 2020; Fernandes et al., 2021; Majumder et al., 2022; Sun et al., 2022).

One of the first methods proposed for the task is the concatenation of context sentences to the sentence to be translated, in either the source language only, or in both source and target languages (Tiedemann and Scherrer, 2017; Agrawal et al., 2018). This method does not require any architectural change and uses a fixed contextual window of sentences. It provides a robust baseline that often achieves performances comparable to that of more sophisticated methods, in particular in high-resource scenarios (Lopes et al., 2020; Sun et al., 2022; Post and Junczys-Dowmunt, 2023). Variants of this approach include discounting the loss generated by the context (Lupo et al., 2022), extending model capacity (Majumder et al., 2022; Post and Junczys-Dowmunt, 2023) or encoding the specific position of the context sentences (Lupo et al., 2023).

Alternative approaches include refining context-agnostic translations (Voita et al., 2019a; Mansimov et al., 2021) and modelling context information with specific NMT architectures (Jean et al., 2017; Li et al., 2020; Bao et al., 2021). More recently, the use of pretrained language models has been explored for the task, using them to encode the context (Wu et al., 2022) or to initialize NMT models (Huang et al., 2023). Other studies directly use Large Language Models to perform translations, showing that competitive results can be obtained with this approach, although they might still make critical errors in certain domains and sometimes perform worse than conventional NMT models (Wang et al., 2023; Karpinska and Iyyer, 2023; Hendy et al., 2023).

Concatenation-based approaches vary regarding their use of context, exploiting either the source context (Zhang et al., 2018; Voita et al., 2018), the target context (Voita et al., 2019a) or both (Bawden et al., 2018; Agrawal et al., 2018; Xu et al., 2021; Majumder et al., 2022). The benefits of using context sentences in both the source and the target languages are also discussed in Müller et al. (2018), for a multi-encoder approach.

Close to the approach we propose in this work, Gete et al. (2023) include a model variant where target data is concatenated to the source sentence, which was shown to be particularly beneficial to address target-level phenomena in Basque-Spanish translation. However, their experiments were limited to one target sentence, i.e. without prepending context on the target side. We show in this work that including the target context in both source and target languages is critical to achieve significant improvements overall.

Since standard NMT evaluation metrics such as BLEU (Papineni et al., 2002) are not well equipped to assess discourse phenomena, several challenge test sets have been developed specifically to measure translations in context, via contrastive evaluations (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019b; Lopes et al., 2020; Nagata and Morishita, 2020; Gete et al., 2022). We include contrastive test sets that cover target-language phenomena such as deixis or lexical cohesion, as well as phenomena where the relevant context informa-

SRC: Так эти фотографии снял ты ? Да . Так что произошло ночью ? [BR] she was there posing , looking at you ?
TGT: Так эти фотографии снял ты ? Да . Так что произошло ночью ? [BR] Она позировала глядя на тебя ?
(And you took those photos, right? Yeah. So what about last night? [BR] She was there posing, looking at you?)

Table 1: Example of input of the tgt-$n$to$n$ model, extracted from the corpus prepared in Voita et al. (2019b).

tion is available in both source and target languages.

3 Exploiting Target Language Data

Our approach operates within a standard NMT architecture. At training time, we simply discard source data from the equation and prepend context sentences in the target language to both the source sentence to be translated and the target reference sentence. In both cases, we add a special token to separate the context, as shown in Table 1. At inference time, the previously translated sentences would be prepended as source context. Due to the nature of the challenge sets, our contrastive results will be based on reference context translations.[1]

The main incentive for choosing target language data instead of source data is the nature of the contextual phenomena of interest for machine translation, as these can be grouped into two broad categories depending on the location of the relevant contextual information.

In a first category would be discourse-level phenomena that require context information in the target language side, typically related to discursive cohesion in a broad sense (see examples *a* and *b* in Table 2). For instance, to maintain lexical cohesion beyond the sentence level, a quality translation should feature lexical repetition when necessary, as it can mark emphasis or support question clarification. Another case is that of names with several possible translations, where translations must remain consistent throughout. Degrees of politeness and linguistic register in general also involve translation alternatives that are linguistically correct in isolation, but require consistency at the document level. In the case of pronouns, when the source antecedent has translation options in different grammatical genders, translation choices should be coherent throughout in the target language. In all of these cases, the relevant information involves previous translations into the target language.

In a second major category are phenomena for which either the source or the target context provides relevant information (examples *c* and *d* in Table 2). This includes word sense disambiguation scenarios, where different types of source or target elements may be relevant to perform disambiguation to some extent, in combination or in isolation. Gender selection would also fall into this category, in those cases where translation options for the relevant contextual antecedent are unique or share the same gender. The resolution of elliptical constructions in the source language, with no equivalent in the target language, may also require context information from the source or the target language. Another instance for this type of phenomena would be the translation of Japanese zero pronouns into English (Nagata and Morishita, 2020), where information on both sides can become relevant to determine the grammatical features of the target pronoun.

Note that, even in those cases where contextual information is present in both the source and target languages, using source information for disambiguation can result in a lack of consistency in the target language, whenever incorrect translations are involved. Bawden et al. (2018) provide a contrastive test for these cases, where part of the source has been translated incorrectly but the translation is still required to be consistent overall.

4 Experimental Setup

4.1 Data

We describe in turn below the datasets used to train and test our NMT models. All selected datasets were normalised, tokenised and truecased using Moses (Koehn et al., 2007) and segmented with BPE (Sennrich et al., 2016), using 32,000 operations. Tables 3 and 4 show corpora statistics for parallel and contrastive datasets respectively.

For Basque–Spanish, we selected the TANDO corpus (Gete et al., 2022), which contains parallel

[1] See Section 7 for a discussion and results with machine-translated context in terms of reference metrics.

(a) Lexical cohesion: name translation
EN: Not for Julia. Julia has a taste for taunting her victims. RU: Не для **Джулии**[*Julia*]. **Юлия***[*Julia*] умеет дразнить своих жертв.
(b) Deixis: register coherence
EU: Ez dago martetarrik zuen artean. Guztiak ari zarete ereduak lotu eta... ES: Ninguno de **ustedes**[form] es marciano. Todos **vosotros estáis***[inf] siguiendo un modelo y... *(None of you are Martians. You are all following a model and...)*
(c) Gender selection
EU: Hori nire **arreba** da. **Berak**[?] zaindu zituen nire argazkiak. *(That's my **sister**. **He/She** took care of my photos.)* ES: Esa es mi **hermana**. **Él*** cuido mis fotos. *(That's my **sister**. **He*** took care of my photos.)*
(d) Verb phrase ellipsis
EN: Veronica, thank you, but you **saw** what happened. We all **did**[?]. RU: Вероника, спасибо, но ты **видела**, что произошло. Мы все **хотели***. *(Veronica, thank you, but you **saw** what happened. We all **wanted*** it.)*

Table 2: Examples of inconsistencies extracted from (Voita et al., 2019b) and (Gete et al., 2022).

	EU-ES	EN-RU
Train	1,753,726	6,000,000
Dev	3,051	10,000
Test	6,078	10,000

Table 3: Parallel corpora statistics (number of sentences)

EU-ES	Size	src	tgt	Dist.
GDR-SRC+TGT	300	✓	✓	≤ 5
COH-TGT	300		✓	≤ 5
EN-RU	**Size**	**src**	**tgt**	**Dist.**
Ellipsis (infl.)	500	✓	✓	≤ 3
Ellipsis (VP)	500	✓	✓	≤ 3
Deixis	2,500		✓	≤ 3
Lex. cohesion	1,500		✓	≤ 3

Table 4: Contrastive test sets: size (number of instances), required context information and distance to the disambiguating information (number of sentences)

data from subtitles, news and literary documents. It includes two contrastive datasets for Basque to Spanish translation. The first one, GDR-SRC+TGT, centres on gender selection, with the disambiguating information present in both the source and target languages. The second one, COH-TGT, is meant to evaluate cases where, despite the absence in the source language of the necessary information to make a correct selection of gender or register, the translation must be contextually coherent using target-side information.

For English–Russian, we used the dataset described in Voita et al. (2019b), based on Open Subtitles excerpts (Lison et al., 2018). It includes 4 large-scale contrastive test sets for English to Russian translation. Two of these tests are related to ellipsis and contain the disambiguating information in both the source and target-side context: Ellipsis (infl.) assesses the selection of correct morphological noun phrase forms in cases where the source verb is elided, whereas Ellipsis (VP) evaluates the ability to predict the verb in Russian from an English sentence in which the verb phrase is elided. In the other two tests, the disambiguating information is only present in the target-side context: Deixis addresses politeness consistency in the target language, without nominal markers, whereas Lexical Cohesion focuses on the consistent translation of named entities in Russian.

4.2 Models

All models in our experiments are based on the Transformer-base architecture (Vaswani et al., 2017), trained with Marian (Junczys-Dowmunt et al., 2018).

As a general baseline, we trained a sentence-level model using all source-target sentence pairs in the selected training datasets for each language pair. We then trained the following context-aware models, varying the type of context sentences prepended to the source and/or the target sentence, and adding a special token to separate the context:

- $n$to1: n-1 source context sentences concatenated to the source sentence, and a single reference target sentence.
- $n$to$n$: n-1 source context sentences concatenated to the source sentence and n-1 target context sentences to the target sentence.
- tgt-$n$to1: n-1 context sentences from the target language concatenated to the source sentence, and a single reference target sentence.
- tgt-$n$to$n$: n-1 context sentences from the target language concatenated to both the source and target sentences.

Given the size of the context for each language pair, we thus have n=6 for Basque–Spanish models and n=4 for English–Russian models. All context-aware models were initialised with the weights of the sentence-level baseline.

5 Results

5.1 Parallel Tests

We first compared models in terms of BLEU on the parallel test sets, using SacreBLEU (Post, 2018)[2]. Statistical significance was computed via paired bootstrap resampling (Koehn, 2004), for $p < 0.05$.[3]

The results are shown in Table 5. In Basque–Spanish, the $n$to$n$ and tgt-$n$to$n$ models performed better than the alternatives, with no statistically significant differences between the two. Both were significantly better than the baseline and the models which used only a single reference in the target language. In English–Russian, the tgt-$n$to$n$ model outperformed all other models, including the standard $n$to$n$ model, although with only a 1.09 BLEU point gain over the latter.

Using only target context data was thus not detrimental in terms of reference metrics on the large parallel test sets used in the experiments, and was even optimal in one language pair. This is at least indicative of an absence of unwarranted side-effects in terms of translation quality.

	EU-ES	EN-RU
Sentence-level	31.20	31.09
$n$to1	29.91	31.48
tgt-$n$to1	29.43	31.03
$n$to$n$	**31.96**	31.20
tgt-$n$to$n$	**31.82**	**32.29**†

Table 5: BLEU results on the parallel test sets.

5.2 Challenge Tests

We evaluated the models in the challenge test sets, both in terms of BLEU and in terms of accuracy of the contrastive evaluation. Statistical significance of accuracy results was computed using McNemar's test (Mcnemar, 1947), for $p < 0.05$. The results are shown in Tables 6 and 7.

Considering both language pairs, the first notable results are the significant improvements obtained with the tgt-$n$to$n$ models on the target-oriented test sets. In terms of accuracy, in EU-ES on the COH-TGT test, this model outperformed the baseline by 27.67 points and the $n$to$n$ model by 16.34 points. In EN-RU, the gains were of 37.44 and 5 points in Deixis against the baseline and $n$to$n$ model, respectively; on the Lexical Cohesion test set, the gains were 3.6 and and 3.54 points, respectively. On these target-oriented test-sets, the tgt-$n$to$n$ model also achieved gains in terms of BLEU scores: +3.72 points in EU-ES, +7.02 in EN-RU on Deixis, and +3.09 in EN-RU on the Lexical cohesion test.

Turning now to the test sets where relevant context information is available in both the source and target languages, the results are more balanced between methods and even apparently large score differences are not always statistically significant, as all these tests are significantly smaller. In EU-ES, there is thus no statistical significance between the two best methods, $n$to$n$ and tgt-$n$to$n$, in terms of accuracy or BLEU. The same was true for the Ellipsis VP results in EN-RU between these two models, with similar BLEU and accuracy scores. On Ellipsis infl., tgt-$n$to$n$ was significantly better than $n$to$n$ in terms of BLEU, with a gain of +3.72 points, whereas the reverse was true on accuracy, with a difference of 5.20 points.

Regarding the other two contextual variants, $n$to1

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

[3] In all tables, best scores given the statistical test at hand are shown in bold; statistically significant results between $n$to$n$ and tgt-$n$to$n$ results are indicated with †.

	GDR-SRC+TGT		COH-TGT	
	BLEU	ACC.	BLEU	ACC.
Sentence-level	36.28	53.67	35.04	54.00
$n$to1	36.82	66.33	33.23	53.00
tgt-$n$to1	36.79	66.33	37.31	74.00
$n$to$n$	**40.45**	**77.67**	35.89	65.33
tgt-$n$to$n$	**39.05**	**72.67**	**39.61**†	**81.67**†

Table 6: BLEU and accuracy results on the Basque–Spanish challenge tests.

	Ellipsis infl.		Ellipsis VP		Deixis		Lex. Cohesion	
	BLEU	ACC.	BLEU	ACC.	BLEU	ACC.	BLEU	ACC.
Sentence-level	30.81	51.80	22.20	27.80	28.10	50.04	**31.52**	45.87
$n$to1	32.69	54.60	**30.24**	**65.40**	28.20	50.04	29.47	45.87
tgt-$n$to1	32.28	53.60	23.59	29.00	28.30	50.56	30.37	45.87
$n$to$n$	36.97	**75.20**†	**29.59**	**62.60**	27.15	82.48	27.89	45.93
tgt-$n$to$n$	**40.69**†	70.00	**30.75**	60.00	**34.17**†	**87.48**†	**30.98**†	**49.47**†

Table 7: BLEU and accuracy results in English–Russian challenge tests.

and tgt-$n$to1, which used no context information in the target side of the input, the results in accuracy were similar overall, performing on a par with the baseline on Lexical Cohesion, Deixis and COH-TGT for $n$to1. This was was expected for the $n$to1 models, as the relevant information is in the target language in these cases, which these models have no access to. For the tgt-$n$to$n$1 model, the gains achieved over the $n$to1 model on COH-TGT in both BLEU and accuracy (also outperforming the $n$to$n$ model) were not unexpected, as the target context information is exploitable by this model, although on the encoder side rather than the decoder side.

Similar gains could have been expected on Lexical Cohesion and Deixis with the tgt-$n$to1 model, but it performed on a par with the baseline and $n$to1 model on these test sets. In terms of lexical cohesion, this might be due to the fact that named entities are usually translated into a single default variant in the training data, a strong tendency reflected by the model[4]. In the COH-TGT test however, register and gender options are all equally valid and more equally distributed (modulo typical bias), which might give more relative weight to contextual information. The same should hold true for the EN-RU Deixis test set, however, it was not the case here. In this case, we hypothesise that this could result from a similar unbalanced register distribution training and contrastive sets, both extracted from OpenSubtitles, again strongly biasing the model towards default translations irrespective of context. Similarly unexpected was the performance of the $n$to1 model on Ellipsis VP, on a par with or outperforming the $n$to$n$ variants. We left further exploration of both $n$to1 models aside, as they were outperformed by the $n$to$n$ variants overall.

From these results, the tgt-$n$to$n$ model proved optimal overall in terms of accuracy and BLEU on the contrastive test sets, matching the strong $n$to$n$ variant where relevant context information is available in both source and target languages, and providing large improvements over all alternatives in the other cases.

6 Using Back-translated Data

When document-level parallel data are lacking, monolingual data in the target language can be exploited within concatenation-based approaches via back-translation (Junczys-Dowmunt, 2019; Sugiyama and Yoshinaga, 2019; Huo et al., 2020). Some level of degradation is expected, depending on the quality of the model used to back-translate the target data, and we also expect the models to be impacted differently: for both the $n$to$n$ and tgt-$n$to$n$ models, the target sentence and its back-translation would be identical, as would be the original target context sentences, but the $n$to$n$ model will also

[4] Some illustrative examples are discussed in Appendix A

require back-translated target context sentences, unlike the tgt-$n$to$n$ model.

For comparison purposes we back-translated the target side of the training data for both language pairs, and trained the two main model variants strictly on the back-translated data. The results are shown in Table 8, contrasting the use of parallel (PA) and back-translated (BT) data. The overall degradation using BT data was more salient in EU-ES than in EN-RU, which is likely due to the differences in training data size and the resulting quality of the respective models. In both cases, the tgt-$n$to$n$ model proved more robust with around 1 and 2 BLEU point gains over the $n$to$n$ model. This is also likely due to the latter being affected by back-translation quality of the translated context, whereas the former only requires the back-translation of the non-context target sentence.[5]

	EU-ES	EN-RU
$n$to$n$ (PA)	**31.96**	31.20
tgt-nton (PA)	**31.82**	**32.29**
$n$to$n$ (BT)	25.46	29.21
tgt-$n$to$n$ (BT)	27.33†	30.10†

Table 8: BLEU results on the parallel test sets using parallel (PA) and back-translated (BT) data.

Accuracy and BLEU results on the contrastive test sets are shown in Table 9 and Table 10 for Basque–Spanish and English–Russian, respectively. In EU-ES on the COH-TGT test, there is marked degradation in terms of BLEU for both models when using BT data, although the tgt-$n$to$n$ model was still closer to the best model; in terms of accuracy, both models maintained parity or achieve slight gains using BT data, with the tgt-$n$to$n$ model still largely outperforming the $n$to$n$ baseline. On the GDR-SRT+TGT test, there were almost no changes in terms of BLEU. In terms of accuracy, only a slight degradation was observed for the $n$to$n$ model using BT data, and slight gains for the tgt-$n$to$n$ model.

In English–Russian, BLEU degradation was observed for the $n$to$n$ model on all but the lexical cohesion test, and for the tgt-$n$to$n$ model on all but the Ellipsis VP test, with only minor losses overall and the largest losses for both models, at around 2 BLEU points, on Deixis. In terms of accuracy, both models achieved gains on Ellipsis infl. and Deixis, and minor losses or parity otherwise.

Overall, the tendencies observed using parallel data are translated to the use of back-translated data, with the tgt-$n$to$n$ model being the top-performing variant overall. Larger test sets would be warranted to assess the performance of these models using BT data, as some gains are somewhat surprising, e.g. those of the $n$to$n$ model on Ellipsis infl. using BT data, which are likely to include errors due to the nature of back-translation. BLEU results in particular are more likely to be representative of underlying tendencies on the parallel test sets, as shown by the losses described in Table 8. Nonetheless, the results on the available datasets in terms of accuracy seem to indicate that the use of BT data is viable, and particularly exploitable by the tgt-$n$to$n$ model considering the large gains obtained on target-language phenomena, and the parity achieved on the other discourse-level phenomena.

7 Machine-translated Target Context

Following standard practice, for all results reported so far, we used the reference target context instead of the machine-translated output. This is meant to remove potential noise in terms of context translation errors and evaluate the approaches on their ability to translate with a correct context. Using reference translations also allows for an evaluation of phenomena where more than one translation in the context would be correct – e.g. *box* translated as *boîte* (fem.) instead of *carton* (masc.) in French – but the contrastive evaluation relies on one of these translations being selected as the correct one and further phenomena, such as coherence, are measured accordingly. A correct but different context translation would be unfairly penalised in these cases.

Nonetheless, in practice, at inference time there are no reference translations, of course. Whereas the $n$to$n$ model should not be impacted at all, since it only translates the source sentences and any translated target material before the generated separator is discarded, the tgt-$n$to$n$ models are more susceptible to suffer from errors in the translation of the context sentences. To measure this aspect, we computed BLEU scores using the machine-translated target sentences with the tgt-$n$to$n$ model. The results are shown in Table 11 on the larger parallel test sets.

From these results, using MT output does not

[5] On practical grounds, the tgt-$n$to$n$ model is also less resource consuming, as the context sentences do not need to be back-translated.

	GDR-SRC+TGT		COH-TGT	
	BLEU	ACC.	BLEU	ACC.
$n$to$n$ (PA)	**40.45**	**77.67**	35.89	65.33
tgt-$n$to$n$ (PA)	**39.05**	**72.67**	**39.61**	**81.67**
$n$to$n$ (BT)	**41.58**	**76.00**	31.02	67.00
tgt-$n$to$n$ (BT)	**40.22**	**74.00**	34.62†	**81.33**†

Table 9: BLEU and accuracy results on Basque–Spanish contrastive tests with parallel (PA) and back-translated (BT) data.

	Ellipsis infl.		Ellipsis VP		Deixis		Lex. cohesion	
	BLEU	ACC.	BLEU	ACC.	BLEU	ACC.	BLEU	ACC.
$n$to$n$ (PA)	36.97	75.20	29.59	62.60	27.15	82.48	27.89	45.93
tgt-$n$to$n$ (PA)	**40.69**	70.00	**30.75**	60.00	**34.17**	**87.48**	**30.98**	**49.47**
$n$to$n$ (BT)	35.63	**78.60**†	28.84	**69.40**†	25.66	83.92	28.29	46.20
tgt-$n$to$n$ (BT)	39.25†	73.60	**31.86**†	57.60	31.84†	**87.84**†	29.81†	**49.20**†

Table 10: BLEU and accuracy results on English–Russian contrastive tests with parallel (PA) and back-translated (BT) data.

	EU-ES	EN-RU
$n$to$n$ (RF)	**31.96**	31.20
tgt-$n$to$n$ (RF)	**31.82**	**32.29**
tgt-$n$to$n$ (MT)	31.08	31.52

Table 11: BLEU results on the parallel test sets using reference (RF) and machine-translated (MT) context.

seem to markedly impact translation quality, at least in terms of BLEU scores. As previously noted, measuring its impact on contrastive accuracy would require challenge sets that take into account different correct choices in the translation of context sentences, a task which we left for future work considering the effort required in designing and preparing this type of dataset. Additionally, a proper assessment of the impact of machine-translated context on the tgt-$n$to$n$ model would need to take into account the quality of the translation model, with larger models expected to minimise context translation errors for this approach.

8 Conclusions

We proposed a novel variant for context-aware NMT, where target-language context is prepended to both source and target sentences. Our results, in terms of BLEU and contrastive accuracy, showed that this approach significantly outperformed state-of-the-art models for target-language phenomena, while achieving parity overall for discourse-level phenomena where the relevant contextual information is in both the source and target languages.

We further evaluated the use of back-translated data, showing that the tendencies observed on parallel data were maintained. We also measured the impact of using machine-translated output instead of reference translations, which could have impacted the proposed approach but were shown to have only a marginal effect, on the parallel test sets at least. In addition, the use of more robust baseline models, trained on larger volumes of data, should mitigate these effects. New challenge datasets might be needed to support a more precise evaluation of these aspects, as current challenge datasets can be dependent on arbitrary context translation decisions depending on the phenomena at hand.

Overall, the proposed approach requires no changes to the standard NMT architecture, supports simplified back-translation where the context need not be back-translated, and provides either significant gains or parity against strong baselines. In future work, we will further explore this approach in different languages and domains, notably testing its limits by seeking specific context-level translation phenomena, for which source context data might actually be of higher relevance, if any, beyond current evaluation suites.

Limitations

The evaluation of possible losses when using machine-translated output was limited to BLEU scores on parallel test sets, as contrastive test sets could not be used in this case due to the necessary arbitrary selection of context translations among various equally correct options. A correct machine-translation choice could thus result in artificially erroneous answers on some contrastive tests. This limited our evaluation of the impact of machine-translated output, which could in theory impact our proposed approach where target translations are used in the source, whereas a standard concatenation-based approach would not be affected. Designing and constructing datasets that support a fair evaluation on these grounds was beyond the scope of this work.

We also used BLEU as our sole reference metric, although its limitations are fairly well known and other metrics such as COMET (Rei et al., 2020) might provide results that better correlate with human judgements in some cases. We did not report reference metrics results beyond BLEU for presentation convenience, as those results correlated strongly with COMET results in our experiments. Additionally, reference-based metrics are not sufficiently precise for document-level translation in general, and should be mainly valued as complementary to the results in terms of contrastive accuracy which we provide in our work.

Ethics Statement

Context-aware machine translation models may help reduce some of the biases of sentence-level models, by more adequately translating cases where a context-agnostic translation would be biased due to training data distribution, in terms of gender for instance. However, this work does not address nor measure the impact of the proposed approach on translation bias specifically.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

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

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, 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.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Harritxu Gete, Thierry Etchegoyhen, and Gorka Labaka. 2023. What works when in context-aware neural machine translation? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 147–156, Tampere, Finland. European Association for Machine Translation.

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. 2022. TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Zhihong Huang, Longyue Wang, Siyou Liu, and Derek F. Wong. 2023. How does pretraining improve discourse-aware translation?

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney.

2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 54–57, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

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.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. 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.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid).

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

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Suvodeep Majumder, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: Pushing the limits of multi-segment models for context-aware translation.

Elman Mansimov, Gábor Melis, and Lei Yu. 2021. Capturing document context inside sentence-level neural machine translation models with self-training. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 143–153, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Quinn Mcnemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

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.

Masaaki Nagata and Makoto Morishita. 2020. A test set for discourse translation from Japanese to English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3704–3709, Marseille, France. European Language Resources Association.

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. *arXiv preprint arXiv:2304.12959v1*.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

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

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

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

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models.

Xueqing Wu, Yingce Xia, Jinhua Zhu, Lijun Wu, Shufang Xie, and Tao Qin. 2022. A study of BERT for context-aware neural machine translation. *Mach. Learn.*, 111(3):917–935.

Mingzhou Xu, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021. Document graph for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8435–8448. Association for Computational Linguistics.

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.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Toward making the most of context in neural machine translation. *CoRR*, abs/2002.07982.

A Lexical Cohesion Results

The results in the English-Russian language pair for the Lexical cohesion challenge test are quite remarkable, as none of the models achieves an accuracy score of 50%. These results could be attributed to the fact that, although the test presents two suitable options for translating proper nouns, the training data for this language pair consists exclusively of OpenSubtitles data, where these names tend to be systematically translated in a certain way, leading the models to exhibit a strong bias towards this translation option. While verifying this theory can be challenging, as it would require identifying all the names and their corresponding translations in the training data, we manually examined some of the names featured in the test.

One example is the name Spence, which has two possible valid translations in the test, Спенсер and Спенс. These are two possible translations of Spencer in the training data, however: whereas Спенс is almost always related to Spence, Спенсер is only a translation of Spence in 3% of the cases; in the other 97% of the cases, this word is a translation of Spencer. It would thus be logical to think that the concatenation-based models relate Spence to Спенс and Spencer to Спенсер, a translation bias which might be difficult to mitigate even when adding contextual information. Table 12 shows a few other examples of cases that might be challenging to handle along these lines.

While our results align with other publications (Zheng et al., 2020; Lupo et al., 2022; Sun et al., 2022), alternative approaches such as Voita et al. (2019a), have achieved better results on this task. Their approach relies on a monolingual repair model which does not rely on source information, thus obviating the obsevred training data bias altogether. Alternatively, models like CADec (Voita et al., 2019b) intentionally introduce artificial errors in their data, potentially making them less conservative and more prone to corrections, while $n$to$n$ models are more influenced by the default translation of the source sentence. Moreover, when these artificial errors are not introduced, the CADec accuracy in this test also falls below 50%, supporting the hypothesis that training data bias is a relevant factor for the observed results on the lexical cohesion test set.

Source	Posible translations
Spence	Спенсер (Spence 3%, Spencer 97%), Спенс (Spence 99%, Spencer 1%)
Darius	Дария (Darius 46%, Daria 54%), Дариуса (Darius 100%)
Sidney	Сидней (Sidney 50%, Sydney 50%), Сидни (Sidney 25%, Sydney 75%)
Hillary	Хиллари (Hillary 92%, Hilary 8%), Хилари (Hillary73%, Hilary 27%)
Fausto	Фауст (Fausto 62%, Faust 38%), Фаусто (Fausto 100%)

Table 12: Examples of names and their plausible translations selected from the challenge test and their relationship in the training data.