Targeted Data Augmentation Improves Context-aware Neural Machine Translation

Harritxu Gete[1,2] hgete@vicomtech.org
Thierry Etchegoyhen[1] tetchegoyhen@vicomtech.org
Gorka Labaka[2,3] gorka.labaka@ehu.eus

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]University of the Basque Country UPV/EHU
[3]HiTZ Basque Center for Language Technologies - Ixa

Abstract

Progress in document-level Machine Translation is hindered by the lack of parallel training data that include context information. In this work, we evaluate the potential of data augmentation techniques to circumvent these limitations, showing that significant gains can be achieved via upsampling, similar context sampling and back-translations, targeted on context-relevant data. We apply these methods on standard document-level datasets in English-German and English-French and demonstrate their relevance to improve the translation of contextual phenomena. In particular, we show that relatively small volumes of targeted data augmentation lead to significant improvements over a strong context-concatenation baseline and standard back-translation of document-level data. We also compare the accuracy of the selected methods depending on data volumes or distance to relevant context information, and explore their use in combination.

1 Introduction

Neural Machine Translation (NMT) models (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) are typically trained and used to translate sentences in isolation, ignoring their context of occurrence. This limitation impends the accurate translation of linguistic phenomena that depend on context information, such as discursive coreference or coherence, among others (Bawden et al., 2018; Lopes et al., 2020). A number of approaches have been devised in NMT to extend the modeling window beyond isolated sentences. These approaches range from extending the input by including context sentences (Tiedemann and Scherrer, 2017) to architectural variants (Jean et al., 2017; Zhang et al., 2018; Voita et al., 2019b; Li et al., 2020). Despite the improvements achieved by these methods, the lack of training data that includes contextual information is hindering progress in the field, with only relatively recent efforts to provide large parallel datasets that preserve document boundaries (Barrault et al., 2019).

Data augmentation is one of the main methods to increase machine translation coverage at the sentence level, typically via back-translation of monolingual data (Sennrich et al., 2016a) or comparable data mining (Sharoff et al., 2014). For document-level NMT, fewer studies have addressed the use of data augmentation to tackle the aforementioned scarcity. Back-translation at the document level has been shown to help context-aware NMT (Junczys-Dowmunt, 2019; Sugiyama and Yoshinaga, 2019; Huo et al., 2020), but its use has been limited to bulk back-translation rather than targeting contextual phenomena. Other data augmentation methods such

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 298–312
September 4–8, 2023, Macau SAR, China.

as data alteration based on coreference resolvers (Stojanovski et al., 2020; Hwang et al., 2021) have also been shown to be useful for the task. Overall, it is currently unclear whether data augmentation that do not rely on bulk back-translation or external tools can provide any benefits for context-aware NMT.

In this work, we explore different approaches to data augmentation for context-aware NMT, which, to the best of our knowledge, have not yet been studied in depth. We thus evaluate the use of upsampling, context sampling and back-translations, targeted on context-relevant data. Our experiments focus on pronoun translation with a single context sentence, to provide initial results in a constrained experimental protocol, and are evaluated on standard datasets, namely ContraPro (Müller et al., 2018) for English-German and the large-scale pronoun test set for English-French (Lopes et al., 2020). We show that significant gains can be achieved by each method over a strong baseline, with relatively small quantities of augmented data, and provide a detailed analysis of these methods in isolation and in combination.

2 Related work

A variety of studies have tackled context-aware approaches within the framework of NMT, analysing the improvements that these models can provide over non-contextual baselines (Li et al., 2020; Ma et al., 2020; Lopes et al., 2020; Lupo et al., 2022; Majumde 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 (Tiedemann and Scherrer, 2017). This simple approach is still one of the most efficient methods to perform context-aware neural machine translation, matching or outperforming more sophisticated ones (Lopes et al., 2020). Alternative methods have involved refining the context-agnostic translations (Xiong et al., 2019; Voita et al., 2019a; Mansimov et al., 2021), or modelling context information with specific NMT architectures (Jean et al., 2017; Zhang et al., 2018; Li et al., 2020; Wang et al., 2017; Tan et al., 2019).

The growing interest in context-aware NMT models has increased the need for parallel data where context information is preserved. Dedicated efforts have been made to increase the availability of this type of data, for instance in recent shared tasks in the WMT series (Barrault et al., 2019). However, context boundaries might not always be recoverable, ensuring continuous contextual information in sentence-aligned datasets can be a costly task, and most of the available relevant data might be limited to specific domains. Data augmentation might thus complement the existing datasets for the variety of possible language pairs and domains.

Over the years, specific efforts have been made to create synthetic data to improve NMT at the sentence-level (Fadaee et al., 2017; Li et al., 2019; Li and Specia, 2019; Xia et al., 2019; Liu et al., 2021). The most widespread method is the use of back-translations, a technique introduced to NMT by Sennrich et al. (2016a) that exploits monolingual corpora by machine-translating target language data into the source language. For document-level NMT, back-translation has been shown to be effective in capturing contextual information, both by translating the original data sentence by sentence (Junczys-Dowmunt, 2019) or by using context-aware models (Sugiyama and Yoshinaga, 2019). In the same vein, Huo et al. (2020) find that document-level models benefit even more from back-translations than their sentence-level counterparts. To our knowledge, back-translations targeted on specific phenomena, as proposed by Fadaee and Monz (2018) for sentence-level models, have not been investigated for context-aware NMT and we include this approach among our data augmentation methods.

Monolingual data have also been exploited for document-level NMT via context-level decoders (Voita et al., 2019b) or systems that learn to improve the translations generated by sentence-level models (Voita et al., 2019a). Other methods augment document-level parallel data by creating synthetic sentence sequences via the concatenation of varying numbers of sentences extracted from aligned document pairs (Popel et al., 2019; Popel, 2020; Nowakowski

et al., 2022). Other forms of data augmentation are antecedent-free augmentation (Stojanovski et al., 2020), which creates new training examples by modifying cases where the antecedent is not present in the available context, or the more recent method of Hwang et al. (2021), which generates faulty data and trains NMT models via contrastive learning. In both cases, a coreference analysis needs to be performed on document pairs. Finally, data augmentation has also been performed for sentence-level models by mining large volumes of comparable data (Sharoff et al., 2014). This type of data has been shown to increase the quality of NMT models for low-resource languages, independently or in combination with back-translations (Gete and Etchegoyhen, 2022). To our knowledge, using similar data for contextual data augmentation has not yet been explored, and we include a variant of this method in our analysis.

Context-aware models are particularly suited to improve the translation of phenomena that directly depend on context information, such as intersentential anaphora resolution, discourse coherence or terminological consistency (Müller et al., 2018). We evaluate our approach on the specific task of adequately translating pronouns in context, for which several specific test sets have been created (Guillou and Hardmeier, 2016; Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Lopes et al., 2020; Gete et al., 2022).

3 Methodology

We aim to generate synthetic parallel data that include relevant information for the translation of specific contextual phenomena. This involves (i) identifying context blocks in document-level data, i.e. parallel sequences consisting of a sentence and its previous context sentence in the source and target languages, and (ii) sampling blocks that contain elements whose translation typically requires context information. Although our approach could be applied to other contextual phenomena as well, we selected pronouns as our linguistic category of interest, specifically the translation of pronouns from English into German and French, given their relevance for document-level translation and the availability of contrastive test sets for precise evaluations (Müller et al., 2018; Lopes et al., 2020). In particular, for English-German, we focused on the pronoun *it* which can be translated as *es* (neutral gender), *er* (masculine) or *sie* (feminine). For English-French, in addition to *it*, which can be translated as *elle* (feminine) or *il* (masculine), we also included *they*, which can be translated as *elles* (feminine) or *ils* (masculine).

We first identify context blocks where the targeted elements occur in the source (src_i) and target (tgt_i) sentences and the preceding source sentence (src_{i-1}) is available. More specifically, we extracted context blocks that met one of the following conditions: (i) *it* in src_i and *es*/*er*/*sie* in tgt_i (EN-DE) (ii) *it* in src_i and *elle*/*il* in tgt_i (EN-FR) (iii) *they* in src_i and *elles*/*ils* in tgt_i (EN-FR). Under this approach, we might sample data where the antecedent of the pronoun is found in the block, but might also extract blocks where the antecedent is not included. These instances can also be useful as they might help balance the data in case of bias. This extraction method avoids having to use coreference annotation tools, which simplifies the data extraction process. To avoid introducing ambiguity in the sampled data, we discarded cases where more than one pronominal translation with different genders appeared in the target sentence.

After sampling the blocks of interest, we create new ones by either duplicating the sampled blocks (*upsampling*), replacing the context sentences randomly or via sentence embedding similarity (*context sampling*), or back-translating the target language blocks (*targeted back-translation*). We describe each method in more details below.

Upsampling. This method (hereafter, UP-SAMP) is the simplest, and consists in repeating the selected blocks multiple times and adding them to the training data. This type of data augmentation could lead to overfitting, i.e. overtraining the model on the upsampled data and learning specific patterns which might be irrelevant in other cases. It may thus happen that the model achieves higher accuracy on the selected data but does not generalise well to other data.

Context Sampling. To avoid the overfitting that may arise from upsampling, context sampling uses context blocks as a basis to create synthetic data. To do this, the sentences src_i and tgt_i remain unchanged, but the English source context (src_{i-1}) is replaced by another sentence from the corpus. To select the substitute sentence, we first retrieve blocks which contain the same target pronoun and may thus contain varying but useful context. We then select the replacement context sentence among the retrieved blocks via one of two methods: random sampling (RDM-SAMP) and similarity sampling (SIM-SAMP). Random sampling is meant to evaluate unconstrained substitution by randomly selecting any context sentence within the candidate blocks. Note that the antecedent is likely to be replaced by a semantically unrelated one, which could impact the final quality of the model. Similarity sampling is performed by selecting the most similar context in terms of cosine similarity using pretrained sentence embeddings.[1]

Targeted Back-translation. Our final method is targeted back-translation (T-BT), where we back-translate specific portions of document-level monolingual data, selecting (tgt_{i-1}, tgt_i) blocks where tgt_i contains one of the targeted pronouns. As in bilingual data extraction, if the sentence contains a pronoun, the pronoun corresponding to the other gender cannot appear in the sentence. The selected blocks are translated into the source language using a context-agnostic NMT model and blocks where the back-translation does not contain a translation of the targeted pronoun are discarded.

4 Experimental setup

4.1 Data

All selected datasets were normalised, tokenised and truecased using Moses (Koehn et al., 2007) scripts and segmented with BPE (Sennrich et al., 2016b), using 32,000 operations. Table 1 describes corpora statistics, indicating the amount of data with context information (DOC-LEVEL) and without (SENT-LEVEL), for parallel and monolingual datasets.

	EN-DE	EN-FR		DE	FR
	DOC-LEVEL	SENT-LEVEL	DOC-LEVEL	DOC-LEVEL	DOC-LEVEL
TRAIN	5,852,458	11,221,790	234,738	58,979,140	106,830,385
DEV	2,999	4,992	5,818	-	-
TEST	6,002	-	1,210	-	-

Table 1: Corpora statistics (number of sentences)

Parallel corpora. For English-German, we follow the setup of Müller et al. (2018) and selected the data from the WMT 2017 news translation task, using newstest2017 and newstest2018 as test sets and the union of newstest2014, newstest2015 and newstest2016 for validation. For English-French, we follow Lopes et al. (2020) and use publicly available sentence-level parallel data to train baseline models. We used Europarl v7, NewsCommentary v10, CommonCrawl, UN, Giga from WMT 2017 and the IWSLT17 TED Talks (Cettolo et al., 2012) processed at the sentence-level. We then fine-tune context-aware models on the document-level IWSLT17 dataset, using the test sets from 2011 to 2014 as dev sets, and 2015 as test sets.

Monolingual corpora. We use NewsCrawl2021 (Barrault et al., 2021) as German monolingual data and OpenSubtitles2018 (Lison et al., 2018) for French.

[1]Embeddings were computed with all-MiniLM-L6-v2 (https://www.sbert.net/docs/pretrained_models.html).

EN-DE		EN-FR	
it→es	221,327	*it→elle*	3,539
it→er	40,238	*it→il*	13,252
it→sie	105,906	*they→elles*	2,886
		they→ils	14,967
TOTAL	367,471		34,644

Table 2: Extracted context data per target category (number of sentences)

Contrastive tests. We evaluate our models using two sets of contrastive tests, both created from OpenSubtitles2018[2] excerpts and aiming to assess a model's ability to rank correct translations over incorrect ones. ContraPro (Müller et al., 2018) enables testing the ability of a model to identify the correct German translation of the English anaphoric pronoun *it* as *es*, *sie* or *er*. It contains 4,000 examples per pronoun and, for 80% of them, the sentence-based antecedent distance is superior to 0. The EN-FR large-scale pronoun test set (hereafter, LSCP) (Lopes et al., 2020) is similar, but in addition to assessing the translation of *it* as *elle* or *il*, it includes the translation of *they* as *elles* or *ils*. It consists of 3,500 examples for each type of pronoun and almost 60% of the examples need contextual information to make the correct choice.

4.2 Models

All models follow the Transformer-base architecture (Vaswani et al., 2017) and were trained with the MarianNMT toolkit (Junczys-Dowmunt et al., 2018). The embeddings for source, target and output layers were tied and optimisation was performed with Adam (Kingma and Ba, 2015), with $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. As baselines, we trained sentence-level models and 2to1 models. The latter is a context-aware approach that extends the input by concatenating the previous sentence without any changes to the model architecture (Tiedemann and Scherrer, 2017), including an additional sentence break token between the context and the current sentence.

For English-German, both the sentence-level model and the 2to1 baseline were trained with the available document-level corpora, and the parameters of the 2to1 model were initialised with those of the sentence-level model. For English-French, due to the lower data volumes, a sentence-level model was first trained with sentence-level data. Following Müller et al. (2018), this model was then fine-tuned with document-level data to obtain a 2to1 model. In addition, 2to1 models are trained also on the augmented data, with varying quantities and different data distributions to balance or maintain the distribution of pronouns in the original datasets.

5 Optimal Variants

We first aimed to establish the optimal selection of data along two lines: (i) balancing the distribution of pronominal categories vs. maintaining an unbalanced distribution, and (ii) varying the amounts of sampled data with each method.

5.1 Distribution Balance

As shown in Table 2, the distribution per pronominal category in the extracted context blocks is unbalanced. To balance the data, we increased the representation of the least represented categories to reach the volumes of the most represented one. For English-French, given the relatively lower data volumes, we raised the amounts of data to a minimum of 45K for all categories, rather than just matching the volumes of the most represented one. For each method, we compared balancing with data augmentation maintaining the original distribution of the

[2]Note that training data were filtered so as not to include examples of the contrastive tests.

training data. To maintain the distribution, n blocks were created for each extracted block, choosing the smallest n so that the amount of data reached the amount in the balanced data. These quantities were reached with $n = 1$ for English-German and $n = 5$ for English-French.

For comparison purposes, we also include results from untargeted back-translation, i.e. standard back-translation of document-level monolingual data. We trained a BT-SMALL model with the same amount of data added to balance the distribution (296K in total for English-German and 145K for English-French) and a larger version, BT-LARGE, with 1.1M and 765K back-translations for English-German and English-French, respectively. Note that, in this case, no selection of the data is performed, so the final distribution does not necessarily maintain the original distribution and is not necessarily balanced.

	TOTAL	ES	ER	SIE	Δ
2TO1	0.58	**0.92**	0.38	0.43	0.54
UP-SAMP (B)	**0.69**	0.81	**0.70**	0.55	0.26
UP-SAMP (O)	0.62	0.91	0.43	0.52	0.48
RDM-SAMP (B)	0.64	0.83	0.55	0.53	0.30
RDM-SAMP (O)	0.58	0.90	0.37	0.48	0.53
SIM-SAMP (B)	0.65	0.82	0.62	0.51	0.31
SIM-SAMP (O)	0.61	0.91	0.42	0.49	0.49
T-BT (B)	0.66	0.71	0.66	**0.60**	0.11
T-BT (O)	0.62	0.88	0.41	0.57	0.47
BT-SMALL	0.59	0.91	0.39	0.48	0.52
BT-LARGE	0.59	**0.92**	0.39	0.47	0.53

Table 3: English-German accuracy results. (B) and (O) indicate balancing and maintaining the original data distribution, respectively. Δ is the difference in accuracy between best and worst categories. Best results for each category are shown in bold.

	TOTAL	ELLE	IL	ELLES	ILS	Δ
2TO1	0.84	0.80	0.92	0.67	0.98	0.31
UP-SAMP (B)	**0.87**	0.90	0.85	0.77	0.96	0.19
UP-SAMP (O)	0.86	0.82	0.92	0.71	0.98	0.27
RDM-SAMP (B)	0.86	0.90	0.83	0.78	0.95	0.17
RDM-SAMP (O)	0.84	0.79	0.91	0.68	0.98	0.30
SIM-SAMP (B)	**0.87**	0.89	0.84	0.78	0.96	0.18
SIM-SAMP (O)	0.85	0.80	0.91	0.69	0.98	0.29
T-BT (B)	0.85	**0.91**	0.79	**0.83**	0.86	0.12
T-BT (O)	0.85	0.76	**0.94**	0.72	0.98	0.26
BT-SMALL	0.84	0.79	0.92	0.65	**0.99**	0.33
BT-LARGE	0.84	0.79	0.92	0.65	**0.99**	0.34

Table 4: English-French accuracy results. (B) and (O) indicate balancing and maintaining original data distribution, respectively. Δ is the difference in accuracy between best and worst categories. Best results for each category are shown in bold.

The results for this first set of experiments are provided in Tables 3 and 4. Balancing reduces the difference in accuracy between the different genders, to a marked extent, and although it has a negative impact on the most represented categories (*es* in German, *il* and *ils* in French), it markedly increases the accuracy for the less represented ones. Overall, balancing clearly improves over keeping the original data distribution, and we thus opted to balance all datasets in the remaining experiments. Of note are the significant total improvements obtained

in English-German, and the smaller ones for English-French, where the baseline 2to1 method already achieves relatively high accuracy. The use of untargeted back-translations, whether in smaller or larger quantities, performed on a par with the baseline, maintaining a distribution of scores very similar to the original one. In both cases, this method was outperformed by targeted data augmentation methods, in all but the top-scoring cases for each language pair (*es* in English-German and *ils* in English-French), where it achieved marginally better scores.

5.2 Data Size

We then turned to measuring the impact of different volumes of augmented data. For English-German, we started from the minimal balanced data size and augmented the data by increments of 100,000, up to 500,000 per category; for English-French, given the smaller amounts, the increments were made on a 10,000 basis, starting from 15,000 and up to 45,000, with an additional increase reaching 200,000 instances per category, to test the impact of larger datasets. Note that not all data increases were feasible with T-BT, as there were not enough data meeting the targeted sampling criteria. Therefore, for this method, the maximum available data were 226,651 instances per pronoun category in English-German and 126,905 in English-French.

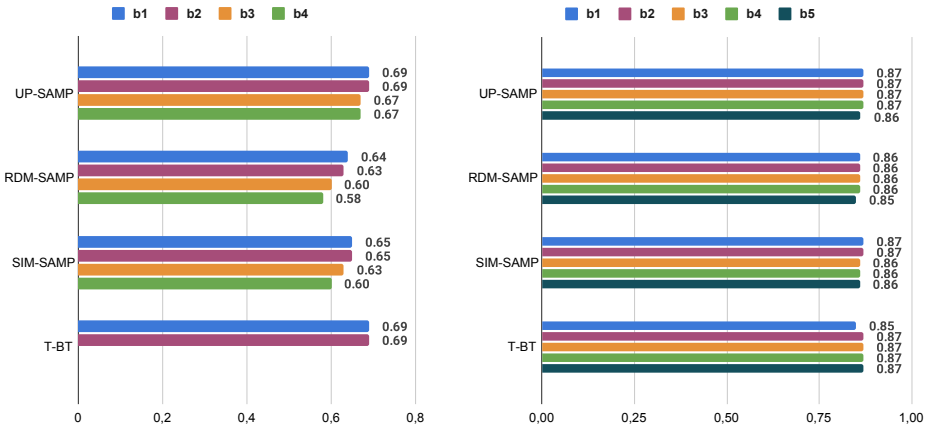

Figure 1: Accuracy results for English-German (left) and English-French (right) as a factor of augmented data size. For English-German, b1=221,327, b2=300K, b3=400K and b4=500K except for T-BT, where b2=226,651. For English-French, b1=15K, b2=25K, b3=35K, b4=4K and b5=200K except for T-BT, where b5=126,905.

Accuracy results obtained with varying amounts of augmented data are shown in Figure 1. Increasing the data size brought no improvement or was detrimental for all models except for the English-French T-BT. Adding data beyond what was needed for data balancing did not improve the upsampling and contextual sampling models, even in a more data-sparse scenario such as English-French. This might be caused by the overfitting arising from upsampling and the noise introduced by sampling methods with incorrect contexts, although a more detailed analysis, beyond the scope of this work, would be needed to establish the determining factors for this behaviour. In the case of T-BT, for English-German the results remained identical, which is not unexpected considering that very little data could be added. In the case of English-French, where there was less initial data, increasing up to 25K instances per category improved accuracy. In what follows, therefore, we opted for the smallest data sizes for each model, except for English-French T-BT where we selected 25K cases per pronominal category.

6 Method Comparison

In this section, we compare the methods selected in the previous section, i.e. balanced 221,327 for English-German and balanced 15K for English-French, except for English-French T-BT, with 25K selected. Additionally, for each language pair we trained a combined model (COMB) where we merged the augmentation blocks from each method and selected a random sample maintaining distribution balance. We discarded the option of using the combination of all data, as this would have resulted in unbalanced data distributions.

6.1 Comparative Accuracy

Accuracy results, including total and pronoun-specific results, are shown in Figure 2. All data augmentation methods improve over both the sentence-level and the 2to1 baselines, although the improvements are more marked in English-German, where the baselines are less accurate than in English-French. The high scores obtained by the English-French baselines may be due to several factors. On the one hand, as previously mentioned, this test set contains over 40% of examples where the context is not necessary to make a correct translation. On the other hand, this is a less varied test than the corresponding test for English-German, since it only includes subject pronouns whose antecedent is a noun. Although a more detailed analysis would be needed to confirm this conjecture, the uniformity and relative simplicity of antecedent-pronoun configuration might be a relevant factor for the rather high scores obtained by the baseline. Improving over these baseline results might thus be a challenge for any method on this test set.

The sampling methods are better at preserving the distribution of the most frequent pronouns, with upsampling outperforming both random and similarity sampling for English-German. For this language pair, T-BT is outperformed by upsampling in most cases but performs better than all other methods on translation of *sie*, the less accurately translated category overall. The combination provides balanced results across categories and achieves the best results on the initially least represented *er* category, but also loses accuracy for the most frequent *es* pronoun. Similar results are obtained for English-French, where it obtains the best results for *elle* and *elles*, and the worst for *il* and *ils*. T-BT and COMB obtain the most balanced results, to the detriment of *ils*, the category with the best results with the other methods.

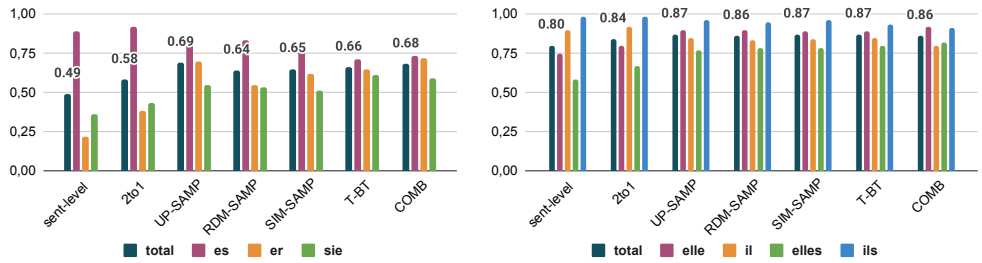

Figure 2: Accuracy results in English-German (left) and English-French (right) for all selected models. Numerical results are indicated for total accuracy.

6.2 Impact of Distance

The results so far indicate that accuracy increases when using the selected data augmentation methods, overall and per category. However, since the contrastive test sets include data where the relevant pronoun antecedent can occur within the same sentence, the extent to which the observed improvements come from an actual improved use of the preceding context is unclear.

In Figure 3, we compare results for cases where the antecedent is in the same sentence

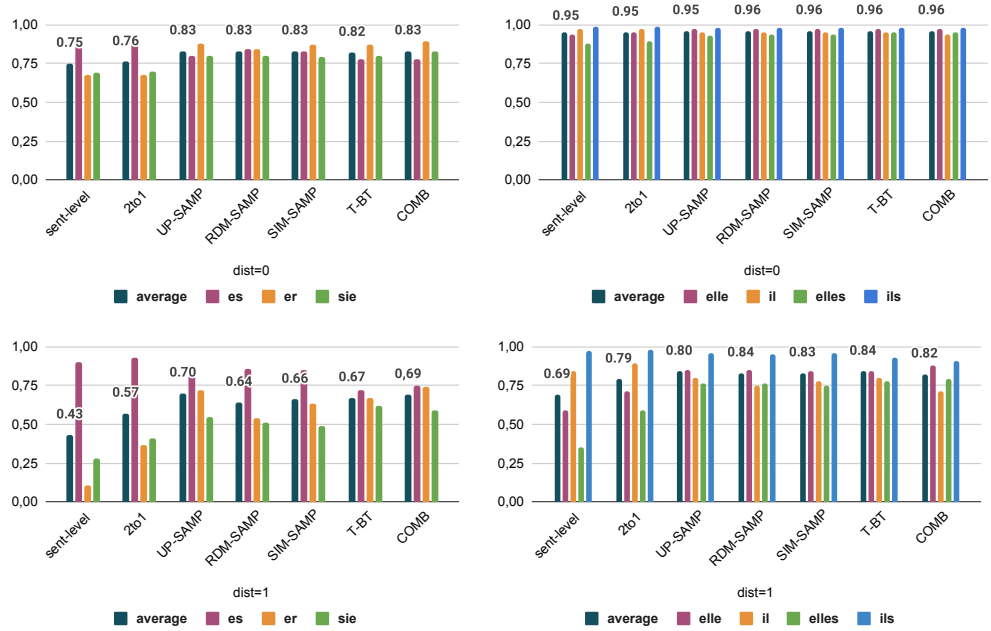

Figure 3: Accuracy results as a factor of antecedent distance in English-German (left) and English-French (right). Numerical results are indicated for average accuracy.

(dist=0) or in the preceding context sentence (dist=1). Here, we indicate the average across categories instead of total accuracy, since categories are distributed differently in the test set depending on antecedent distance. For both language pairs, the improvements are markedly larger across categories when the relevant context is in the preceding sentence, although all methods also match or improve over the baselines when the antecedent occurs within the same sentence. These results thus indicate that the selected data augmentation methods do improve context appraisal beyond the current sentence, with additional improvements at the sentence level. Untangling the precise impact of the augmented data in both cases would require additional experiments which we leave for future work.

6.3 Impact on translation metrics

Finally, since data augmentation may impact the resulting models in terms of general translation quality, we computed BLEU (Papineni et al., 2002) scores on both sentence-level and document-level test sets. The scores were computed with the SacreBLEU[3] toolkit (Post, 2018) and statistical significance was computed via paired bootstrap resampling (Koehn, 2004). The results are shown in Table 5.

Overall, T-BT was the optimal method preserving general translation quality, improving over both baselines in most cases. These differences are more marked in the case of English-French, where T-BT was the only method that improved over the two baselines in all cases. Upsampling induced BLEU loss across the board in English-German when compared to the sentence-level baseline, a result which may be due to the overfitting resulting from this method. Both random and similarity sampling performed worse than T-BT in general, although they slightly improved over the 2to1 baselines on several test sets. Finally, COMB obtained relatively balanced results across test sets, outperforming both baselines in most cases.

[3]signature: nrefs:1—case:mixed—eff:no—tok:13a—smooth:exp—version:2.0.0

The different methods we examined in this work do not seem to negatively impact the models' general translation capability, and may even improve over both sentence-level and 2to1 models in this respect. Combining these results with those achieved on the contrastive test sets, it appears that the data augmentation techniques evaluated in this work can thus contribute to improving translation quality of context-aware NMT models overall.

	EN-DE			EN-FR	
	wmt2017	wmt2018	ContraPro	iwslt17	ContraPro
SENTENCE-LEVEL	27.7	41.1	22.7	41.2	27.7
2TO1	26.8	40.7	23.4	42.6	28.7
UP-SAMP	26.8†	40.1†‡	24.8†‡	42.6†	29.2†‡
RDM-SAMP	27.4‡	40.7	24.5†‡	42.2†‡	29.1†‡
SIM-SAMP	27.5‡	40.4†	24.7†‡	42.4†	29.1†‡
T-BT	28.0‡	41.7†‡	**24.9**†‡	**42.9**†‡	**29.7**†‡
COMB	27.8‡	41.1	**25.0**†‡	42.4†	29.3†‡
BT-SMALL	28.3†‡	41.7†‡	24.1†‡	42.4†	28.9†‡
BT-LARGE	**28.8**†‡	**42.4**†‡	24.4†‡	42.2†	28.8†‡

Table 5: BLEU results. † and ‡ indicate statistically significant results ($p < 0.05$) against the sentence-level and 2to1 baselines, respectively; best performing systems, without statistically significant differences between them, are shown in bold.

7 Conclusions

In this work, we described three different data augmentation techniques for context-aware NMT and evaluated them in isolation and in combination over standard sentence-level and document-level test sets. Specifically, we created synthetic data centred on improving pronoun translation in English-German and English-French, as a test case for an approach which could be applied to other contextual phenomena as well, provided they feature overt elements that may be targeted.

The methods we examined included upsampling, context sampling with both random and similar context substitution, and back-translations, all targeted on specific data featuring different pronominal types. All methods improved over a strong concatenation baseline, in terms of accuracy on contrastive test sets, while also achieving parity or improving in terms of BLEU scores in most cases. Accuracy improvements were markedly larger on the English-German contrastive sets, as high scores could already be obtained by the baseline on the English-French test sets. We leave for future work an exploration of alternative contrastive datasets and models with a wider contextual window. We demonstrated that balancing the data and using minimal volumes was optimal overall, and showed that the improvements were mainly obtained by leveraging contextual information in preceding sentences. All methods were shown to perform markedly better than simply back-translating document-level data, indicating that targeted data augmentation might be a research path worth exploring further for context-aware NMT.

Finally, among the selected methods, targeted back-translation proved a simple and effective approach which performed well across the board, although it can be outperformed in terms of accuracy on specific categories. This method does not require external tools such as coreference resolvers and can significantly improve the results of a 2to1 model with relatively small amounts of data, as measured in contrastive evaluations as well as evaluations in terms of BLEU. The combination of data from the different examined methods may also be considered a viable alternative, as it resulted in balanced improvements over categories overall.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors (2021). *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (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.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (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.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Fadaee, M. and Monz, C. (2018). Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Gete, H. and Etchegoyhen, T. (2022). Making the most of comparable corpora in neural machine translation: a case study. *Lang. Resour. Evaluation*, 56(3):943–971.

Gete, H., Etchegoyhen, T., Ponce, D., Labaka, G., Aranberri, N., Corral, A., Saralegi, X., Ellakuria, I., and Martin, M. (2022). TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France. European Language Resources Association.

Guillou, L. and Hardmeier, C. (2016). PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (2018). A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Huo, J., Herold, C., Gao, Y., Dahlmann, L., Khadivi, S., and Ney, H. (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.

Hwang, Y., Yun, H., and Jung, K. (2021). Contrastive learning for context-aware neural machine translation using coreference information. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1135–1144, Online. Association for Computational Linguistics.

Jean, S., Lauly, S., Firat, O., and Cho, K. (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.

Junczys-Dowmunt, M. (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.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Koehn, P. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 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.

Li, B., Liu, H., Wang, Z., Jiang, Y., Xiao, T., Zhu, J., Liu, T., and Li, C. (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.

Li, G., Liu, L., Huang, G., Zhu, C., and Zhao, T. (2019). Understanding data augmentation in neural machine translation: Two perspectives towards generalization. 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 5689–5695, Hong Kong, China. Association for Computational Linguistics.

Li, Z. and Specia, L. (2019). Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336, Hong Kong, China. Association for Computational Linguistics.

Lison, P., Tiedemann, J., and Kouylekov, M. (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).

Liu, Q., Kusner, M., and Blunsom, P. (2021). Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. T. (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.

Lupo, L., Dinarelli, M., and Besacier, L. (2022). Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830—842, Abu Dhabi. Association for Computational Linguistics.

Ma, S., Zhang, D., and Zhou, M. (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.

Majumde, S., Lauly, S., Nadejde, M., Federico, M., and Dinu, G. (2022). A baseline revisited: Pushing the limits of multi-segment models for context-aware translation. *arXiv preprint arXiv:2210.10906v2*.

Mansimov, E., Melis, G., and Yu, L. (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.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (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.

Nowakowski, A., Pałka, G., Guttmann, K., and Pokrywka, M. (2022). Adam mickiewicz university at wmt 2022: Ner-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326—334, Abu Dhabi. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Popel, M. (2020). CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Popel, M., Macháček, D., Auersperger, M., Bojar, O., and Pecina, P. (2019). English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy. Association for Computational Linguistics.

Post, M. (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.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). 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.

Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P. (2014). Building and using comparable corpora. In *Springer Berlin Heidelberg*.

Stojanovski, D., Krojer, B., Peskov, D., and Fraser, A. (2020). ContraCAT: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sugiyama, A. and Yoshinaga, N. (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.

Sun, Z., Wang, M., Zhou, H., Zhao, C., Huang, S., Chen, J., and Li, L. (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.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (2019). Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

Tiedemann, J. and Scherrer, Y. (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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Voita, E., Sennrich, R., and Titov, I. (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.

Voita, E., Sennrich, R., and Titov, I. (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.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017). Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Xia, M., Kong, X., Anastasopoulos, A., and Neubig, G. (2019). Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Xiong, H., He, Z., Wu, H., and Wang, H. (2019). Modeling coherence for discourse neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7338–7345. AAAI Press.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (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.