Promoting Target Data in Context-aware Neural Machine Translation

Harritxu Gete[1,2*] **Thierry Etchegoyhen**[1*]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]University of the Basque Country UPV/EHU
{hgete,tetchegoyhen}@vicomtech.org

Abstract

Standard context-aware neural machine translation (NMT) typically relies on parallel document-level data, exploiting both source and target contexts. Concatenation-based approaches in particular, still a strong baseline for document-level NMT, prepend source and/or target context sentences to the sentences to be translated, with model variants that exploit equal amounts of source and target data on each side achieving state-of-the-art results. In this work, we investigate whether target data should be further promoted within standard concatenation-based approaches, as most document-level phenomena rely on information that is present on the target language side. We evaluate novel concatenation-based variants where the target context is prepended to the source language, either in isolation or in combination with the source context. Experimental results in English-Russian and Basque-Spanish show that including target context in the source leads to large improvements on target language phenomena. On source-dependent phenomena, using only target language context in the source achieves parity with state-of-the-art concatenation approaches, or slightly underperforms, whereas combining source and target context on the source side leads to significant gains across the board.

*These authors contributed equally to this work.

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 on 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). Among the various approaches to context-aware NMT, 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; Sun et al., 2022; Post and Junczys-Dowmunt, 2023).

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 as the core information to model context-awareness. However, most 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, in this work we aim to explore the impact of promoting target language data in standard context-aware NMT.

Along these lines, we explore a simple

concatenation-based approach which consists in simply prepending context sentences from the target language to the source sentence to be translated, in isolation or in combination with source context. The underlying intuition is that contextual phenomena would be mainly modelled at the decoder level via target-side context information, whereas, on the encoder side, context data will be either ignored and copied, as foreign data, or also associated with source information to further model context. Using target language context data on the source side also enables the use of a standard NMT architecture and concatenation-based approach to context-aware NMT.

We show that replacing source context sentences with the target context already leads to significant gains for discourse-level phenomena that depend on target-language information, while achieving either parity or moderate degradation in contrastive accuracy on other phenomena. Combining both source and target context sentences on the source side leads to consistent significant improvements across the board. 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 the source and/or target language context.

In addition to accuracy results on specific phenomena, we compare the overall translation quality on parallel test sets as well. We also measure the impact of using either reference or machine-translated output as context at inference time, with only minor loss observed with the latter in our experiments. Finally, we evaluate the use of back-translated data, with similar comparative gains as those obtained using parallel document-level data. Overall, our experimental results indicate that promoting target context data within a standard NMT architecture can be a promising alternative for context-aware machine translation.

2 Related Work

One of the first methods proposed for document-level NMT 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; Gete and Etchegoyhen, 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), to initialize NMT models (Huang et al., 2023) or fusing the language model with a sentence-level translation model (Petrick et al., 2023). Directly using pretrained language models to perform translation can achieve competitive results, although these models might still produce critical errors 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. Fernandes et al. (2021) conclude that concatenation-based models make more use of the target context than the source context, but Jin et al. (2023) show that the effectiveness of the target context versus the source context is highly dependent on the language pair involved. Close to the target-based approach we explore in this work, Scherrer et al. (2019) and Gete et al. (2023) include variants where target data is concatenated to the source sentence, notably showing that the target context is equally as important than source context, and particularly beneficial to address target-level phenomena. 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 con-

(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 1: Examples of document-level inconsistencies extracted from (Voita et al., 2019b) and (Gete et al., 2022).

text 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 accuracy on 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; Currey 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 information is available in both source and target languages.

3 Exploiting Target Language Context

The main incentive for the promotion of target context data is the nature of the contextual phenomena of interest for machine translation, as these can be grouped into four broad categories depending on the location of the relevant contextual information.

In a first category would be discourse-level phenomena that require context information on the target language side, typically related to discursive cohesion in a broad sense (see examples *a* and *b* in Table 1). 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 equally 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 in the target language.

In a second major category are phenomena for which the relevant context information is in both the source and the target context (examples *c* and *d* in Table 1). This includes word sense disambiguation scenarios, where different types of source or target elements may be relevant to perform disambiguation. 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,

(a)	ES: Hablé con mi **amiga**[fem]. Dijo que sí. EN: I talked to my **friend**[?]. **She/He*** said yes.
(b)	EN: **You** can't leave me! Don't go away! ES: ¡No **puede** dejarme! ¡No **se vaya/te vayas***!

Table 2: Example of ambiguity where source context is necessary for disambiguation, in isolation (a) or in combination with the target context (b).

2020), where information on both sides can be relevant to determine the grammatical features of the target pronoun. Note that, even when 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.[1]

A third class of context-dependent phenomena exists, where source data are the only source of disambiguating information. This involves cases where the context includes the translation of a word marked for a specific category (e.g., gender) into a unmarked one, while the source sentence to be translated involves insufficient source information (e.g., a dropped pronoun) that needs to be translated into a marked element (e.g., a pronoun marked for gender). A typical example is provided in Table 2 *a*. In such a case, there would be insufficient information in the target language, as the proper translation of the dropped subject pronoun into *she* could only be determined from the gender of the source context antecedent *amiga* (*friend)*.

Finally, a fourth broad category contains constructions where the source and target context need to be processed in combination for a correct translation. In the example *b* in Table 2, the source context subject *you* does not provide information about register, and neither does the target context in Spanish, since the verb *puede* can indicate either third person in informal register or second person in polite register. However, the source context indicates second person. Therefore combining both sources of context information, it can be derived that the translation should be second person in polite form.

Any target-only approach, such as monolingual repair (Voita et al., 2019a) or the target-only variant we also explore in this work, would only generate the correct translation in the latter two classes of cases by either chance or training bias. Although these cases exist, it is unclear how widespread they actually are, compared to the other two main classes of contextual phenomena described above. In what follows, we set to compare the relative importance of source and target data across the main phenomena as represented in the selected document-level test suites.

4 Promoting Target Language Data

To explore the promotion of target language data, we simply prepend the target context sentences to the source sentence to be translated, either discarding or maintaining the source context sentences. On the target side, we evaluate the use of empty context as well as maintaining the target context sentences. We add a special token to separate the concatenated context sentences in all cases.

At inference time, in practice the previously translated sentences would be prepended as context. Since context translations can feature various degrees of correctness, we assess the approach under both ideal and average conditions. On parallel test sets, we measure the use of both correct reference context sentences (Section 6.1) and machine-translated ones (Section 8). On the contrastive test sets, only reference translations are used, as is standard practice, since target context coherence requirements prevent the use of non-reference context translations for fair evaluations (see the discussion in Section 8).

The prepended target-language data will need to be processed by the source language encoder under this approach, which might generate unwarranted noise. We hypothesise however that the encoder will essentially treat foreign language subwords as tokens to be copied directly into the target language, a typically simple operation for standard NMT models. We use BPE models jointly learned on merged source and target language data to facilitate this part of the process. Overall, the proposed approach provides the means to exploit target language data on the decoder side, without

[1] 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.

any change to model architecture, while introducing data that might be easily processed via copying on the source side.

5 Experimental Setup

5.1 Data

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

	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)

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

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)

5.2 Models

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

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 different variants of concatenation-based 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.

We use the following convention to denote the models: $n$to$n$ uses the same amount of source and target data on each side, and represents the state-of-the-art baseline; tgt-$n$to$n$ uses target language data on both sides, discarding source context altogether; $n$to1 and tgt-$n$to1 are variants of the previous models that use no context sentences in the target language; finally, src+tgt-$n$to$n$ and tgt+src-$n$to$n$ are variants where target context sentences are combined with source context sentences, by prepending them after or before the latter, respectively. For convenience, we will refer to the tgt-$n$to$n$, src+tgt-$n$to$n$ and tgt+src-$n$to$n$ variants as X-tgt-$n$to$n$, as they share the use of target context on both sides. In Appendix A, we provide a diagram to illustrate data composition for each model.

Given the size of the context for each dataset, we 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.

Note that we discarded 1to$n$ models, as they present two main challenges. Within a standard concatenation approach, we would be tasking the model to learn a transformation from a single source sentence to both the context and the target sentence, although the target context cannot be derived from the source sentence, obviously. Alternatively, a 1to$n$ model could be designed via changes in the NMT architecture, with forced decoding over the specified target context at both training and inference time. The required architectural changes were beyond the scope of this work, although this type of model might be worth exploring in more details.

6 Results

6.1 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$n$, tgt-$n$to$n$, and src+tgt-$n$to$n$ models performed better than the alternatives, with no statistically significant differences between the three, with the tgt+src-$n$to$n$ achieving slightly lower results. All three were notably significantly better than the baseline and the models which used only a single reference in the target language. In English–Russian, all X-tgt-$n$to$n$ model variants, that included target context data on the source side, outperformed all other models, including the standard $n$to$n$ model.

	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**
src+tgt-$n$to$n$	**31.94**	**32.32**
tgt+src-$n$to$n$	31.56	**32.49**

Table 5: BLEU results on the parallel test sets.

Sentence-level metrics are typically insufficient to assess translation quality at the document level (Wong and Kit, 2012), and conclusions should not be drawn from the above results regarding context-aware ability of the different models. They do however indicate several tendencies at the sentence level. First, the proposed use of target context data on both sides was not detrimental in terms of translation quality, as the X-tgt-$n$to$n$ models performed on a par with, or better than, the other variants. Secondly, the lower results obtained by the $n$to1 and tgt-$n$to1 models seem to indicate that (i) removing target context data on the decoder side can be detrimental, as in EU-ES, and (ii) using source or target language data on the encoder side can lead to similar BLEU results, as was the case in both language pairs.

Note that the results above were obtained with reference translations, in an ideal scenario where the context is correctly translated. In Section 8, we present additional results using machine-translated context, to measure the impact of eventual errors in target context translation.

6.2 Challenge Tests

We evaluated the different models on 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 gains achieved by the src+tgt-$n$to$n$ and tgt+src-$n$to$n$ models, which outperformed all other variants overall, in terms of both BLEU scores and contrastive accuracy. The tgt-$n$to$n$ model, where source context was discarded altogether, also outperformed the baselines in terms of BLEU in all but one case, and either matched the other two target-based variants in half of the scenarios, or was outperformed by these variants in the other three cases. In terms of contrastive accuracy, it also outperformed the baselines by a wide margin on target-oriented phenomena while achieving parity or resulting in accuracy loss on other phenomena. Overall, the best performing and most consistent variant across datasets and metrics was the src+tgt-$n$to$n$ variant.

On all target-related phenomena, the X-tgt-$n$to$n$ models outperformed all alternatives, and in particular the standard $n$to$n$ variant by large margins. In terms of accuracy, in EU-ES on the COH-TGT test, the tgt-$n$to$n$ model already outperformed the baseline by 27.67 points and the $n$to$n$ model by

[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.

	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**
src+tgt-$n$to$n$	**41.29**	**78.67**	**40.23**	**84.67**
tgt+src-$n$to$n$	**42.35**	**78.67**	**39.86**	**82.67**

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
src+tgt-$n$to$n$	**40.98**	**77.20**	**35.84**	**77.60**	**34.38**	**87.48**	**31.75**	**53.07**
tgt+src-$n$to$n$	**42.02**	**75.60**	**34.46**	**74.88**	**34.07**	**88.28**	**31.33**	51.00

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

16.34 points, with even higher accuracy gains for the best-performing src+tgt-$n$to$n$ model (+19.34). In EN-RU, on *Deixis* gains of up to 38.24 and 5.8 points were achieved against the baseline and $n$to$n$ model, respectively; on the *Lexical Cohesion* test set, the gains reached 7.2 and and 7.14 points, respectively. On these target-oriented test-sets, all X-tgt-$n$to$n$ model also achieved comparable gains in terms of BLEU scores, with a maximum against the $n$to$n$ model of +4.34 points in EU-ES, +7.23 in EN-RU on *Deixis*, and +3.86 in EN-RU on the *Lexical cohesion* test.

Turning now to the test sets where relevant context information is available in either both the source and target languages, or perhaps only in the source language in some cases, the results are more balanced between the $n$to$n$ baseline and the X-tgt-$n$to$n$ variants, although the src+tgt-$n$to$n$ achieved the best results overall in terms of both BLEU and accuracy. On *Ellipsis VP*, the latter notably achieved gains of 15 accuracy points, with the tgt+src-$n$to$n$ variant a close second at +12.28. On *Ellipsis infl.* and GDR-SRC-TGT, the gains were more limited, with a maximum of +1 and and +2 accuracy points for the src+tgt-$n$to$n$ model against the $n$to$n$ baseline, respectively, although significant BLEU gains of up to +3.3 and +5.05 were observed on these test sets, respectively.

Unsurprisingly, on these three datasets where source information is a relevant factor, in combination or in isolation, the tgt-$n$to$n$ model underperformed, though in accuracy only and to a limited extent on *Ellipsis VP*, for instance. This variant also significantly outperformed the $n$to$n$ baseline in terms of BLEU on *Ellipsis infl.*, with a 3.60 points gain. To further determine the impact of source and target context and more precisely assess the limits of this type of model, more fine-grained challenge tests would be needed to distinguish between cases that can solely be resolved with source context information and those where either side of context provides sufficient information.

Regarding the other two contextual variants, $n$to1 and tgt-$n$to1, which used no context information on the target side of the input, the results in accuracy were similar overall, performing on a par with the sentence-level baseline on *Lexical Cohesion*, *Deixis* and COH-TGT for $n$to1. This 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.

Overall, promoting target data in a

concatenation-based approach achieved large improvements across the board over the sentence-level and $n$to$n$ baselines. Replacing source context data altogether with the target context already improved significantly on target-context phenomena, while achieving relatively close results in the other cases. Combining source and target context provided the best balance however, achieving the best results in all cases. In particular, the src+tgt-$n$to$n$ proved optimal and we discarded the slightly worse tgt+src-$n$to$n$ variant in the remainder of this work.

7 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: the target sentence and its back-translation would be identical for all models, as would be the original target context sentences, but the $n$to$n$ and the src+tgt-$n$to$n$ models also require back-translated context, unlike the tgt-$n$to$n$ model.

For comparison purposes we back-translated the target side of the training data for both language pairs, using a sentence-level model trained on the parallel data, and trained the main model variants strictly on the back-translated data.[4] The results are shown in Tables 8, 9 and 10, 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 X-tgt-$n$to$n$ variants proved more robust than the $n$to$n$ model. This is likely due to the latter having as context only the back-translation of the target context, while the former contain, alone or in combination with the back-translation, the original target context.

Overall, the tendencies observed using parallel data were replicated with back-translated data, with the src+tgt-$n$to$n$ model being the top-performing variant across the board, and the tgt-$n$to$n$ a close second on target-context phenomena but performing worse than the $n$to$n$ model in accuracy on the GDR-SRC+TGT and *Ellipsis infl.* with BT data. Perhaps more surprising are the results achieved by the src+tgt-$n$to$n$ model, trained on BT data, on the *Lexical cohesion* test set, where it outperformed the same variant trained on parallel data by 13 points. Additional datasets might be warranted to further assess the tendencies for these models, but 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 X-tgt-$n$to$n$ models overall. We conjecture that this is mainly due to the fact that these approaches promote target language data which are in essence correct, while discarding or reducing the role of source context data which are likely to feature back-translation errors.

	EU-ES	EN-RU
Sentence-level (PA)	31.20	31.09
$n$to$n$ (PA)	**31.96**	31.20
tgt-$n$to$n$ (PA)	**31.82**	**32.29**
src+tgt-$n$to$n$ (PA)	**31.94**	**32.32**
$n$to$n$ (BT)	25.46	29.21
tgt-$n$to$n$ (BT)	27.33	**30.10**
src+tgt-$n$to$n$ (BT)	**31.27**	29.39

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

8 Machine-translated Target Context

Following standard practice, so far we used the reference target context instead of the machine-translated output in our evaluations. This is meant to remove potential noise in terms of context translation errors and evaluate the approaches on their translation accuracy given a correct context. Using reference translations also allows for an evaluation of phenomena where more than one context translation 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 and contextual phenomena, such as coherence, are evaluated accordingly. A correct but different context translation would unfairly affect the evaluation.

Still, in practice, at inference time there are no reference translations, of course. Whereas X-to1

[4]Note that we did not mix back-translated data with the original parallel data, to strictly contrast the approaches in their ability to exploit monolingual back-translated data.

	GDR-SRC+TGT		COH-TGT	
	BLEU	ACC.	BLEU	ACC.
Sentence-level	36.28	53.67	35.04	54.00
$n$to$n$ (PA)	**40.45**	**77.67**	35.89	65.33
tgt-$n$to$n$ (PA)	39.05	72.67	**39.61**	**81.67**
src+tgt-$n$to$n$ (PA)	**41.25**	**78.67**	**40.23**	**84.67**
$n$to$n$ (BT)	41.58	**76.00**	31.02	67.00
tgt-$n$to$n$ (BT)	40.22	74.00	34.62	**81.33**
src+tgt-$n$to$n$ (BT)	**45.67**	**77.33**	**42.67**	**84.67**

Table 9: 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.
Sentence-level	30.81	51.80	22.20	27.80	28.10	50.04	31.52	45.87
$n$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
src+tgt-$n$to$n$ (PA)	**40.98**	**77.20**	**35.84**	**77.60**	**34.38**	**87.48**	**31.75**	**53.07**
$n$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
src+tgt-$n$to$n$ (BT)	**41.96**	**81.20**	**35.23**	**76.00**	**31.63**	**87.36**	**31.68**	**66.07**

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

	EU-ES	EN-RU
Sentence-level	31.20	31.09
$n$to$n$	**31.96**	31.20
tgt-$n$to$n$ (RF)	**31.82**	**32.29**
tgt-$n$to$n$ (MT)	31.08	31.52
src+tgt-$n$to$n$ (RF)	**31.94**	**32.32**
src+tgt-$n$to$n$ (MT)	30.93	31.31

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

model should not be impacted at all, the X-tgt-$n$to$n$ models are susceptible to suffer from errors in the translation of the context. To measure this aspect, we computed BLEU scores using machine-translated target sentences for X-tgt-$n$to$n$ models. The results are shown in Table 11.

Using MT output resulted in a slight degradation for EU-ES, with results on a par with the sentence-level baseline and at most 1.01 points loss compared to the use of reference translations. For EN-RU, all models achieved comparable results except those that relied on reference translations, with gains of approximately 1 point for the latter. As previously noted, the BLEU metric is known to be deficient for context-aware model evaluation, and contrastive tests provide more precise benchmarks. However, measuring MT context in terms of contrastive accuracy is not a valid option, as challenge tests rely on specific context translation choices, and the reference context is provided instead in standard practice. Note that $n$to$n$ models would also be impacted in terms of contrastive accuracy, since MT output would affect decoding.[5]

Evaluating approaches based on promoting target data in a practical scenario with imperfect machine-translated context thus faces important limitations with current document-level evaluation protocols. A proper assessment of the impact of machine-translated context would also need to take into account the quality of the translation model itself, with larger models expected to minimise context translation errors in this type of approach. We leave these aspects for future research.

[5]For completeness, in Appendix B we provide results in terms of BLEU and accuracy on the challenge tests using machine-translated context.

		GDR-SRC+TGT				COH-TGT		
Dist	% cases	$n$to$n$	tgt-$n$to$n$	src+tgt-$n$to$n$	% cases	$n$to$n$	tgt-$n$to$n$	src+tgt-$n$to$n$
1	64.67%	77.32	70.10	76.80	62.34%	69.52	85.03	86.10
2	20.67%	91.23	85.48	85.48	20.67%	66.13	90.32	85.48
3	9.33%	72.41	71.43	71.43	9.67%	51.72	72.41	75.86
4	2.00%	57.14	57.14	85.71	6.00%	50.00	83.33	83.33
5	3.33%	66.67	55.56	88.89	1.33%	25.00	50.00	75.00

Table 12: Accuracy results in Basque–Spanish according to relevant context distance.

		Deixis				Lex. Cohesion		
Dist	% cases	$n$to$n$	tgt-$n$to$n$	src+tgt-$n$to$n$	% cases	$n$to$n$	tgt-$n$to$n$	src+tgt-$n$to$n$
1	33.33%	88.66	90.49	89.63	42.75%	46.27	51.45	57.53
2	33.33%	85.82	90.07	91.02	31.50%	45.87	47.39	50.00
3	33.33%	73.02	81.89	81.77	25.75%	45.43	48.56	49.09

Table 13: Accuracy results in English–Russian according to relevant context distance.

9 Accuracy At Distance

The results so far were measured considering context as a whole. To achieve a more fine-grained view of the differences between approaches, we computed their accuracy in terms of the distance between the current sentence and the disambiguating context information, expressed in number of sentences. The results are shown in Tables 12 and 13, indicating the distance and the percentages of cases in the corresponding dataset.

The main observable tendency is that of the decreasing accuracy over distance for the $n$to$n$ model, in all cases but GDR-SRC+TGT at distance 2 (where all models perform better), in contrast with the significantly more robust accuracy of the src+tgt-$n$to$n$ model at larger distances, for Basque-Spanish in particular. The tgt-$n$to$n$ model exhibits mixed tendencies, improving or maintaining accuracy over distance 1 in some cases, but also degrading at larger distances (GDR-SRC+TGT or COH-TGT, at dist=5). Note though that larger distances are under-represented in the Basque-Spanish test sets, and may thus not be as representative.

10 Conclusions

In this work, we investigated the promotion of target context data within a standard concatenation-based approach to context-aware neural machine translation. The main incentive revolves around the fact that, for most contextual phenomena of interest for document-level machine translation, the relevant information is either in the target language or distributed on the source and target sides.

We studied simple model variants where target context sentences are concatenated to the source sentence, either in isolation or in combination with the source context. Our results in Basque-Spanish and English-Russian, over five datasets showcasing different types of contextual phenomena, showed large improvements in terms of contrastive accuracy and BLEU scores. Models where the source context was discarded altogether achieved parity or slightly underperformed on phenomena involving both source and target contexts. The variants based on augmenting the source context with target data achieved the best results across the board and were also shown to be more accurate in handling context at larger distances.

We further evaluated the use of back-translated data, with models merging target and source matching or outperforming variants trained on parallel data. We also measured the impact of using machine-translated context, although only in a limited way given current evaluation protocols for context-aware models, with slight degradation observed in terms of BLEU. The use of more robust baseline models, trained on larger volumes of data, could mitigate the observed effects.

The proposed approach promoting target data requires no changes to the standard NMT architecture and provides significant gains over strong baselines. Although it also implies larger contexts when merging source and target context, it might be worth further exploring this type of approach and the respective roles of source and target context data in neural machine translation.

References

Agrawal, Ruchit, 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)*.

Bahdanau, Dzmitry, 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*.

Bao, Guangsheng, 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, August. Association for Computational Linguistics.

Barrault, Loïc, 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, August. Association for Computational Linguistics.

Bawden, Rachel, 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, June. Association for Computational Linguistics.

Currey, Anna, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Fernandes, Patrick, 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, August. Association for Computational Linguistics.

Gete, Harritxu and Thierry Etchegoyhen. 2023. An evaluation of source factors in concatenation-based context-aware neural machine translation. In Mitkov, Ruslan and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 399–407, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.

Gete, Harritxu, 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.

Gete, Harritxu, 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, June. European Association for Machine Translation.

Hendy, Amr, 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.

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

Huo, Jingjing, 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, November. Association for Computational Linguistics.

Jean, Sebastien, 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, September. Association for Computational Linguistics.

Jin, Linghao, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore, December. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, 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, July. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin. 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, August. Association for Computational Linguistics.

Karpinska, Marzena and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Koehn, Philipp, 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.

Koehn, Philipp. 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, July. Association for Computational Linguistics.

Läubli, Samuel, 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, October-November. Association for Computational Linguistics.

Li, Bei, 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, July. Association for Computational Linguistics.

Lison, Pierre, 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, May. European Language Resources Association (ELRA).

Lopes, António, 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, November. European Association for Machine Translation.

Lupo, Lorenzo, 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), December.

Lupo, Lorenzo, 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, May.

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

Mansimov, Elman, 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, November. Association for Computational Linguistics.

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

Müller, Mathias, 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, October. Association for Computational Linguistics.

Nagata, Masaaki 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, May. European Language Resources Association.

Papineni, Kishore, 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, July. Association for Computational Linguistics.

Petrick, Frithjof, Christian Herold, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2023. Document-level language models for machine translation. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings*

of the Eighth Conference on Machine Translation, pages 375–391, Singapore, December. Association for Computational Linguistics.

Post, Matt and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv:2304.12959v1*.

Post, Matt. 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, October. Association for Computational Linguistics.

Scherrer, Yves, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China, November. Association for Computational Linguistics.

Sennrich, Rico, 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, August. Association for Computational Linguistics.

Sugiyama, Amane 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, November. Association for Computational Linguistics.

Sun, Zewei, 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, May. Association for Computational Linguistics.

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

Tiedemann, Jörg 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, September. Association for Computational Linguistics.

Vaswani, Ashish, 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.

Voita, Elena, 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, July. Association for Computational Linguistics.

Voita, Elena, 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, November. Association for Computational Linguistics.

Voita, Elena, 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, July. Association for Computational Linguistics.

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

Wong, Billy T. M. and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July. Association for Computational Linguistics.

Wu, Xueqing, 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.

Xu, Mingzhou, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021. Document graph for neural machine translation. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *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.

Zhang, Jiacheng, 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, October-November. Association for Computational Linguistics.

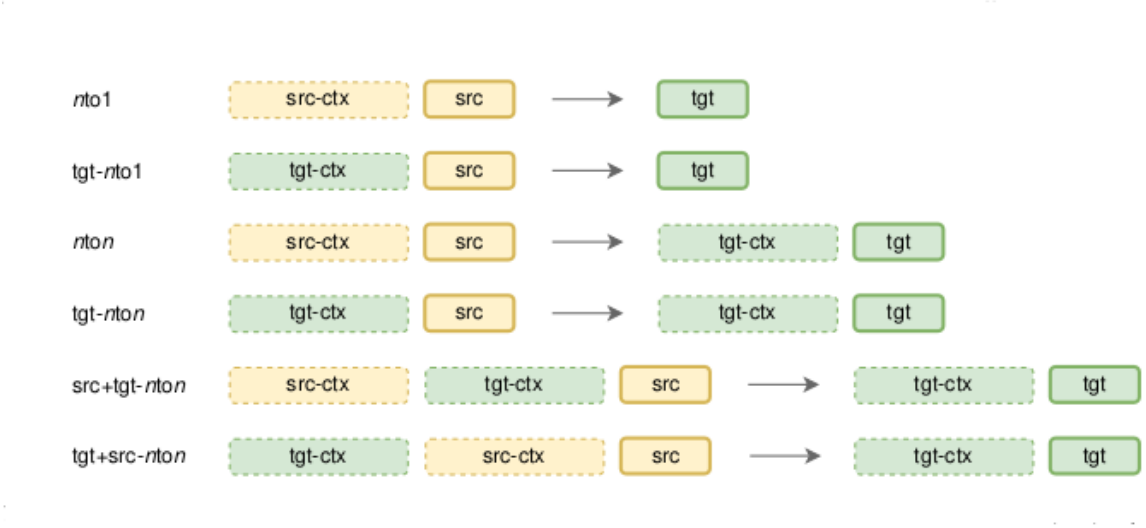

Figure 1: Schematic representation of a training instance for the different models. The yellow blocks represent the source language and the green blocks the target language. The dashed lines indicate context sentences; the continuous lines indicate the current sentence and its translation.

A Models Overview

To clarify the differences between model variants, Figure 1 provides a schematic view of the composition of a training instance for each type of concatenation-based model. We show the main building blocks and their ordering for both source and target sides.

B Machine-translated Target Context on Challenge Tests

To complement the results in Section 8, we evaluated the models on the challenge test sets using the machine-translated context instead of the reference translation in the test. Although this would be the process at inference time, as previously noted the challenge test sets depend on pre-established translation choices, in particular for coherence. A machine-translated context sentence might be entirely correct but differ from the specific translation choice the test has been designed for. The reference target context is thus typically provided as is on these test sets for standard approaches such as the $n$to$n$ model and we followed this protocol for our main results.

With these caveats in mind, we computed results in terms of BLEU and accuracy using machine translated-context on a subset of the challenge tests, with the results shown in Table 14. For this evaluation, we discarded the tests where the disambiguating information is present only in the target context, as this would lead to erroneous results, for the reasons mentioned above. Thus, the evaluation was restricted to the GDR-SRC+TGT test for Basque-Spanish, and on the ellipsis-related tests for English-Russian. Although the contrastive results on these challenge tests might still be impacted by differing translation choices, the source context might contain sufficient information to compensate for these variations.

Using MT output impacted all the models that promoted the target context, in terms of both BLEU and accuracy scores, except in Basque-Spanish on BLEU where the loss was not statistically significant. However, these variants still outperformed the sentence-level baselines in a significant way across the board.

In English-Russian, the src+tgt-$n$to$n$ model using machine-translated context achieved better results than all other models on *Ellipsis VP*, excepting the same variant using reference translations. It was notably better than the $n$to$n$ and the tgt-$n$to$n$ models with reference target context. The situation is reversed on *Ellipsis infl.*, with significant losses for the src+tgt-$n$to$n$ (MT) model compared to src+tgt-$n$to$n$ (RF), and the $n$to$n$ model achieving better results with MT context. Note that the $n$to$n$ model also incurred significant losses in terms of accuracy when using MT context in this case. This is not unexpected, as the decoding process involves the target context in these models, with cascading divergences between the machine-translated target context and the expected context in the contrastive test. Note that this type of model is not impacted by the use of MT output in terms

	EU-ES		EN-RU			
	GDR-SRC+TGT		Ellipsis infl.		Ellipsis VP	
	BLEU	ACC.	BLEU	ACC.	BLEU	ACC.
Sentence-level	36.28	53.67	30.81	51.80	22.20	27.80
$n$to$n$ (RF)	**40.45**	**77.67**	36.97	**75.20**	29.59	62.60
$n$to$n$ (MT)	**40.45**	**74.33**	36.97	67.40	29.59	63.20
tgt-$n$to$n$ (RF)	39.05	72.67	**40.69**	70.00	30.75	60.00
tgt-$n$to$n$ (MT)	37.45	69.33	34.44	62.40	30.18	55.20
src+tgt-$n$to$n$ (RF)	**41.25**	**78.67**	**40.98**	**77.20**	**35.84**	**77.60**
src+tgt-$n$to$n$ (MT)	**39.63**	73.33	36.40	62.20	33.36	71.40

Table 14: Results on contrastive tests using reference (RF) and machine-translated (MT) context.

of BLEU, however, as the translated context is discarded after translation in non-contrastive evaluations.

In Basque-Spanish, the slight loss in BLEU between src+tgt-$n$to$n$ (RF) and src+tgt-$n$to$n$ (MT) was not statistically significant. In terms of accuracy, the losses were notable between these two models however, at over 5 points, but marginal between the src+tgt-$n$to$n$ (MT) and the $n$to$n$ (MT) models (1 point).

As previously discussed, contrastive tests are meant for a specific context, and evaluations with machine-translated output are only tentative. Different evaluation protocols would be needed to evaluate the use of MT context in a more principled and robust manner.