Focused Concatenation for Context-Aware Neural Machine Translation

Lorenzo Lupo, Marco Dinarelli, Laurent Besacier

Focused Concatenation for Context-Aware Neural Machine Translation

Lorenzo Lupo[1] **Marco Dinarelli**[1] **Laurent Besacier**[2]
[1]Université Grenoble Alpes, France
[2]Naver Labs Europe, France
lorenzo.lupo@univ-grenoble-alpes.fr
marco.dinarelli@univ-grenoble-alpes.fr
laurent.besacier@naverlabs.com

Abstract

A straightforward approach to context-aware neural machine translation consists in feeding the standard encoder-decoder architecture with a window of consecutive sentences, formed by the current sentence and a number of sentences from its context concatenated to it. In this work, we propose an improved concatenation approach that encourages the model to focus on the translation of the current sentence, discounting the loss generated by target context. We also propose an additional improvement that strengthen the notion of sentence boundaries and of relative sentence distance, facilitating model compliance to the context-discounted objective. We evaluate our approach with both average-translation quality metrics and contrastive test sets for the translation of inter-sentential discourse phenomena, proving its superiority to the vanilla concatenation approach and other sophisticated context-aware systems.

1 Introduction

While current neural machine translation (NMT) systems have reached close-to-human quality in the translation of decontextualized sentences (Wu et al., 2016), they still have a wide margin of improvement ahead when it comes to translating full documents (Läubli et al., 2018). Many works tried to reduce this margin, proposing various approaches to context-aware NMT (CANMT)[1]. A common taxonomy (Kim et al., 2019; Li et al., 2020) divides them in two broad categories: multi-encoding approaches and concatenation (single-encoding) approaches. Despite its simplicity, the concatenation approaches have been shown to achieve competitive or superior performance to more sophisticated, multi-encoding systems (Lopes et al., 2020; Ma et al., 2021). Nonetheless, it has been shown that Transformer-based NMT systems (Vaswani et al., 2017) struggle to learn locality properties (Hardmeier, 2012; Rizzi, 2013) of both the language itself and the source-target alignment when the input sequence grows in length, as in the case of concatenation (Bao et al., 2021). Unsurprisingly, the presence of context makes learning harder for concatenation models by distracting attention. Moreover, we know from recent literature that NMT systems require context for a sparse set of inter-sentential discourse phenomena only (Voita et al., 2019; Lupo et al., 2022). Therefore, it is desirable to make concatenation models more focused on local linguistic phenomena, belonging to the current sentence, while also processing its context for enabling inter-sentential contextualization whenever it is needed. We propose an improved concatenation approach to CANMT that is more focused on the translation of the current sentence by means of two simple, parameter-free solutions:

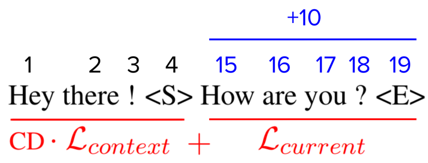

Figure 1: Example of the proposed approach applied over a window of 2 sentences, with context discount CD and segment-shifted positions by a factor of 10.

- Context-discounting: a simple modification of the NMT loss that improves context-aware translation of a sentence by making the model less distracted by its concatenated context;
- Segment-shifted positions: a simple, parameter-free modification of position embeddings, that facilitates the achievement of the context-discounted objective by supporting the learning of locality properties in the document translation task.

We support our solutions with extensive experi-

[1]Unless otherwise specified, we refer to *context* as the sentences that precede or follow a *current* sentence to be translated, within the same document.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 830–842
Abu Dhabi, December 7–8, 2022.

ments, analysis and benchmarking.

2 Background

2.1 Multi-encoding approaches

Multi-encoding models couple a self-standing sentence-level NMT system, with parameters θ_S, with additional parameters θ_C that encode and integrate the context of the current sentence, either on source side, target side, or both. The full context-aware architecture has parameters $\Theta = [\theta_S; \theta_C]$. Multi-encoding models differ from each other in the way they encode the context or integrate its representations with those of the current sentence. For instance, the representations coming from the context encoder can be integrated with the encoding of the current sentence outside the decoder (Maruf et al., 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019; Zheng et al., 2020) or inside the decoder (Tu et al., 2018; Kuang et al., 2018; Bawden et al., 2018; Voita et al., 2019; Tan et al., 2019), by making it attending to the context representations directly, using its internal representation of the decoded history as query.

2.2 Single-encoder approaches

The concatenation approaches are the simplest in terms of architecture, as they mainly consist in concatenating each (current) source sentence with its context before feeding it to the standard encoder-decoder architecture (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019; Agrawal et al., 2018; Ma et al., 2020), without the addition of extra learnable parameters. The decoding can then be limited to the current sentence, although decoding the full target concatenation is more effective thanks to the availability of target context. A typical strategy to train a concatenation approach and generate translations is by sliding windows (Tiedemann and Scherrer, 2017). An sKtoK model decodes the translation $\boldsymbol{y}^j_K$ of a source window $\boldsymbol{x}^j_K$, formed by K consecutive sentences belonging to the same document: the current (jth) sentence and $K-1$ sentences concatenated as source-side context. Besides the end-of-sequence token <E>, another special token <S> is introduced to mark sentence boundaries in the concatenation:

$$\boldsymbol{x}^j_K = \boldsymbol{x}^{j-K+1}\text{<S>}\boldsymbol{x}^{j-K+2}\text{<S>}\dots\text{<S>}\boldsymbol{x}^{j-1}\text{<S>}\boldsymbol{x}^{j}\text{<E>}$$
$$\boldsymbol{y}^j_K = \boldsymbol{y}^{j-K+1}\text{<S>}\boldsymbol{y}^{j-K+2}\text{<S>}\dots\text{<S>}\boldsymbol{y}^{j-1}\text{<S>}\boldsymbol{y}^{j}\text{<E>}$$

Both past and future contexts can be concatenated to the current pair $\boldsymbol{x}^j, \boldsymbol{y}^j$, although in this work we consider only the past context, for simplicity. At training time, the loss is calculated over the whole output $\boldsymbol{y}^j_K$, but only the translation $\boldsymbol{y}^j$ of the current sentence is kept at inference time, while the translation of the context is discarded. Then, the window is slid by one position forward to repeat the process for the $(j+1)$th sentence and its context. Concatenation approaches are trained by optimizing the same objective function as standard NMT over a window of sentences:

$$\mathcal{L}(\boldsymbol{x}^j_K, \boldsymbol{y}^j_K) = \sum_{t=1}^{|\boldsymbol{y}^j_K|} \log P(y^j_{K,t} | \boldsymbol{y}^j_{K,<t}, \boldsymbol{x}^j_K), \quad (1)$$

so that the likelihood of the current target sentence is conditioned on source and target context.

2.3 Closing the gap

Concatenation approaches have the advantage of treating the task of CANMT in the same way as context-agnostic NMT, which eases learning because the learnable parameters responsible for inter-sentential contextualization are the same that undertake intra-sentential contextualization. Indeed, learning the parameters responsible for inter-sentential contextualization in multi-encoding approaches (θ_C) has been shown to be challenging because the training signal is sparse and the task of retrieving useful context elements difficult (Lupo et al., 2022). Nonetheless, encoding current and context sentences together comes at a cost. In fact, when sequences are long the risk of paying attention to irrelevant elements increases. Paying attention to the "wrong tokens" can harm their intra and inter-sentential contextualization, associating them to the wrong latent features. Indeed, Liu et al. (2020) and Sun et al. (2022) showed that learning to translate long sequences, comprised of many sentences, fails without the use of large-scale pre-training or data-augmentation (e.g., like Junczys-Dowmunt (2019) and Ma et al. (2021) did). Bao et al. (2021) provided some evidence about this leaning difficulty, showing that failed models, i.e., models stuck in local minima with a high validation loss, present a distribution of attention weights that is flatter (with higher entropy), both in the encoder and the decoder, than the distribution occurring in models that converge to lower validation loss. In other words, attention struggles to learn the locality properties of both the language itself and the source-target alignment (Hardmeier, 2012; Rizzi, 2013). As a solution, Zhang et al. (2020) and Bao et al. (2021) propose two slightly different masking methods that allow both the encoding of the current sentence concatenated with context, and the separate encoding of each sentence in window. The representations generated by the two encoding schemes are then integrated together, at the cost of adding extra learnable parameters to the standard Transformer architecture.

3 Proposed approach

3.1 Context discounting

Evidently, Equation 1 defines an objective function that does not factor in the fact that we only care about the translation of the current sentence $\boldsymbol{x}^j$, because the context translation will be discarded during inference. Moreover, as discussed above, we need attention to stay focused locally, relying on context only for the disambiguation of relatively sparse inter-sentential discourse phenomena that are ambiguous at sentence level. Hence, we propose to encourage the model to focus on the translation of the current sentence $\boldsymbol{x}^j$ by applying a discount $0 \leq \text{CD} < 1$ to the loss generated by context tokens:

$$\begin{aligned}\mathcal{L}_{\text{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) &= \text{CD}\cdot\mathcal{L}_{context} + \mathcal{L}_{current} \qquad (2)\\ &= \text{CD}\cdot\mathcal{L}(\boldsymbol{x}_{K-1}^{j-1}, \boldsymbol{y}_{K-1}^{j-1}) + \mathcal{L}(\boldsymbol{x}^j, \boldsymbol{y}^j).\end{aligned}$$

This is equivalent to consider an sKtoK concatenation approach as the result of a multi-task sequence-to-sequence setting (Luong et al., 2016), where an sKto1 model performs the *reference task* of translating the current sentence given a concatenation of its source with K-1 context sentences, while the translation of the context sentences is added as a secondary, complementary task. The reference task is assigned a bigger weight than the secondary task in the multi-task composite loss. As we will see in Section 4.5, this simple modification of the loss allows the model to learn a self-attentive mechanism that is less distracted by noisy context information, thus achieving net improvements in the translation of inter-sentential discourse phenomena occurring in the current sentence (Section 4.3), and helping concatenation systems to generalize to wider context after training (Section 4.5.3).

3.2 Segment-shifted positions

Context discounting pushes the model to discriminate between the current sentence and the context. Such discrimination can be undertaken by cross-referencing the information provided by two elements: sentence separation tokens <S>, and sinusoidal position encodings, as defined in (Vaswani et al., 2017). In order to facilitate this task, we propose to provide the model with extra information about sentence boundaries and their relative distance. (Devlin et al., 2019) achieve this goal by adding segment embeddings to every token representation in input to the model, on top of token and position embeddings, such that every segment embedding represents the sentence position in the window of sentences. However, we propose an alternative solution that does not require any extra learnable parameter nor memory allocation: segment-shifted positions. As shown in Figure 1, we apply a constant shift after every separation token <S>, so that the resulting token position is equal to its original position plus a total shift depending on the chosen constant *shift* and the index $k = 1, 2, ..., K$ of the sentence the token belongs to: $t' = t + k * shift$. As a result, the position distance between tokens belonging to different sentences is increased. For example, the distance between the first token of the current sentence and the last token of the preceding context sentence increases from 1 to 1 + *shift*. By increasing the distance between sinusoidal position embeddings[2] of tokens belonging to different sentences, their dot product, which is at the core of the attention mechanism, becomes smaller, possibly resulting in smaller attention weights. In other words, the resulting attention becomes more localized, as confirmed by the empirical analysis reported in Section 4.6.1. In Section 4.3, we present results of segment-shifted positions, and then compare them with both sinusoidal segment embeddings and learned segment embeddings in Section 4.6.2.

4 Experiments

4.1 Setup[3]

We conduct experiments with two language pairs and domains. For En→Ru, we adopt a document-level corpus released by Voita et al. (2019), based on OpenSubtitles2018 (with dev and test sets), comprised of 1.5M parallel sentences. For En→De, we train models on TED talks subtitles released by IWSLT17 (Cettolo et al., 2012). Models are tested

[2]Positions can be shifted by segment also in the case of learned position embeddings, both absolute and relative. We leave such experiments for future works.

[3]See Appendix A for more details.

on IWSLT17's test set 2015, while test-sets 2011-2014 are used for development, following related works in the literature.

Besides evaluating average translation quality with BLEU[4] (Papineni et al., 2002) and COMET[5] (Rei et al., 2020), we employ two contrastive test suites for the evaluation of the translation of inter-sentential discourse phenomena. For En→Ru, we adopt Voita et al. (2019)'s test suite for evaluation on deixis, lexical cohesion, verb-phrase ellipsis and inflection ellipsis. This test suite is comprised of a development set with examples of deixis and lexical cohesion, that we adopted for a preliminary analysis of context discounting. For En→De, we evaluate models on ambiguous pronoun translation with ContraPro (Müller et al., 2018), a large contrastive set of ambiguous pronouns whose antecedents belong to context. In order to validate the improvements achieved by our approaches on the test sets, we perform statistical significance tests, detailed in Annex A.1.

We experiment with two models: 1) **base**: a context-agnostic baseline following *Transformer-base* (Vaswani et al., 2017); 2) **s4to4**: a context-aware concatenation approach with the exact same architecture as *base*, but that adopts sliding windows of 4 concatenated sentences as source and target. An implementation of these models and the proposed approach can be found on github.[6]

4.2 Preliminary analysis

As a preliminary analysis, we evaluate the impact of various values of context discounting on the performance of concatenation approaches with sliding windows, in order to choose one value for all the subsequent experiments. We train En→Ru s4to4 models with context discounts ranging from 1 (no context discounting) to 0 (context loss is completely ignored): $\texttt{CD} = 1.0, 0.9, 0.7, 0.5, 0.3, 0.1, 0.01, 0$. We evaluate these models on the development sets by means of their average loss calculated over the current target sentence (*current loss*) and the average accuracy on the disambiguation of discourse phenomena. The results are plotted on Figure 2. We find out that the stronger the context discounting, the better the performance, with an improving trend from $\texttt{CD} = 1$ to $\texttt{CD} = 0.01$. Performance drops

[4]Moses' *multi-bleu-detok* (Koehn et al., 2007) for De, *multi-bleu* for lowercased Ru as Voita et al. (2019).

[5]Default model: wmt20-comet-da.

[6]https://github.com/lorelupo/focused-concat

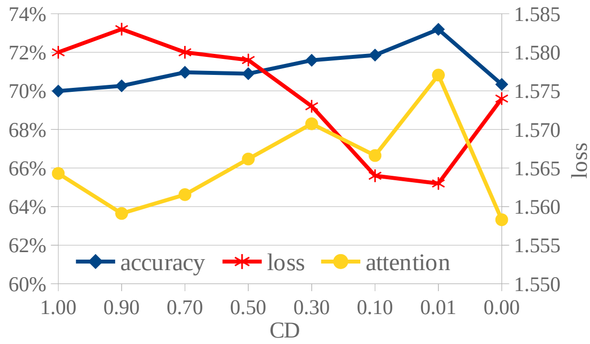

Figure 2: Evaluation of En→Ru s4to4 trained with various levels of context discounting, ranging from 1 to 0. We plot the best *current loss* obtained by each model on the development set (red), and its average accuracy on the development portion of the contrastive set on discourse phenomena (blue). In yellow, the average portion of attention that is focused on the current sentence (see Section 4.5.2).

on the extreme case of $\texttt{CD} = 0$, likely because too much training signal is lost in this situation (all the training signal coming from the context is completely ignored). As such, we set $\texttt{CD} = 0.01$ for all of our following experiments.

4.3 Main results

Tables 1 and 2 display the main evaluation results measured in terms of accuracy on contrastive test sets (Disc.) and BLEU, for the En→Ru and En→De language pairs, respectively. We first observe that s4to4 is a strong context-aware baseline as it improves accuracy on contrastive sets by a large margin compared to the context-agnostic *base*, as already reported by previous works (Voita et al., 2019; Zhang et al., 2020; Lopes et al., 2020).

Average translation quality as measured by BLEU is virtually the same for all models. Indeed, our main focus is on contrastive evaluation of discourse translation, since average translation quality metrics like BLEU have been repeatedly shown to be ill-equipped to detect improvements in CANMT (Hardmeier, 2012). Learned average translation quality metrics like COMET might be more sensitive to inter-sentential discourse phenomena when applied at document-level, as we do. However, COMET differences are also negligible: all models perform on par according to statistical significance tests, except for the En→Ru model with context discount and segment shifting, that outperforms all the others with statistical significance.

When evaluating the accuracy on inter-sentential discourse phenomena, instead, we remark relevant

En→Ru system	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.40***	**73.55***	32.37	0.334*

Table 1: Accuracy on the En→Ru contrastive set for the evaluation of discourse phenomena (Disc., %), and BLEU score on the corresponding test set. The accuracy on Disc. is detailed on its left with the accuracy on each of the 4 discourse phenomena evaluated in the contrastive set. The symbol * denotes statistically significant (p < 0.05) improvements w.r.t. base and s4to4.

En→De system	$d = 1$	$d = 2$	$d = 3$	$d > 3$	Disc.	BLEU	COMET
base	32.89	43.97	47.99	70.58	37.27	29.63	0.546
s4to4	68.89	74.96	79.58	**87.78**	71.35	29.48	0.536
s4to4 + CD	**72.86***	75.96	80.10	84.38	74.31*	29.32	0.522
s4to4 + shift + CD	72.56*	**77.15***	**80.27**	86.65	**74.39***	29.20	0.528

Table 2: Accuracy on the En→De contrastive sets for the evaluation of discourse phenomena (Disc., %), and BLEU score on the corresponding test sets. The accuracy on Disc. is detailed on its left with the accuracy on anaphoric pronouns with antecedents at different distances $d = 1, 2, ...$ (in number of sentences). The symbol * denotes statistically significant (p < 0.05) improvements w.r.t. base and s4to4.

performance improvements. In fact, adding a 0.01 context discounting (+ CD) improves the accuracy on all of the 4 discourse phenomena under evaluation in En→Ru, and for all distances of pronoun's antecedents in En→De, with the sole exception of $d > 3$, proving to be an effective solution. Adding segment-shifted positions further improves performance for 3 discourse phenomena out of 4, and for pronouns with antecedents at distances $d = 1, 2$, showing that sliding windows systems often benefit from enhanced sentence position information in order to achieve the discounted CANMT objective. For both language pairs, we adopt a segment-shifting equal to the average sentence length, calculated over the entire training corpus, i.e., +8 positions for En→Ru and +21 positions for En→De. Experiments with other shifting values are reported in Section 4.6.3.

As a further experiment, we apply our solutions to concatenation models with concatenated windows shorter than 4 sentences,[7] and evaluate them in the En→Ru setting. The results presented in Table 3 show that context discounting is effective for s2to2 and s3to3 too, while adding segment-shifted positions only helps s2to2 + CD. As in the case of s4to4, BLEU only displays negligible fluctuations.

[7]We cannot evaluate with more sentences because 4 is the maximum size of documents in the test sets specialized on discourse phenomena.

System	Disc.	BLEU
s2to2	59.10	32.73
s2to2 + CD	60.28*	32.69
s2to2 + shift + CD	**60.54***	32.41
s3to3	65.58	32.34
s3to3 + CD	**67.02***	32.42
s3to3 + shift + CD	66.98*	32.45

Table 3: Accuracy on the En→Ru contrastive set for the evaluation of discourse phenomena (Disc., %), and BLEU score on the test set. The symbol * denotes statistically significant (p < 0.05) improvements w.r.t. s2to2/s3to3. Our approach is effective for different concatenation windows.

4.4 Benchmarking

For a wider contextualization of our results, we compare in Table 4 our best system with other CANMT systems from the literature. For the En→Ru language pair, we compare with all the systems from the literature that were trained and evaluated under the same experimental conditions as ours, to the best of our knowledge. In particular, we report the results by Chen et al. (2021), Sun et al. (2022)' *MR Doc2Doc*, Zheng et al. (2020), Kang et al. (2020)'s *CADec + DCS-pf* and Zhang et al. (2020). All of them are sophisticated CANMT systems that add extra trainable parameters to the

System			En→Ru					En→De		
	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	d=1	d=2	d=3	d>3	Disc.
Chen et al. (2021)	62.30	47.90	64.90	36.00	55.61	n.a.	n.a.	n.a.	n.a.	n.a.
Sun et al. (2022)	64.70	46.30	65.90	53.00	58.13	n.a.	n.a.	n.a.	n.a.	n.a.
Zheng et al. (2020)	61.30	58.10	72.20	80.00	63.30	n.a.	n.a.	n.a.	n.a.	n.a.
Kang et al. (2020)	79.20	62.00	71.80	80.80	73.46	n.a.	n.a.	n.a.	n.a.	n.a.
Zhang et al. (2020)	**91.00**	46.90	78.20	**82.20**	**75.61**	n.a.	n.a.	n.a.	n.a.	n.a.
Maruf et al. (2019)	n.a.	n.a.	n.a.	n.a.	n.a.	34.70	46.40	51.10	70.10	39.15
Voita et al. (2018)	n.a.	n.a.	n.a.	n.a.	n.a.	39.00	48.00	54.00	66.00	42.55
Stojanovski and Fraser (2019)	n.a.	n.a.	n.a.	n.a.	n.a.	53.00	46.00	50.00	71.00	52.55
Lupo et al. (2022)	n.a.	n.a.	n.a.	n.a.	n.a.	56.50	44.90	48.70	73.30	54.98
Müller et al. (2018)	n.a.	n.a.	n.a.	n.a.	n.a.	58.00	55.00	55.00	75.00	58.13
s4to4 + shift + CD (ours)	85.76	**48.33**	**81.40**	80.40	73.56	**72.56**	**77.15**	**80.27**	**86.65**	**74.39**

Table 4: Benchmarking: accuracy (%) on the contrastive sets for the evaluation of discourse phenomena (Disc., %).

Transformer architecture. Despite being the simplest and the only parameter free approach, our method outperforms all the others on lexical cohesion and noun phrase inflection based on elided context, while it is only second to Zhang et al. (2020) on deixis and verb-phrase ellipsis. BLEU scores were not available for comparison on the same test set, except for Zhang et al. (2020), which scored 31.84 BLEU points against the 32.45 BLEU points of our method.

For the En→De language pair, we compare to the literature performing evaluation on Müller et al. (2018)'s test set and providing details about their accuracy on pronouns with antecedents at $d > 1$. In particular: Maruf et al. (2019)'s best offline system, Stojanovski and Fraser (2019)'s *pron-25→pron-0**, Lupo et al. (2022)'s *K1-d&r*, Müller et al. (2018)'s *s-hier-to-2.tied* and their evaluation of Voita et al. (2018)'s architecture.[8] All of these works but Maruf et al. (2019) adopt the much larger WMT17[9] dataset for training. Despite this advantage, our system outperforms each of them on all the discourse phenomena under evaluation, by a large margin.

Notably, from this comparison it might seem that our approach is proposed in opposition to the others reported in Table 4, but it can actually be complimentary to many of them, such as (Zhang et al., 2020)'s, hopefully in a synergistic way. We encourage future research to investigate this possibility.

[8]Whenever the cited works present and evaluate multiple systems, we compare to the best performing one. To the best of our knowledge, we are including all the relevant works available in the literature. BLEU scores are not compared because, besides using different training data, the cited works don't adopt the same test set neither, with the sole exception of (Lupo et al., 2022).

[9]http://www.statmt.org/wmt17/translation-task.html

4.5 Analysis of context-discounting

4.5.1 Loss distribution

In this section, we analyze the impact of context discounting on the ability of the model to predict the translation of the current sentence. On the left side of Figure 3 we plotted the evolution along training epochs of the loss calculated on the current target sentence (*current loss*), for the En→Ru language pair. The right side, instead, represents the ratio between the *current loss* and the average loss-per-sentence calculated on the context sentences belonging to the same sliding window. These results support empirically our idea of context discounting as a solution to improve model performance on the current sentence. They also confirm that a strong discounting works best. Interestingly, predictions are improved on the current sentence (left) partially as a result of a trade-off with context quality (right). In fact, the current/context loss ratio of context-discounted models increases along training even when the *current loss* is decreasing, indicating that, at the beginning of training, context discounting pushes the model to only care about current predictions, but later it allows for good predictions of the context too. Such behavior is in line with the intuition that a good translation of the current sentence, even if strongly prioritized, also requires a good translation of the context. Otherwise, it is not possible to systematically solve the translation ambiguities referring to context.

4.5.2 Attention distribution

In this section we show some empirical evidence in favor of our intuition that context-discounting improves performance by helping the self-attentive mechanism to be more focused on the current sentence (less distracted by context). We analyzed

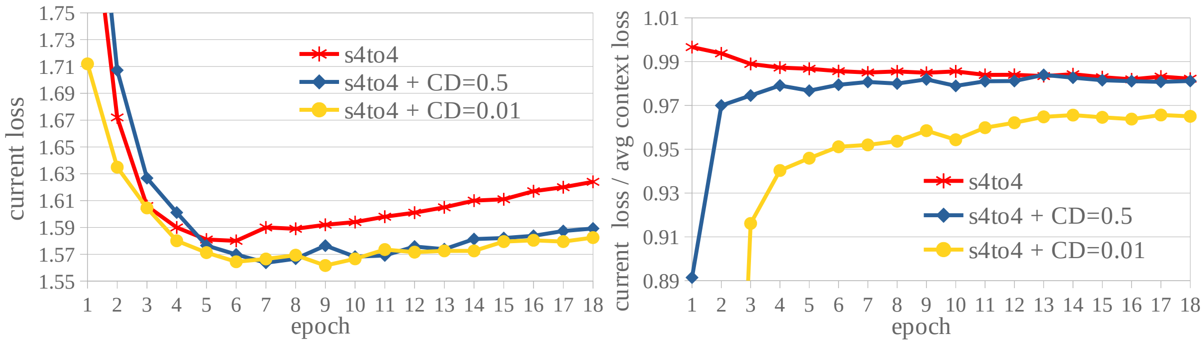

Figure 3: Context discounting enables better predictions of the current sentence (lower validation loss, on the left) at the expense of context sentences (lower current/context validation loss ratio, on the right). Language pair: En→Ru.

the distribution of the self-attention weights generated by the queries belonging to the current sentence (*current queries*), and how it is impacted by context discounting. Figure 2 clearly shows that context-discounting impacts the distribution of attention weights by skewing it towards the current sentence: a higher percentage of the total attention from *current queries* is directed towards tokens belonging to the (same) current sentence. As expected, the higher the context-discounting, the higher the portion of attention that is not dispersed towards context. The limit case of $CD = 0$ is not aligned with this trend, however. We suspect that the attention distribution is more flat in this case because the model encounters learning difficulties due to the training signal from the context being completely ignored (c.f. Bao et al. (2021) on non-fully-converged models having a flatter attention distribution).

4.5.3 Robustness

Figure 4 shows that the s2to2 model is not robust to the translation of concatenation windows longer than those seen during training, i.e. longer than 2 sentences. Indeed, s2to2 loses 9.23 BLEU points when translating the same test set with windows of 3 sentences, and 12.14 BLEU points when translating with windows of 4. Instead, the context discounted model (blue bars) is very robust to unseen context lengths, being capable of translating them with minor degradation in average translation quality (-0.68 and -1.06 BLEU points for windows of 3 and 4, respectively). We observe a similar trend for s3to3, that loses 1.74 BLEU points when tested with windows of size 4, but recovers completely when equipped with context-discounting. The increased robustness of the concatenation models

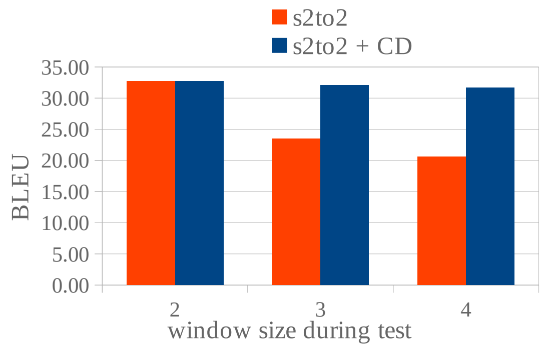

Figure 4: Our approach improves robustness of En→Ru s2to2 to window sizes unseen during training.

w.r.t. context size suggests once again that context discounting helps the models focusing on the current sentence.

4.6 Analysis of segment-shifted positions

4.6.1 Attention distribution

As a complementary evaluation, we tested if segment-shifted positions work as intended, i.e., by helping context-discounted models to learn the locality properties of both the language itself and the source-target alignment (Hardmeier, 2012; Rizzi, 2013). In other words, we expect segment-shifted positions to result in a more localized attention-distribution, in each of the sentences belonging to the concatenated sequence. To this aim, we computed the average entropy of the distribution of attention weights generated by all queries (both from current and context sentences), in both self and cross-attention. Results are shown in Table 5: context-discounting slightly reduces the average entropy, and this effect is amplified with the adoption of segment-shifted positions. Segment-shifted positions make attention more focused locally, as intended, which explains why the job of context

System	Attn entropy
s4to4	2.293
s4to4 + CD	2.276
s4to4 + shift + CD	**2.251**

Table 5: Average entropy of self and cross-attention weights decreases with the help of context-discounting and segment-shifted positions. All of the three values are different from one another with statistical significance ($p<0.01$).

	En→Ru		En→De	
System	Disc.	BLEU	Disc.	BLEU
s4to4 + shift + CD	73.56	32.45	**74.39**	29.20
s4to4 + lrn + CD	**73.68**	32.45	72.14	28.35
s4to4 + sin + CD	73.48	32.53	73.88	29.23

Table 6: Comparison between segment-shifted positions, learned segment embeddings and sinusoidal segment embeddings. Approaches are evaluated with accuracy on contrastive sets for the evaluation of discourse phenomena (Disc., %), and BLEU score on test sets. Differences across models are not statistically significant ($p>0.05$), except for s4to4+lrn+CD on En→De.

discounting is eased by this solution.

4.6.2 Comparison with segment-embeddings

In this section we compare our parameter-free approach to include explicit information on segment position (segment-shifted positions), with learned segment embeddings (Devlin et al., 2019), and sinusoidal segment embeddings. The latter are added to token and position embeddings at input, in the very same way as learned segment embeddings, with the only difference that their parameters are not learned but defined in the same way as sinusoidal position embeddings (Vaswani et al., 2017). In order to evaluate which approach helps best with context-discounting, we trained a context-discounted concatenation model with learned segment embeddings (s4to4+lrn+CD), and one with sinusoidal segment embeddings (s4to4+sin+CD), and compared them with s4to4+shift+CD. The results reported in Table 6 do not display any statistically significant differences across the three alternatives ($p>0.05$), except for learned embeddings, that underperform with statistical significance the other two variants on En→De. Instead, sinusoidal segment embeddings are competitive with segment-shifted positions on both language pairs. We leave a more in-depth analysis of segment-embeddings for concatenation approaches to future works.

System	Shift	Disc.	BLEU
s4to4 + shift + CD	100.00	73.46	32.41
s4to4 + shift + CD	avg-sequence	**73.86**	32.37
s4to4 + shift + CD	avg-corpus	73.56	32.45

Table 7: Accuracy on the En→Ru contrastive set for the evaluation of discourse phenomena (Disc., %), and BLEU score on the test set. Differences across models are not statistically significant ($p>0.05$).

4.6.3 Segment-shifting variants

In the experiments reported above, we always adopt a shifting value equal to the average sentence length calculated over the entire training corpus (avg-corpus), i.e., +8 positions for En→Ru, +21 positions for En→De. In this section we evaluate two alternative strategies for the selection of the shifting value: 1) applying a big shift of 100 units, one order of magnitude bigger than the average sentence length in the corpus (100); 2) applying a shifting value equal to the average sentence length of each window, calculated dynamically for each window of 4 concatenated sentences (avg-sequence). The results of this study are reported in Table 7. We do not observe relevant differences in average translation quality (BLEU) nor accuracy on the translation of discourse phenomena, and therefore confirm that the avg-corpus approach is a good alternative.

5 Conclusions

We presented a simple, parameter-free modification of the NMT objective for context-aware translation with sliding windows of concatenated sentences: context discounting. We analyzed the impact of our approach in the trade-off between current sentence predictions and context sentence predictions, showing that context discounting helps the model to focus on the current sentence, as intended. As a result, the concatenation model significantly improves its ability to disambiguate inter-sentential discourse phenomena, and becomes more robust to different context sizes. As an additional inductive bias towards locality, we equipped our model with segment-shifted positions, marking more explicitly the boundaries between sentences. This solution brings further improvements on targeted evaluation metrics. In the attempt of explaining the empirical functioning of the proposed solutions, we analysed their impact on the distribution of the attention weights, showing that they make it more focused and skewed towards the current sentence, as intended.

Limitations and future works

Our experiments are limited to the use case of short concatenated windows (up to 4 sentences). This is enough for capturing most of the ambiguous inter-sentential discourse phenomena, that usually span across a few sentences only (Müller et al., 2018; Voita et al., 2019; Lupo et al., 2022). However, recent works suggest that longer context windows might be helpful to increase average translation quality (BLEU) of concatenation approaches (Junczys-Dowmunt, 2019; Bao et al., 2021; Sun et al., 2022), and long-range discourse phenomena could be handled. We hope to investigate the impact of context discounting on longer sequences in future works. We also encourage to test the effectiveness of our approach on a wider range of data scenarios: from very limited document-level data to very abundant, including back translation (Ma et al., 2021) and monolingual pre-training techniques (Junczys-Dowmunt, 2019; Sun et al., 2022), to understand whether these methods are only alternative to context discounting or there exist synergies. Furthermore, experimenting with future context is also needed (c.f. Wong et al. (2020)).

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This work has been partially supported by the Multidisciplinary Institute in Artificial Intelligence MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

Ruchit Rajeshkumar Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20, Alacant, Spain.

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

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 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.

Linqing Chen, Junhui Li, Zhengxian Gong, Boxing Chen, Weihua Luo, Min Zhang, and Guodong Zhou. 2021. Breaking the corpus bottleneck for context-aware neural machine translation with cross-task pre-training. 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 2851–2861, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Hardmeier. 2012. Discourse in Statistical Machine Translation. A Survey and a Case Study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11). 00039 Number: 11 Publisher: Presses universitaires de Caen.

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

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the*

2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

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

Zhiyi Ma, Sergey Edunov, and Michael Auli. 2021. A Comparison of Approaches to Document-level Machine Translation. *arXiv:2101.11040 [cs]*. ArXiv: 2101.11040.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. 03511.

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

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. *arXiv:1701.06548 [cs]*. 00464 arXiv: 1701.06548.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70. 00112 arXiv: 1804.00247.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Luigi Rizzi. 2013. Locality. *Lingua*, 130:169–186.

Dario Stojanovski and Alexander Fraser. 2019. Improving anaphora resolution in neural machine translation using curriculum learning. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 140–150, Dublin, Ireland. European Association for Machine Translation.

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

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 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.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 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.

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.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*. 00000 arXiv: 1609.08144.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3983–3989. ijcai.org.

A Details on experimental setup

All models are implemented in *fairseq* (Ott et al., 2019) and follow the *Transformer-base* architecture (Vaswani et al., 2017): hidden size of 512, feed forward size of 2048, 6 layers, 8 attention heads, total 60.7M parameters. They are trained on 4 Tesla V100, with a fixed batch size of approximately 32k tokens for En→Ru and 16k for En→De. As it has been shown that Transformers need a large batch size for achieving the best performance (Popel and Bojar, 2018). We stop training after 12 consecutive non-improving validation steps (in terms of loss on dev), and we average the weights of the 5 checkpoints that are closest to the best performing checkpoint, included. We train models with the optimizer configuration and learning rate (LR) schedule described in Vaswani et al. (2017). The maximum LR is optimized for each model over the search space $\{7e-4, 9e-4, 1e-3, 3e-3\}$. The LR achieving the best loss on the validation set after convergence was selected. We use label smoothing with an epsilon value of 0.1 (Pereyra et al., 2017) for all settings. We adopt strong model regularization (dropout=0.3) following Kim et al. (2019) and Ma et al. (2021). At inference time, we use beam search with a beam of 4 for all models. We adopt a length penalty 0.6 for all models. The other hyperparameters were set according to the relevant literature (Vaswani et al., 2017; Popel and Bojar, 2018; Voita et al., 2019; Ma et al., 2021; Lopes et al., 2020).

A.1 Statistical hypothesis tests

We perform statistical hypothesis testing with McNemar's test McNemar (1947) for comparing accuracy results on the contrastive test sets. For comparing BLEU performances and mean entropy (Table 5), we use approximate randomization (Riezler and Maxwell, 2005) with 10000 and 1000 permutations, respectively. For COMET, the official library[10] has a built in tool for the calculation of statistical significance with Paired T-Test and bootstrap resampling (Koehn, 2004).

[10]https://github.com/Unbabel/COMET

B Details on experimental results

In this section, we report more details about the results presented in our Tables.

B.1 Evaluation of the translation of discourse phenomena

For each model that we evaluated by its accuracy on the contrastive sets for the evaluation of discourse phenomena (Disc., %), we include in Table 8 the accuracy achieved on the different subsets of the contrastive sets, as already done for Tables 1, 2 and 4. For the En→Ru set (Voita et al., 2019), we report the accuracy on each of the 4 discourse phenomena under evaluation; for the En→De test set (Müller et al., 2018), the accuracy on anaphoric pronouns with antecedents at different distances $d = 1, 2, ...$ (in number of sentences). As it can be noticed, our approach mostly outperform baselines and other variants on the majority of the evaluation subsets. We also include the column Disc$_{\text{avg}}$, which is calculated, for both language pairs, as the average of the 4 columns before the vertical dashed line.

$$Disc. = \frac{d1 * 7075 + d2 * 1510 + d3 * 573 + (d > 3) * 442}{9600},$$
$$Disc_{\text{avg}} = \frac{d1 + d2 + d3 + d > 3}{4}.$$

Disc$_{\text{avg}}$ represents the average accuracy on the disambiguation of the discourse phenomena present in the contrastive sets, as if they were all present with the same frequency. Instead, Disc. represents the overall accuracy on the contrastive set, which is equivalent to the average over the same 4 columns, but weighted by the sample size (last row) of each penomenon represented by the columns. While Disc. is a proxy of the ability to correctly translate a distribution of inter-sentential discourse phenomena as represented in the contrastive set, Disc$_{\text{avg}}$ is a proxy for the average ability to translate each of the inter-sentential phenomena under evaluation. Interestingly, Disc$_{\text{avg}}$ captures more evidently than Disc. the improvement achieved by adding segment-shifted positions to the context-discounted concatenation models. Finally, Disc$_{all-d}$ is calculated like Disc. but it also take into account pronouns whose antecedent belong to the same sentence ($d = 0$, i.e., they don't require context).

	En→Ru						En→De							
System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	Disc_{avg}	d=0	d=1	d=2	d=3	d>3	Disc.	Disc_{avg}	Disc_{all-d}
base	50.00	45.87	51.80	27.00	46.64	47.67	68.75	32.89	43.97	47.99	70.58	37.27	48.86	43.57
s4to4	85.80	46.13	79.60	73.20	72.02	71.18	75.20	68.89	74.96	79.58	**87.78**	71.35	77.80	72.12
s4to4 + CD	**87.16**	46.40	81.00	78.20	73.42	73.19	**76.66**	**72.86**	75.96	80.10	84.38	74.31	78.33	**74.78**
s4to4 + shift + CD	85.76	**48.33**	**81.40**	**80.40**	**73.56**	**73.97**	75.25	72.56	**77.15**	**80.27**	86.65	**74.39**	**79.16**	74.56
s4to4 + sin + CD	87.96	**46.80**	78.00	**76.60**	73.48	72.34	**76.75**	**71.83**	**76.82**	**80.97**	**87.55**	**73.88**	**79.29**	**74.46**
s4to4 + lrn + CD	**88.12**	46.47	**81.20**	75.60	**73.68**	**72.85**	73.91	70.21	75.29	77.66	85.06	72.14	77.06	72.49
s4to4 + 100 + CD	**85.60**	**48.73**	**80.80**	**79.60**	**73.46**	**73.68**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s4to4 + avg-seq + CD	84.84	46.20	77.60	73.00	71.34	70.41	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s2to2	61.84	46.07	74.60	69.00	59.10	62.88	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s2to2 + CD	**62.88**	46.27	78.00	**71.60**	60.28	64.69	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s2to2 + shift + CD	62.60	**46.60**	**81.20**	71.40	**60.54**	**65.45**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s3to3	73.52	45.87	78.00	72.60	65.58	66.45	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s3to3 + CD	73.88	**46.80**	**82.40**	**78.00**	**67.02**	**67.45**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
s3to3 + shift + CD	**75.24**	46.07	79.40	76.00	66.98	68.45	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Chen et al. (2021)	62.30	47.90	64.90	36.00	55.61	52.78	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Sun et al. (2022)	64.70	46.30	65.90	53.00	58.13	57.48	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Zheng et al. (2020)	61.30	58.10	72.20	80.00	63.30	67.90	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Kang et al. (2020)	79.20	62.00	71.80	80.80	73.46	73.45	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Zhang et al. (2020)	**91.00**	46.90	78.20	**82.20**	**75.61**	**74.58**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
(Maruf et al., 2019)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	68.60	34.70	46.40	51.10	70.10	39.15	50.58	45.04
(Müller et al., 2018)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	75.00	39.00	48.00	54.00	66.00	42.55	51.75	49.04
(Stojanovski and Fraser, 2019)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	74.00	53.00	46.00	50.00	71.00	52.55	55.00	56.84
(Lupo et al., 2022)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	**81.10**	56.50	44.90	48.70	73.30	54.98	55.85	**60.21**
(Müller et al., 2018)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	65.00	**58.00**	**55.00**	**55.00**	**75.00**	**58.13**	**60.75**	59.51
Sample size	2500	1500	500	500	5000	5000	2400	7075	1510	573	442	9600	9600	12000

Table 8: Accuracy on contrastive sets for the evaluation of discourse phenomena (Disc., %) and on their subsets: for En→Ru, the accuracy on each of the 4 discourse phenomena under evaluation; for En→De, the accuracy on anaphoric pronouns with antecedents at different distances $d = 1, 2, ...$ (in number of sentences). Disc_{all-d}, includes also $d = 0$. Disc_{avg} denotes the average of the 4 accuracies before the dashed line.