# Context-Aware Monolingual Repair for Neural Machine Translation

**Elena Voita**[1,2]   **Rico Sennrich**[4,3]   **Ivan Titov**[3,2]

[1]Yandex, Russia   [2]University of Amsterdam, Netherlands
[3]University of Edinburgh, Scotland   [4]University of Zurich, Switzerland
`lena-voita@yandex-team.ru`
`sennrich@cl.uzh.ch   ititov@inf.ed.ac.uk`

## Abstract

Modern sentence-level NMT systems often produce plausible translations of isolated sentences. However, when put in context, these translations may end up being inconsistent with each other. We propose a monolingual DocRepair model to correct inconsistencies between sentence-level translations. DocRepair performs automatic post-editing on a sequence of sentence-level translations, refining translations of sentences in context of each other. For training, the DocRepair model requires only monolingual document-level data in the target language. It is trained as a monolingual sequence-to-sequence model that maps inconsistent groups of sentences into consistent ones. The consistent groups come from the original training data; the inconsistent groups are obtained by sampling round-trip translations for each isolated sentence. We show that this approach successfully imitates inconsistencies we aim to fix: using contrastive evaluation, we show large improvements in the translation of several contextual phenomena in an English→Russian translation task, as well as improvements in the BLEU score. We also conduct a human evaluation and show a strong preference of the annotators to corrected translations over the baseline ones. Moreover, we analyze which discourse phenomena are hard to capture using monolingual data only.[1]

## 1 Introduction

Machine translation has made remarkable progress, and studies claiming it to reach a human parity are starting to appear (Hassan et al., 2018). However, when evaluating translations of the whole documents rather than isolated sentences, human raters show a stronger preference for human over machine translation (Läubli et al., 2018). These findings emphasize the need to shift towards context-aware machine translation both from modeling and evaluation perspective.

Most previous work on context-aware NMT assumed that either all the bilingual data is available at the document level (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Agrawal et al., 2018; Kuang et al., 2018; Miculicich et al., 2018) or at least its fraction (Voita et al., 2019). But in practical scenarios, document-level parallel data is often scarce, which is one of the challenges when building a context-aware system.

We introduce an approach to context-aware machine translation using only monolingual document-level data. In our setting, a separate monolingual sequence-to-sequence model (DocRepair) is used to correct sentence-level translations of adjacent sentences. The key idea is to use monolingual data to imitate typical inconsistencies between context-agnostic translations of isolated sentences. The DocRepair model is trained to map inconsistent groups of sentences into consistent ones. The consistent groups come from the original training data; the inconsistent groups are obtained by sampling round-trip translations for each isolated sentence.

To validate the performance of our model, we use three kinds of evaluation: the BLEU score, contrastive evaluation of translation of several discourse phenomena (Voita et al., 2019), and human evaluation. We show strong improvements for all metrics.

We analyze which discourse phenomena are hard to capture using monolingual data only. Using contrastive test sets for targeted evaluation of several contextual phenomena, we compare the performance of the models trained on round-trip

---

[1]The code and data sets (including round-trip translations) are available at `https://github.com/lena-voita/good-translation-wrong-in-context`.

translations and genuine document-level parallel data. Among the four phenomena in the test sets we use (deixis, lexical cohesion, VP ellipsis and ellipsis which affects NP inflection) we find VP ellipsis to be the hardest phenomenon to be captured using round-trip translations.

Our key contributions are as follows:

- we introduce the first approach to context-aware machine translation using only monolingual document-level data;

- our approach shows substantial improvements in translation quality as measured by BLEU, targeted contrastive evaluation of several discourse phenomena and human evaluation;

- we show which discourse phenomena are hard to capture using monolingual data only.

## 2 Our Approach: Document-level Repair

We propose a monolingual DocRepair model to correct inconsistencies between sentence-level translations of a context-agnostic MT system. It does not use any states of a trained MT model whose outputs it corrects and therefore can in principle be trained to correct translations from any black-box MT system.

The DocRepair model requires only monolingual document-level data in the target language. It is a monolingual sequence-to-sequence model that maps inconsistent groups of sentences into consistent ones. Consistent groups come from monolingual document-level data. To obtain inconsistent groups, each sentence in a group is replaced with its round-trip translation produced in isolation from context. More formally, forming a training minibatch for the DocRepair model involves the following steps (see also Figure 1):

1. sample several groups of sentences from the monolingual data;

2. for each sentence in a group, (i) translate it using a target-to-source MT model, (ii) sample a translation of this back-translated sentence in the source language using a source-to-target MT model;

3. using these round-trip translations of isolated sentences, form an inconsistent version of the initial groups;
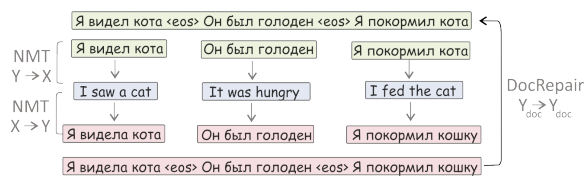


Figure 1: Training procedure of DocRepair. First, round-trip translations of individual sentences are produced to form an inconsistent text fragment (in the example, both genders of the speaker and the cat became inconsistent). Then, a repair model is trained to produce an original text from the inconsistent one.
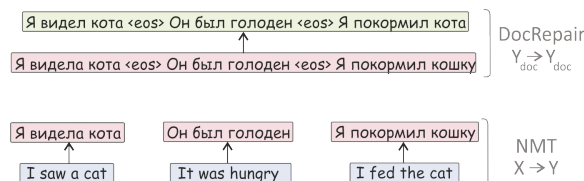


Figure 2: The process of producing document-level translations at test time is two-step: (1) sentences are translated independently using a sentence-level model, (2) DocRepair model corrects translation of the resulting text fragment.

4. use inconsistent groups as input for the DocRepair model, consistent ones as output.

At test time, the process of getting document-level translations is two-step (Figure 2):

1. produce translations of isolated sentences using a context-agnostic MT model;

2. apply the DocRepair model to a sequence of context-agnostic translations to correct inconsistencies between translations.

In the scope of the current work, the DocRepair model is the standard sequence-to-sequence Transformer. Sentences in a group are concatenated using a reserved token-separator between sentences.[2] The Transformer is trained to correct these long inconsistent pseudo-sentences into consistent ones. The token-separator is then removed from corrected translations.

## 3 Evaluation of Contextual Phenomena

We use contrastive test sets for evaluation of discourse phenomena for English-Russian by Voita et al. (2019). These test sets allow for testing different kinds of phenomena which, as we show, can

---

[2]In preliminary experiments, we observed that this performs better than concatenating sentences without a separator.

|  | | distance | | |
|  | total | 1 | 2 | 3 |
|---|---|---|---|---|
| **deixis** | 3000 | 1000 | 1000 | 1000 |
| **lex. cohesion** | 2000 | 855 | 630 | 515 |
| **ellipsis (infl.)** | 500 | | | |
| **ellipsis (VP)** | 500 | | | |

Table 1: Size of test sets: total number of test instances and with regard to the distance between sentences requiring consistency (in the number of sentences). For ellipsis, the two sets correspond to whether a model has to predict correct NP inflection, or correct verb sense (VP ellipsis).

be captured from monolingual data with varying success. In this section, we provide test sets statistics and briefly describe the tested phenomena. For more details, the reader is referred to Voita et al. (2019).

## 3.1 Test sets

There are four test sets in the suite. Each test set contains contrastive examples. It is specifically designed to test the ability of a system to adapt to contextual information and handle the phenomenon under consideration. Each test instance consists of a true example (a sequence of sentences and their reference translation from the data) and several contrastive translations which differ from the true one only in one specific aspect. All contrastive translations are correct and plausible translations at the sentence level, and only context reveals the inconsistencies between them. The system is asked to score each candidate translation, and we compute the system accuracy as the proportion of times the true translation is preferred to the contrastive ones. Test set statistics are shown in Table 1. The suites for deixis and lexical cohesion are split into development and test sets, with 500 examples from each used for validation purposes and the rest for testing. Convergence of both consistency scores on these development sets and BLEU score on a general development set are used as early stopping criteria in models training. For ellipsis, there is no dedicated development set, so we evaluate on all the ellipsis data and do not use it for development.

## 3.2 Phenomena overview

**Deixis** Deictic words or phrases, are referential expressions whose denotation depends on context.

This includes personal deixis ("I", "you"), place deixis ("here", "there"), and discourse deixis, where parts of the discourse are referenced ("that's a good question"). The test set examples are all related to person deixis, specifically the T-V distinction between informal and formal you (Latin "tu" and "vos") in the Russian translations, and test for consistency in this respect.

**Ellipsis** Ellipsis is the omission from a clause of one or more words that are nevertheless understood in the context of the remaining elements. In machine translation, elliptical constructions in the source language pose a problem in two situations. First, if the target language does not allow the same types of ellipsis, requiring the elided material to be predicted from context. Second, if the elided material affects the syntax of the sentence. For example, in Russian the grammatical function of a noun phrase, and thus its inflection, may depend on the elided verb, or, conversely, the verb inflection may depend on the elided subject.

There are two different test sets for ellipsis. One contains examples where a morphological form of a noun group in the last sentence can not be understood without context beyond the sentence level ("ellipsis (infl.)" in Table 1). Another includes cases of verb phrase ellipsis in English, which does not exist in Russian, thus requires predicting the verb when translating into Russian ("ellipsis (VP)" in Table 1).

**Lexical cohesion** The test set focuses on reiteration of named entities. Where several translations of a named entity are possible, a model has to prefer consistent translations over inconsistent ones.

## 4 Experimental Setup

### 4.1 Data preprocessing

We use the publicly available OpenSubtitles2018 corpus (Lison et al., 2018) for English and Russian. For a fair comparison with previous work, we train the baseline MT system on the data released by Voita et al. (2019). Namely, our MT system is trained on 6m instances. These are sentence pairs with a relative time overlap of subtitle frames between source and target language subtitles of at least 0.9.

We gathered 30m groups of 4 consecutive sentences as our monolingual data. We used only documents not containing groups of sentences from general development and test sets as well as from contrastive test sets. The main results we report

are for the model trained on all 30m fragments.

We use the tokenization provided by the corpus and use `multi-bleu.perl`[3] on lowercased data to compute BLEU score. We use beam search with a beam of 4.

Sentences were encoded using byte-pair encoding (Sennrich et al., 2016b), with source and target vocabularies of about 32000 tokens. Translation pairs were batched together by approximate sequence length. Each training batch contained a set of translation pairs containing approximately 15000[4] source tokens. It has been shown that Transformer's performance depends heavily on batch size (Popel and Bojar, 2018), and we chose a large batch size to ensure the best performance. In training context-aware models, for early stopping we use both convergence in BLEU score on the general development set and scores on the consistency development sets. After training, we average the 5 latest checkpoints.

### 4.2 Models

The baseline model, the model used for back-translation, and the DocRepair model are all *Transformer base* models (Vaswani et al., 2017). More precisely, the number of layers is $N = 6$ with $h = 8$ parallel attention layers, or heads. The dimensionality of input and output is $d_{model} = 512$, and the inner-layer of a feed-forward networks has dimensionality $d_{ff} = 2048$. We use regularization as described in Vaswani et al. (2017).

As a second baseline, we use the two-pass CADec model (Voita et al., 2019). The first pass produces sentence-level translations. The second pass takes both the first-pass translation and representations of the context sentences as input and returns contextualized translations. CADec requires document-level parallel training data, while DocRepair only needs monolingual training data.

### 4.3 Generating round-trip translations

On the selected 6m instances we train sentence-level translation models in both directions. To create training data for DocRepair, we proceed as follows. The Russian monolingual data is first translated into English, using the Russian→English

| model | BLEU |
|-------|------|
| baseline | 33.91 |
| CADec | 33.86 |
| sentence-level repair | 34.12 |
| DocRepair | **34.60** |

Table 2: BLEU scores. For CADec, the original implementation was used.

model and beam search with beam size of 4. Then, we use the English→Russian model to sample translations with temperature of 0.5. For each sentence, we precompute 20 sampled translations and randomly choose one of them when forming a training minibatch for DocRepair. Also, in training, we replace each token in the input with a random one with the probability of 10%.

### 4.4 Optimizer

As in Vaswani et al. (2017), we use the Adam optimizer (Kingma and Ba, 2015), the parameters are $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\varepsilon = 10^9$. We vary the learning rate over the course of training using the formula:

$$l_{rate} = scale \cdot \min(step\_num^{0.5},$$
$$step\_num \cdot warmup\_steps^{1.5}),$$

where $warmup\_steps = 16000$ and $scale = 4$.

## 5 Results

### 5.1 General results

The BLEU scores are provided in Table 2 (we evaluate translations of 4-sentence fragments). To see which part of the improvement is due to fixing agreement between sentences rather than simply sentence-level post-editing, we train the same repair model at the sentence level. Each sentence in a group is now corrected separately, then they are put back together in a group. One can see that most of the improvement comes from accounting for extra-sentential dependencies. DocRepair outperforms the baseline and CADec by 0.7 BLEU, and its sentence-level repair version by 0.5 BLEU.

### 5.2 Consistency results

Scores on the phenomena test sets are provided in Table 3. For deixis, lexical cohesion and ellipsis (infl.) we see substantial improvements over both the baseline and CADec. The largest improvement over CADec (22.5 percentage points)

---

[3]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/generic

[4]This can be reached by using several of GPUs or by accumulating the gradients for several batches and then making an update.

| model | deixis | lex. c | ell. infl. | ell. VP |
|---|---|---|---|---|
| baseline | 50.0 | 45.9 | 53.0 | 28.4 |
| CADec | 81.6 | 58.1 | 72.2 | **80.0** |
| DocRepair | **91.8** | **80.6** | **86.4** | 75.2 |
| | +10.2 | +22.5 | +14.4 | -4.8 |

Table 3: Results on contrastive test sets for specific contextual phenomena (deixis, lexical consistency, ellipsis (inflection), and VP ellipsis).

(a) **EN** No one believed me. But she did.

　**RU** Мне никто не верил. Но она сказала.

---

(b)**RU** Никто мне не верил. Но она верила.

　**EN** No one believed me. But she believed.

　**RU** Мне никто не верил. Но она поверила.

Figure 3: (a) Example of a discrepancy caused by VP ellipsis: correct meaning is "believe", but MT produces сказала ("say"). (b) Example of producing round-trip translations. From top to bottom: target, first translation, round-trip translation. When translating from Russian, main verbs are unlikely to be translated into auxiliary ones in English, and VP ellipsis is not present.

is for lexical cohesion. However, there is a drop of almost 5 percentage points for VP ellipsis. We hypothesize that this is because it is hard to learn to correct inconsistencies in translations caused by VP ellipsis relying on monolingual data alone. Figure 3(a) shows an example of inconsistency caused by VP ellipsis in English. There is no VP ellipsis in Russian, and when translating auxiliary "did" the model has to guess the main verb. Figure 3(b) shows steps of generating round-trip translations for the target side of the previous example. When translating from Russian, main verbs are unlikely to be translated as the auxiliary "do" in English, and hence the VP ellipsis is rarely present on the English side. This implies the model trained using the round-trip translations will not be exposed to many VP ellipsis examples in training. We discuss this further in Section 6.2.

Table 4 provides scores for deixis and lexical cohesion separately for different distances between sentences requiring consistency. It can be seen, that the performance of DocRepair degrades less than that of CADec when the distance between sentences requiring consistency gets larger.

| | | distance | | |
|---|---|---|---|---|
| | total | 1 | 2 | 3 |
| **deixis** | | | | |
| baseline | 50.0 | 50.0 | 50.0 | 50.0 |
| CADec | 81.6 | 84.6 | 84.4 | 75.9 |
| DocRepair | **91.8** | **94.8** | **93.1** | **87.7** |
| | + 10.2 | +10.2 | +8.7 | +11.8 |
| **lexical cohesion** | | | | |
| baseline | 45.9 | 46.1 | 45.9 | 45.4 |
| CADec | 58.1 | 63.2 | 52.0 | 56.7 |
| DocRepair | **80.6** | **83.0** | **78.5** | **79.4** |
| | + 22.5 | +20.2 | +26.5 | +22.3 |

Table 4: Detailed accuracy on deixis and lexical cohesion test sets.

| all | equal | better | worse |
|---|---|---|---|
| 700 | 367 | 242 | 90 |
| 100% | 52% | 35% | 13% |

Table 5: Human evaluation results, comparing DocRepair with baseline.

### 5.3 Human evaluation

We conduct a human evaluation on random 700 examples from our general test set. We picked only examples where a DocRepair translation is not a full copy of the baseline one.[5]

The annotators were provided an original group of sentences in English and two translations: baseline context-agnostic one and the one corrected by the DocRepair model. Translations were presented in random order with no indication which model they came from. The task is to pick one of the three options: (1) the first translation is better, (2) the second translation is better, (3) the translations are of equal quality. The annotators were asked to avoid the third answer if they are able to give preference to one of the translations. No other guidelines were given.

The results are provided in Table 5. In about 52% of the cases annotators marked translations as having equal quality. Among the cases where one of the translations was marked better than the other, the DocRepair translation was marked better in 73% of the cases. This shows a strong preference of the annotators for corrected translations over the baseline ones.

---

[5]As we further discuss in Section 7, DocRepair does not change the base translation at all in about 20% of the cases.

| | BLEU | deixis | lex. c. | ellipsis |
|---|---|---|---|---|
| 2.5m | 34.15 | 89.2 | 75.5 | 81.8 / 71.6 |
| 5m | 34.44 | 90.3 | 77.7 | 83.6 / 74.0 |
| 30m | **34.60** | **91.8** | **80.6** | **86.4 / 75.2** |

Table 6: Results for DocRepair trained on different amount of data. For ellipsis, we show inflection/VP scores.

| data | deixis | lex. c. | ell. infl. | ell. VP |
|---|---|---|---|---|
| one-way | 85.4 | 63.4 | 79.8 | 73.4 |
| round-trip | 84.0 | 61.7 | 78.4 | 67.8 |

Table 7: Consistency scores for the DocRepair model trained on 2.5m instances, among which 1.5m are parallel instances. Compare round-trip and one-way translations of the parallel part.

| data | BLEU | deixis | lex. c. | ellipsis |
|---|---|---|---|---|
| from mon. | 34.15 | 89.2 | 75.5 | 81.8 / 71.6 |
| from par. | 33.70 | 84.0 | 61.7 | 78.4 / 67.8 |

Table 8: DocRepair trained on 2.5m instances, either randomly chosen from monolingual data or from the part where each utterance in a group has a translation.

## 6 Varying Training Data

In this section, we discuss the influence of the training data chosen for document-level models. In all experiments, we used the DocRepair model.

### 6.1 The amount of training data

Table 6 provides BLEU and consistency scores for the DocRepair model trained on different amount of data. We see that even when using a dataset of moderate size (e.g., 5m fragments) we can achieve performance comparable to the model trained on a large amount of data (30m fragments). Moreover, we notice that deixis scores are less sensitive to the amount of training data than lexical cohesion and ellipsis scores. The reason might be that, as we observed in our previous work (Voita et al., 2019), inconsistencies in translations due to the presence of deictic words and phrases are more frequent in this dataset than other types of inconsistencies. Also, as we show in Section 7, this is the phenomenon the model learns faster in training.

### 6.2 One-way vs round-trip translations

In this section, we discuss the limitations of using only monolingual data to model inconsistencies between sentence-level translations. In Section 5.2 we observed a drop in performance on VP ellipsis for DocRepair compared to CADec, which was trained on parallel data. We hypothesized that this is due to the differences between one-way and round-trip translations, and now we test this hypothesis. To do so, we fix the dataset and vary the way in which the input for DocRepair is generated: round-trip or one-way translations. The latter assumes that document-level data is parallel, and translations are sampled from the source side of the sentences in a group rather than from their back-translations. For parallel data, we take 1.5m parallel instances which were used for CADec training and add 1m instances from our monolingual data. For segments in the parallel

part, we either sample translations from the source side or use round-trip translations. The results are provided in Table 7.

The model trained on one-way translations is slightly better than the one trained on round-trip translations. As expected, VP ellipsis is the hardest phenomena to be captured using round-trip translations, and the DocRepair model trained on one-way translated data gains 6% accuracy on this test set. This shows that the DocRepair model benefits from having access to non-synthetic English data. This results in exposing DocRepair at training time to Russian translations which suffer from the same inconsistencies as the ones it will have to correct at test time.

### 6.3 Filtering: monolingual (no filtering) or parallel

Note that the scores of the DocRepair model trained on 2.5m instances randomly chosen from monolingual data (Table 6) are different from the ones for the model trained on 2.5m instances combined from parallel and monolingual data (Table 7). For convenience, we show these two in Table 8.

The domain, the dataset these two data samples were gathered from, and the way we generated training data for DocRepair (round-trip translations) are all the same. The only difference lies in how the data was filtered. For parallel data, as in the previous work (Voita et al., 2018), we picked only sentence pairs with large relative time overlap of subtitle frames between source-language and target-language subtitles. This is necessary to

ensure the quality of translation data: one needs groups of consecutive sentences in the target language where every sentence has a reliable translation.

Table 8 shows that the quality of the model trained on data which came from the parallel part is worse than the one trained on monolingual data. This indicates that requiring each sentence in a group to have a reliable translation changes the distribution of the data, which might be not beneficial for translation quality and provides extra motivation for using monolingual data.

## 7 Learning Dynamics

Let us now look into how the process of DocRepair training progresses. Figure 4a shows how the BLEU scores with the reference translation and with the baseline context-agnostic translation (i.e. the input for the DocRepair model) are changing during training. First, the model quickly learns to copy baseline translations: the BLEU score with the baseline is very high. Then it gradually learns to change them, which leads to an improvement in BLEU with the reference translation and a drop in BLEU with the baseline. Importantly, the model is reluctant to make changes: the BLEU score between translations of the converged model and the baseline is 82.5. We count the number of changed sentences in every 4-sentence fragment in the test set and plot the histogram in Figure 4b. In over than 20% of the cases the model has not changed base translations at all. In almost 40%, it modified only one sentence and left the remaining 3 sentences unchanged. The model changed more than half sentences in a group in only 14% of the cases. Several examples of the DocRepair translations are shown in Figure 6.

Figure 5 shows how consistency scores are changing in training.[6] For deixis, the model achieves the final quality quite quickly; for the rest, it needs a large number of training steps to converge.

## 8 Related Work

Our work is most closely related to two lines of research: automatic post-editing (APE) and document-level machine translation.
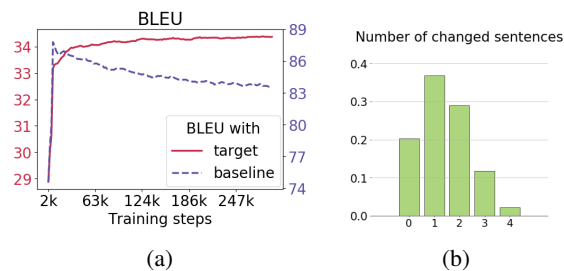


Figure 4: (a) BLEU scores progression in training. BLEU evaluated with the target translations and with the context-agnostic baseline translations (which DocRepair learns to correct). (b) Distribution in the test set of the number of changed sentences in 4-sentence fragments.
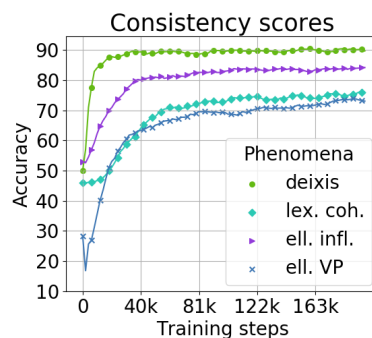


Figure 5: Consistency scores progression in training.

### 8.1 Automatic post-editing

Our model can be regarded as an automatic post-editing system – a system designed to fix systematic MT errors that is decoupled from the main MT system. Automatic post-editing has a long history, including rule-based (Knight and Chander, 1994), statistical (Simard et al., 2007) and neural approaches (Junczys-Dowmunt and Grundkiewicz, 2016; Pal et al., 2016; Freitag et al., 2019).

In terms of architectures, modern approaches use neural sequence-to-sequence models, either multi-source architectures that consider both the original source and the baseline translation (Junczys-Dowmunt and Grundkiewicz, 2016; Pal et al., 2016), or monolingual repair systems, as in Freitag et al. (2019), which is concurrent work to ours. True post-editing datasets are typically small and expensive to create (Specia et al., 2017), hence synthetic training data has been created that uses original monolingual data as output for the sequence-to-sequence model, paired with an automatic back-translation (Sennrich et al., 2016a) and/or round-trip translation as its input(s) (Junczys-Dowmunt and Grundkiewicz, 2016; Fre-

---

[6]Deixis and lexical cohesion scores are evaluated on the development sets which were used in training for the stopping criteria. Ellipsis test sets were not used at training time; the scores are shown here only for visualization purposes.

(a) **EN** Don't you have some souls to save? It's a nice brooch Rita's got. Did you get it from the stolen property, too? How long's it been going on, you and her?

**RU** Разве тебе не нужно спасать души? У Риты красивая брошь. Ты тоже забрал его из украденной собственности? Как долго это продолжается, ты и она?

**RU̅** Разве тебе не нужно спасать души? У Риты красивая брошь. Ты тоже взял ее из украденной собственности? Как долго это продолжается, ты и она?

---

(b) **EN** Y'all working together. Glenn, I've thought for the longest time that you're one of the best in the business. Well, thank you, Rayna. And I love these demos you sent over.

**RU** Вы работаете вместе. Гленн, я долгое время думал, что ты один из лучших в своем деле. Спасибо, Рейна. И мне нравятся эти демо, которые ты прислал.

**RU̅** Вы работаете вместе. Гленн, я долгое время думала, что ты один из лучших в своем деле. Спасибо, Рейна. И мне нравятся эти демо, которые ты прислал.

---

(c) **EN** Who's the new guy? I got to go, all right? Yeah, me too. All right.

**RU** Кто этот новичок? Мне нужно идти, хорошо? Да, я тоже. Хорошо.

**RU̅** Кто этот новичок? Мне нужно идти, хорошо? Да, мне тоже. Хорошо.

Figure 6: Examples of the DocRepair translations. First is the baseline translation, then – corrected by the DocRepair. The differences between translations are underlined. (a) corrected wrong translation of "it": violated gender agreement with the antecedent. (b) corrected wrong gender (marked on a verb): from the third sentence it's clear that the speaker in the second one is feminine (Rayna), but the baseline translation was masculine. (c) corrected wrong morphological form of the pronoun, which was not understood with the elided verb in the third sentence.

itag et al., 2019).

While previous work on automatic post-editing operated on the sentence level, the main novelty of this work is that our DocRepair model operates on groups of sentences and is thus able to fix consistency errors caused by the context-agnostic baseline MT system. We consider this strategy of sentence-level baseline translation and context-aware monolingual repair attractive when parallel document-level data is scarce.

For training, the DocRepair model only requires monolingual document-level data. While we create synthetic training data via round-trip translation similarly to earlier work (Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019), note that we purposefully use sentence-level MT systems for this to create the types of consistency errors that we aim to fix with the context-aware DocRepair model. Not all types of consistency errors that we want to fix emerge from a round-trip translation, so access to parallel document-level data can be useful (Section 6.2).

## 8.2 Document-level NMT

Neural models of MT that go beyond the sentence-level are an active research area (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Agrawal et al., 2018; Miculicich et al., 2018; Kuang et al., 2018; Voita et al., 2019). Typically, the main MT system is modified to take additional context as its input. One limitation of these approaches is that they assume that parallel document-level training data is available.

Closest to our work are two-pass models for document-level NMT (Xiong et al., 2019; Voita et al., 2019), where a second, context-aware model takes the translation and hidden representations of the sentence-level first-pass model as its input. The second-pass model can in principle be trained on a subset of the parallel training data (Voita et al., 2019), somewhat relaxing the assumption that all training data is at the document level.

Our work is different from this previous work in two main respects. Firstly, we show that consistency can be improved with only monolingual document-level training data. Secondly, the

DocRepair model is decoupled from the first-pass MT system, which improves its portability.

# 9 Conclusions

We introduce the first approach to context-aware machine translation using only monolingual document-level data. We propose a monolingual DocRepair model to correct inconsistencies between sentence-level translations. The model performs automatic post-editing on a sequence of sentence-level translations, refining translations of sentences in context of each other. Our approach results in substantial improvements in translation quality as measured by BLEU, targeted contrastive evaluation of several discourse phenomena and human evaluation. Moreover, we perform error analysis and detect which discourse phenomena are hard to capture using only monolingual document-level data. While in the current work we used text fragments of 4 sentences, in future work we would like to consider longer contexts.

# References

Ruchit Agrawal, Turchi Marco, and Negri Matteo. 2018. Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, USA. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. Ape at scale and its implications on mt evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. In *arXiv:1803.05567*. ArXiv: 1803.05567.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. ArXiv: 1704.05135.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representation (ICLR 2015)*.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI '94, pages 779–784, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual*

*Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany. Association for Computational Linguistics.

Martin Popel and Ondrej Bojar. 2018. Training Tips for the Transformer Model. pages 43–70.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, Rochester, New York. Association for Computational Linguistics.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Vivien Macketanz, Inguna Skadiņa, Matteo Negri, , and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Machine Translation Summit XVI*, pages 55–71. Asia-Pacific Association for Machine Translation.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DISCOMT'17, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, Los Angeles.

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

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

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP'17, pages 2816–2821, Denmark, Copenhagen. Association for Computational Linguistics.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7338–7345.