Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models

Anonymous ACL submission

Abstract

Multi-encoder models are a broad family of context-aware neural machine translation systems that aims to improve translation quality by encoding document-level contextual information alongside the current sentence. The context encoding is undertaken by *contextual parameters*, trained on document-level data. In this work, we discuss the difficulty of training these parameters effectively, due to the sparsity of the words in need of context (i.e., the training signal), and their relevant context. We propose to pre-train the contextual parameters over split sentence pairs, which makes an efficient use of the available data for two reasons. Firstly, it increases the contextual training signal by breaking intra-sentential syntactic relations, and thus pushing the model to search the context for disambiguating clues more frequently. Secondly, it eases the retrieval of relevant context, since context segments become shorter. We propose four different splitting methods, and evaluate our approach with BLEU and contrastive test sets. Results show that it consistently improves learning of contextual parameters, both in low and high resource settings.

1 Introduction

Neural machine translation (NMT) has seen substantial improvements in recent years, fostered by the advent of the Transformer model (Vaswani et al., 2017). A remaining challenge for modern machine translation (MT) is the ability to contextualize translation of the current sentence with the other sentences in the document (Läubli et al., 2018). For this reason, contextual NMT has recently triggered a lot of attention and many approaches have been proposed in the literature. A common taxonomy (Kim et al., 2019; Li et al., 2020) divides them in two broad categories: single-encoder (concatenation) approaches (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Ma et al., 2020; Zhang et al., 2020) and multi-encoder approaches (Jean et al., 2017; Tu et al., 2017; Bawden et al., 2018; Miculicich et al., 2018; Voita et al., 2018; Maruf et al., 2019a; Zheng et al., 2020). Multi-encoder models are more flexible and can be more efficient than concatenation approaches, but they have been criticized as being mere regularization methods (Kim et al., 2019; Li et al., 2020). In some cases, they have even been shown to perform worse than sentence-level systems on discourse-aware targeted test suites (Lopes et al., 2020).

In this work, we address this criticism by showing that training multi-encoder models is difficult because of two reasons: (i) the sparsity of *contextual training signal*, i.e. the signal that pushes systems to translate in a context-aware fashion, which comes from the words that need context to be correctly translated; (ii) the sparsity of relevant context words, the ones needed to disambiguate translation. A trivial way to improve context-aware learning is by increasing the amount of document-level training data. Large document-level parallel corpora are not always available, but some works have proposed data augmentation techniques to remedy this lack (Sugiyama and Yoshinaga, 2019; Stojanovski et al., 2020; Huo et al., 2020). However, as we will show in our experimental section, this solution is not efficient and often sub-optimal. We therefore introduce a novel pre-training strategy, *divide and rule (d&r)*, that is based on a simple and yet powerful technique to augment the contextual training signal and to ease learning efficiently: splitting parallel sentences in segments (see Figure 1). Simply put, feeding a context-aware model with a sequence of incomplete, shorter, consecutive segments, forces it to look for context (i.e., surrounding segments) more frequently, and makes it easier to retrieve relevant context because segments are shorter. This results in faster and improved learning. We pre-train multi-encoder models on split datasets and evaluate them in two ways: BLEU score, and

$S^{i,1}$ He said that it was a project of peace
$S^{i,2}$ and unity and that it brought people together .
$T^{i,1}$ *Il disait que c' était un projet de paix*
$T^{i,2}$ *et d' unité et qu' il réunissait les gens .*

$S^{j,1}$ I think single-cell organisms **are**
$S^{j,2}$ possible within two years .
$T^{j,1}$ *Je pense que les organismes unicellulaires*
$T^{j,2}$ ***sont** possibles dans 2 ans .*

Figure 1: Example of sentence pairs from En→Fr IWSLT17, after being tokenized and split in the middle. After the splitting, some syntactic relations span across two segments (underlined). Also, some source-side words are not parallel with their reference (**in bold**).

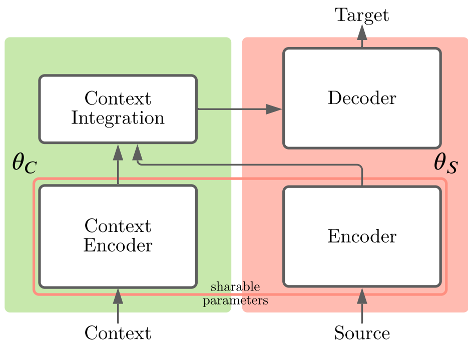

Figure 2: Multi-encoder approach integrating context outside the decoder.

contrastive test sets for discourse phenomena.

Our main contributions are the following: (i) we show that context-aware multi-encoder models need to be trained carefully, because the contextual training signal is sparse, as well as the context elements useful for contextualization; (ii) we propose the *d&r* pre-training strategy, which fosters training of contextual parameters by splitting sentences into segments, with four splitting variants; (iii) we support this strategy with an analysis of the impact of splitting on the distribution of discourse phenomena; (iv) we demonstrate that this strategy is both effective and efficient, as it allows multi-encoder models to learn better and faster than by simply increasing the training data.

2 Background

2.1 Single-encoder approaches

The most straightforward approach to context-aware NMT consists in concatenating the context to the current sentence before feeding it to the standard encoder-decoder architecture (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019; Ma et al., 2020; Zhang et al., 2020). A special token is introduced to mark the boundaries between sentences. Generation can then follow two strategies: the *many-to-many* strategy consists in translating all the source sentences, and then discarding contextual sentences; the *many-to-one* strategy consists in translating the current sentence only. The modeling capacity of concatenation methods is limited to few sentences because the complexity of attention scales quadratically with sentence length, although some recent works try to solve this constraint (Tay et al., 2020).

2.2 Multi-encoder approaches

Multi-encoder models couple a self-standing sentence-level NMT system, with parameters θ_S, with additional parameters for modeling the context either on source side, target side, or both. We refer to these parameters as the *contextual parameters* θ_C. The full context-aware architecture has parameters $\Theta = [\theta_S; \theta_C]$. Most of the multi-encoder models can be described as instances of two architectural families (Kim et al., 2019), that only differ in the way the encoded representations of the context and the current sentence are integrated.

Outside integration. In this approach, depicted in Figure 2, the encoded representations are merged outside the decoder (Maruf et al., 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019a; Zheng et al., 2020). This can happen in different ways, such as by simple concatenation of the encodings, or with a gated sum.

Inside integration. Here the decoder attends to the context representations directly, using its internal representation of the decoded history as query (Tu et al., 2018; Kuang et al., 2018; Bawden et al., 2018; Voita et al., 2019b; Tan et al., 2019).

Many of these works found useful to share parameters of current-sentence and context encoders (Voita et al., 2018; Li et al., 2020). In this way, the amount of contextual parameters to learn, $|\theta_C|$, and the computational cost are drastically reduced. Shared representation can also be cached to be re-used and further processed by contextual parameters without the need of re-encoding sentences from scratch, which represents an advantage with respect to single-encoder approaches. Most of the approaches proposed in the literature focus on a few previous sentences, where most of the relevant context is concentrated.

Two-step training. Multi-encoder models are commonly trained following a two-step strategy (Tu et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Li et al., 2020). The first step consists in training θ_S independently on a sentence-level parallel corpus $\mathcal{C}_S$. Secondarily, contextual parameters θ_C are trained on a document-level parallel corpus $\mathcal{C}_D$, while fine-tuning or freezing θ_S. Note that $\mathcal{C}_S$ can also include sentences from $\mathcal{C}_D$.

2.3 Targeted evaluation

Most proponents of novel NMT systems evaluate them by computing BLEU (Papineni et al., 2002) on the test data. However, BLEU is ill-equipped to capture the improvements achieved by context-aware MT (Hardmeier, 2012), because contextualization can improve the translation of only a small fraction of the words in a document, while most of the words can be correctly translated without knowing the context. For instance, only a fraction of the anaphoric pronouns in a document has its nominal antecedent outside its own sentence. However, despite being sparse, these few cases strongly impact the quality of translation (Läubli et al., 2018; Popescu-Belis, 2019). Consequently, a number of discourse-targeted test sets and automatic metrics have been proposed to measure improvements in context-aware MT (Maruf et al., 2019b), the most widely adopted ones being contrastive test sets.

Contrastive test sets (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019a) consist of a number of source sentences, each paired with a correct translation and some incorrect ones. Models are assessed on their ability to rank first the correct translation. In many cases, this can be identified only by looking at context, which is provided for both source and target sides. Therefore, the ranking accuracy reflects the context-modeling ability of the evaluated translation system.

3 The double challenge of sparsity

Some works criticized multi-encoder methods (Kim et al., 2019; Li et al., 2020), arguing that they do not improve sentence-level baselines in terms of BLEU when the baseline is well regularized. When there are improvements, it is argued that the context-encoder simply works as a noise-generator that makes training more robust, and the improvements are not due to better context-modeling. Along this path, Lopes et al. (2020) showed that multi-encoder architectures struggle to model contextual information, and even deteriorate the performance of a sentence-level baseline on contrastive test sets. In fact, many proponents of multi-encoder models only show BLEU improvements, without providing any kind of targeted evaluation. This doesn't allow a direct evaluation of their context-modeling capability. We posit that training the contextual parameters of multi-encoder models is non-trivial because of two challenges: (i) the sparsity of the training signal, which comes from the words that need context to be correctly translated (most of the words of a sentence can be translated without context); (ii) the sparsity of context words that are useful for contextualization (most of the context is useless). As such, missing the right experimental setting can bring to unsuccessful training and unconvincing results.

More data? A trivial way to offset sparsity is to increase the volume of training data. In fact, existing works that report strong results with targeted evaluation train their contextual parameters with millions of document-level sentence pairs (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019b; Zheng et al., 2020; Wong et al., 2020; Kang et al., 2020). In contrast, many works in the literature train models with the TED talks' subtitles released by the IWSLT shared tasks (Cettolo et al., 2012), which only consist of a couple of hundred thousand parallel sentences. In the experimental section, we will show that IWSLT's subtitles are not sufficient to effectively train multi-encoder models. It follows that one can not make fair comparisons between alternative architectures in such experimental settings. On the other hand, we will give an empirical confirmation to the intuition that increasing the volume of training data helps learning contextual parameters. However, this solution is inefficient and only partial for the double sparsity problem, and it is not always possible: large document-level training sets may not be available in many languages. In the following section, we propose a pre-training solution that makes an efficient use of the available data for learning contextual-parameters effectively.

4 Proposed Approach

One way to simulate document-level data is to split sentences in two or more segments (Luong et al., 2016). In this way intra-sentential syntactic relations are broken, and a word previously disambiguated by looking at its neighbours in the

Algorithm 1: Split parallel corpus

input: Parallel corpus $\mathcal{C}$, minimum source length l_{min}, function $\mathrm{wheresplit}()$
for $i = 1, \ldots, |\mathcal{C}|$ **do**
  if $len(S^i) \geq l_{min}$ **then**
    $m_S, m_T = \mathrm{wheresplit}(S^i, T^i, ...)$
    $S^{i,1} = S^i_{<m_S}$ and $S^{i,2} = S^i_{\geq m_S}$
    $T^{i,1} = T^i_{<m_T}$ and $T^{i,2} = T^i_{\geq m_T}$
  end if
end for
return Split corpus $\mathcal{C}_D$

sentence, now requires contextual information in order to be correctly translated. Moreover, splitting sentences increases the concentration of relevant context words, as we will show in Section 4.2. Within the framework of MT, if we split the source sentence, its corresponding reference has to be split too. The proposed approach, *divide and rule* (*d&r*), consists in pre-training the model on a dataset $\mathcal{C}_D$ that results from splitting all the sentences of a parallel corpus $\mathcal{C}$ that have at least l_{min} tokens, as described by Algorithm 1. Each source-side sentence S^i, with index $i = 1, ..., |\mathcal{C}|$, is split into $S^{i,1}$ and $S^{i,2}$. Its corresponding reference T^i is split into $T^{i,1}$ and $T^{i,2}$. The resulting corpus is a document-level parallel corpus $\mathcal{C}_D$, such that, if the original corpus C was itself document-level, then $\mathcal{C}_D$ keeps the same document boundaries as C. Figure 1 illustrates two examples of parallel sentences that are split in the middle. In both examples, a context-aware system needs to look at $S^{i,1}$ for translating $S^{i,2}$ correctly, i.e. to look at past context. In the first one, the English neuter pronoun “it" could be translated into “il" or “elle", according to the gender of its antecedent (there is no singular neuter 3rd-person in French). The antecedent “a project", which is in the previous segment, allows to disambiguate it into “il". In the second example, the adjective “possible” can be correctly translated into its plural version “possibles” by looking back at the noun it refers to: “organisms”.

4.1 Splitting methods

In Algoritm 1, the $\mathrm{wheresplit}$ function returns the token indices m_S and m_T of S^i and T^i, where the sentence is split. In this work, we propose and experiment with four variants of this function.

Middle-split. The simplest strategy is to split both the source and the target in the middle. In this case, $\mathrm{wheresplit} = \mathrm{middlesplit}(S^i, T^i)$ returns $m_S = \lfloor len(S^i)/2 \rfloor$ and $m_T = \lfloor len(T^i)/2 \rfloor$. Following this method, it can happen that $S^{i,j}$ and $T^{i,j}$, with $j = 1, 2$, are not parallel, as illustrated in the second example of Figure 1. The verb “are” belongs to $S^{i,1}$, but its translation “sont” does not belong to its corresponding reference segment $T^{i,1}$. This problem arises whenever the splitting separates a set of words from their reference, which end up in the other segment. Clearly, this method requires that the two languages do not have strong syntactic divergence, to avoid too large mismatches between $S^{i,j}$ and $T^{i,j}$, with $j = 1, 2$.

Aligned-split. As a solution to the misalignment problem between source and target segments, we can calculate word alignments A^i, and use them to inform our splitting strategy by setting $\mathrm{wheresplit} = \mathrm{alignedsplit}(S^i, T^i, A^i)$, where $\mathrm{alignedsplit}$ splits each sentence close to the middle, while avoiding to separate aligned words in different segments.

Synt-split. The objective of splitting being to break intra-sentential syntactic and semantic relations in order to force the model to exploit the context more frequently, we can run an NLP toolkit over the training set to retrieve relations L (e.g. syntactic dependencies or coreferences), and then by defining $\mathrm{wheresplit} = \mathrm{syntsplit}(S^i, T^i, L^i)$ so that it splits sentences close to the middle, while ensuring that at least a relation is broken whenever possible. Since not all relations raise translation ambiguities when broken, one can choose which of them must be prioritized; in this work we chose pronominal coreferences.

Multi-split. The aforementioned methods can be extended to splitting sentences in more than two segments. The more we split sentences the more likely it is that context is needed for each segment, hence increasing training signal for contextual parameters.

For more details, we refer to Section 6.3, to Appendix A and to our code (will be open-sourced).

4.2 Impact on discourse phenomena

To give an explicit picture of how and why splitting sentences helps learning contextual parameters, we processed the source training data of IWSLT17 with CoreNLP (Manning et al., 2014) and we computed some statistics on coreference chains and dependency parse trees, before and after applying the *middle-split* method. Statistics show how splitting

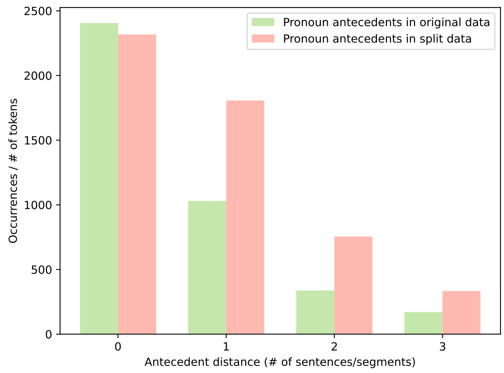

Figure 3: IWSLT's number of antecedents of anaphoric pronouns at a given distance in terms of sentences/segments, normalized by the number of tokens that the model needs to attend for resolving the coreference.

the sentences of a document helps in two ways:

More cases. Splitting generates new cases that require context for disambiguation, making training signal more abundant. When syntactic dependencies are split in two segments, the model needs to access the context for reconstructing the syntactic structure of the source sentence and correctly translate it, as shown in Figure 1. In order to have an idea of the magnitude of this effect, we calculated the percentage of the sentences where the splitting method breaks at least one syntactic dependency between the main verb of the sentence (the root) and : (i) the subject or object (18.1% of the sentences); (ii) any complement (9.5%); (iii) any modifier (9.3%). If we consider all the dependencies with the root, except punctuations, we find that in 84.8% of the sentences at least a syntactic dependency is broken. Given such high proportion, the *middle-split* variant is in fact a good approximation of a syntactically supported splitting approach. These cases add up to the many other cases of broken relations, such as coreferences, which make the overall contextual training signal more abundant.

Denser cases. The splitting also has the effect of shortening the average length of text sequences, which eases the job of context-aware systems because they have to attend to fewer words while looking for context. In Figure 3, we show how many antecedents of an anaphoric pronoun are present in the data at a given distance d, expressed as number of sentences from the current one for original data, and number of segments for split data. $d = 0$ means that both the pronoun and its antecedent are in the same sentence (or segment); $d = 1$ means that the antecedent is in previous sentence (or segment), and so on. We show statistics up to $d = 3$, which is the maximum context distance that we experiment with. The absolute number of antecedents is normalized by the average length of a sentence or segment. The resulting bar plot shows that splitting sentences into segments makes pronominal antecedents more dense in the set of context tokens that the model is attending, which fosters the learning of contextual parameters. The same effect applies to the other discourse phenomena that require contextual disambiguation.[1]

5 Experimental setup

5.1 Data

We conduct experiments for three language pairs English→Russian/German/French on different domains. Following Kim et al. (2019), we pre-train sentence-level baselines on large sentence-level parallel data to make them as robust as possible. In particular, we employ data released by Voita et al. (2019b) for En→Ru (6.0M sentences from OpenSubtitles2018 (Lison et al., 2018)), data from the WMT17[2] news translation shared task for En→De (∼5.2M sentences), and data from WMT14[3] for En→Fr (∼35.8M sentences). We train the contextual parameters of context-aware models in two settings, while freezing the rest of ther parameters:

High resource data. For En→Ru, it consists of the document-level data released by Voita et al. (2019b). For the other two language pairs, we build the training set by assembling (i) News-Commentary-v12 for En→De and News-Commentary-v9 for En→Fr; (ii) Europarl-v7 for En→De/Fr; (iii) TED talks subtitles released by IWSLT17 (Cettolo et al., 2012) for En→De/Fr.

Low resource data. For En→Ru, it consists of 1/10th of a random shuffle of the high resource data. For En→De/Fr, we use IWSLT17's TED talks alone.

The resulting size of the two training settings after pre-processing is reported in Table 1. In the case of En→De/Fr, baselines and context-aware models that were trained on high resources are also fine-tuned on IWSLT17, so that both high and low resource settings can be bench-marked on the IWSLT17's test set 2015. Test-sets 2011-2014 are used as development set. For En→Ru, we use the

[1]More details are available in Appendix B, along with the same statistics for Opensubtitles2018.

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

[3]http://www.statmt.org/wmt14/translation-task.html

	En→Ru	En→De	En→Fr
Low Res	0.15M (8.3)	0.20M (20.8)	0.23M (21.0)
High Res	1.50M (8.3)	2.29M (27.29)	2.31M (27.6)

Table 1: Millions of sentence pairs used for training context-aware models, and their average source length.

dev and test sets provided by Voita et al. (2019b). [4]

5.2 Evaluation

Besides evaluating average translation quality with BLEU (Papineni et al., 2002),[5] we employ three contrastive test suites for the evaluation of translation of discourse phenomena.[6]

En→Ru EllipsisVP (Voita et al., 2019b). Consisting of 500 examples from OpenSubtitles2018, each containing multiple contrastive hypothesis to evaluate the translation of verb phrase ellipses. Source sentences contain an auxiliary verb (e.g. "do") and an omitted main verb, which can be imputed thanks to one of th needine three context sentences. Voita et al. (2019b) proposed test sets for the evaluation of other discourse phenomena, but we do not use them because they are conceived for systems using also target-side context.

En→De ContraPro (Müller et al., 2018). A large-scale test set from OpenSubtitles2018 (Lison et al., 2018), that measures translation accuracy of the English anaphoric pronoun *it* into the corresponding German translations *er*, *sie* or *es*. Examples are balanced across the three pronoun classes (4,000 examples each). Each example requires identification of the pronominal antecedent, either in the source or target side, that can be found in the current sentence or any of the previous ones.

En→Fr ContraPro (Lopes et al., 2020). A large-scale test set from OpenSubtitles2018, completely analogous to the previous one but focused on the translation of two English pronouns: *it* and *they*. It consists of 3,500 examples for each target pronoun type: *il* or *elle* for *it*, *ils* or *elles* for *they*.

5.3 Models

We experiment with three models:

K0. A sentence-level baseline, following the *Transformer-base* by Vaswani et al. (2017).

K1. A context aware multi-encoder architecture with *outside integration* (see Section 2.2), that encodes a single past source sentence as context.

K3. A context aware multi-encoder architecture with *outside integration*, that encodes three past source sentences as context.[7]

For both *K1* and *K3*, sentence-level parameters θ_S follow the *Transformer-base* configuration (hidden size of 512, feed forward size of 2048, 6 layers, 8 attention heads, total of 60.7M parameters), while contextual parameters θ_C follow hierarchical architecture with source-side encoder proposed by Miculicich et al. (2018) (hidden size of 512, feed forward size of 2048, 8 attention heads, total of 4.7M parameters).[8] Context-aware models are trained following the *two-step strategy* described in Section 2.2. Sentence-level parameters θ_S of both *K1* and *K3* are initialized with *K0* and freezed. This has the advantage of saving time and computation, since only a small fraction of parameters (θ_C) is trained (4.7M over a total of 65.2M).

6 Results and Analysis

6.1 Training contextual parameters is hard

In this section we provide evidence about the difficulty of training contextual parameters on document-level data. In the second block of Table 2, after the results of the sentence-level baseline *K0*, we report performance of context-aware models trained on original document-level data, comparing low and high resource settings. When trained on the low resources, models display good BLEU on the test set, generally without strong degradation with respect to *K0*, or even with some improvements. However, such marginal fluctuations in BLEU are difficult to interpret, as they do not necessarily correspond to better or worse translation (Freitag et al., 2020). Accuracy on the contrastive test sets also increases marginally over baseline, if at all, for En→De/Fr. *K1* even shows a slight degradation of performance over the sentence-level baseline for En→Fr. These results highlight the struggle of contextual parameters to learn an appropriate use of context, other than acting as mere regularizers, as it was suggested by Kim et al. (2019) and Li et al. (2020). On Russian instead, models display some improvements

[4] We report in Appendix C a re-cap of the datasets used and details about pre-processing.

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

[6] Whenever relevant, we calculate the statistical significance of the differences between models' accuracies with the paired McNemar test (McNemar, 1947).

[7] Although the splitting does not increase the number of inter-segment phenomena for $d > 1$, it strengthens the signal by making it more dense (see Section 4.2). Thus, *K3* and any wider-context model can profit from the proposed approach.

[8] Details can be found in Appendix C

Model	Setting	En→De BLEU	En→De **ContraPro**↑	En→Fr BLEU	En→Fr **ContraPro**↑	En→Ru BLEU	En→Ru **Ellipsis-VP**↑	Avg. Train Hours
Concat2to1	Low Res	33.41	47.38	41.27	80.42	31.12	31.00	1.9
Concat2to1	High Res	33.05	59.49	40.99	85.57	29.92	62.6	7.3
Zhang2018	Low Res	31.03	42.60	40.95	59.00	n.a.	n.a.	n.a.
K0	-	32.97	46.37	41.63	79.46	31.37	25.40	-
K1	Low Res	33.14	47.05	41.93	79.24	30.89	32.20	2.9
K3	Low Res	32.86	46.48	41.40	80.53	31.00	29.20	3.5
K1	High Res	33.16	57.75	41.65	84.32	31.15	44.00	13.0
K3	High Res	33.1	51.14	41.95	82.94	31.23	39.20	16.8
K1-d&r	Low Res	33.44	60.21*	41.78	84.06	31.09	47.00*	6.7
K3-d&r	Low Res	33.36	56.22*	41.68	85.50*	32.12	46.60*	6.4
K1-d&r	High Res	32.82	**61.09***	41.81	84.17	31.09	59.40*	16.5
K3-d&r	High Res	33.07	59.56*	41.91	**85.66***	31.27	**60.40***	22.3

Table 2: BLEU score on testsets and accuracy (%) on contrastive sets. The last column reports the average context-aware training time (in hours), including the time for *d&r* pre-training. The symbol * denotes statistically significant (p<0.01) improvements w.r.t non-*d&r* couterparts (second block) and *K0*.

w.r.t. *K0*. This aligns with our expectations, since En→Ru Low Res has a volume of inter-sentential discourse phenomena such as coreferences that is comparable with En→De/Fr Low Res, but sentences are 2.5x shorter.[9] In other words, the *double challenge of sparsity* is mitigated on this corpus. When trained on high resources, systems show substantial improvements in their context-modeling capabilities, on all language pairs. Instead, BLEU improves of a few decimal points only, showing its limits to measure improvements in context-aware translation. These results confirm the intuition discussed in Section 3: increasing the volume of data is a first solution to overcome sparsity.

For the sake of benchmarking, we report in the first block results obtained by two other source-side context-aware models trained on low resource data, following the same procedure.[10] *Concat2to1* (our implementation) is a single-encoder approach that concatenates the previous sentence to the current one as context, and outputs the translation for the current one. *Zhang2018* is a multi-encoder model that looks at 2 previous sentences as context, proposed by Zhang et al. (2018).[11] *Concat2to1*'s performance on test suites are comparable to *K1/3* on Low Res, or slightly better since concatenation models are less affected by the problem of sparsity. This advantage is better highlighted on the high resource setting, in which *Concat2to1* is stronger on the test suites (although BLEU lacks behind). *Zhang2018* performs very poorly, confirming the difficulty of multi-encoder models to learn contextualization on low resources and without any help against the problem of sparsity.

6.2 Divide and rule

In this section, we show that the proposed pre-training strategy is a more efficient answer to the double challenge of sparsity than simply adding more data, and one that allows improvements when resources are abundant too. The third block of Table 2 displays performance of models that have undergone *d&r* pre-training on the same document-level data as the models in the previous block, but where sentences were split in two segments following the *middle-split* method with $l_{min} = 7$. After *d&r* pre-training, models have been finetuned on the original, non-split data. The pre-training proves to be very effective, as all models show strong improvements in terms of accuracy on the test suites, with the sole exception of *K1-d&r* on En→Fr High Res. The average improvement is of **+10.79** accuracy points on Low Res, **+8.49** on High Res, showing that *d&r* brings strong improvements even when data are abundant. Interestingly, improvements are not uniformly distributed across language pairs and domains: **+17.20** on average for En→Ru, **+8.67** for En→De, **+3.09** for En→Fr. In terms of BLEU instead, we keep seeing minor fluctuations. This confirms that, while context-aware translation improves dramatically, the average translation quality measured with BLEU stays more or less constant.[12] It is now clear that a proper

[9] See Table 1; more details can be found in Appendix B

[10] We do not compare with target-side approaches as we experimented with source-side only.

[11] Results reported are by Lopes et al. (2020)

[12] To verify that the improvements on test suites after *d&r* pre-training really come from a better use of context, we present in Appendix D an analysis of pronoun translation

	En→De			
	Middle↑	Aligned↑	Synt↑	Multi↑
K1-d&r	60.21	+0.69*	-2.67*	-
K3-d&r	56.22	-1.38*	+1.33*	+1.13*
	En→Fr			
K1-d&r	84.06	+0.27	+0.15	-
K3-d&r	85.50	+0.20	+0.33**	-0.09

Table 3: Comparison of accuracy of context-aware pronoun translation (ContraPro) by *d&r* pre-trained models with the *middle-split* method (first column) and the other proposed methods (relative difference). *: $p < 0.01$, **: $p < 0.05$.

comparison between single and multi-encoder models can not be done without proper training of the multi-encoders' contextual parameters, which targets the problem of sparsity. Here, *d&r* pre-training allows *K1/3* to achieve results on test suites comparable to *Concat2to1* (*K3* is consistently better), along with better BLEU scores (except for *K1* on german).[13] A comparison between -*d&r* models trained on Low Res against models trained on High Res without *d&r* shows another quality of the *d&r* pre-training strategy: efficiency. The same context-aware models achieve superior performances with 1/10th of the document-level data and a much shorter training time (last column).

6.3 Impact of the splitting method

Following Section 4.1, we study the impact of using a different splitting method other than *middle-split*. All the variants are applied to the En→De/Fr low resource setting (IWSLT), with $l_{min} = 7$, and the *d&r* pre-trained models are evaluated on ContraPro. The *aligned-split* method is based on alignments learned with *fast_align* (Dyer et al., 2013), while for the *synt-split* method we retrieve intra-sentential pronominal coreferences with CoreNLP (Manning et al., 2014), and we try to split them wherever present in a sentence-pair. If there are multiple-occurrences in the same sentence, we split as close to the middle as possible, while attempting to break the maximum number of coreferences.[14] Finally, for the *multi-split* method, we split sentence-pairs in a half for $len(S^i) \geq 7$, and also in three segments of identical size for $len(S^i) \geq 15$. The performance differences between models pre-trained with *middle-split* and the other variants are reported in Table 3. As we can see, splitting variants allow small improvements in 7 cases out of 10, although variations are marginal: the simple *middle-split* method seems to be close to optimal already. This observation can be explained by multiple elements. Firstly, *middle-split* produces segment pairs that are already well aligned: most of the source and target segments are aligned with the exception of one or two words, and the fact of having only a few misplaced words might act as a regularization factor. Secondly, *middle-split* breaks a syntactic relation for the vast majority of sentences already, as explained in Section 4.1, which means that improvements achieved with syntactically driven splitting can only be marginal. Thirdly, splitting in more than one segment can be beneficial in some cases, because it allows to break more syntactic relations and increase density of signal, but it also increases the risk of misalignment between source and target, and might make the task too hard. Finally, tools like *fast_align* and CoreNLP are characterized by a certain language-dependent error rate, which affects the performance of the methods. In conclusion, *d&r* pre-training with *middle-split* seems to be the most convenient alternative for most use-cases because of its efficacy, its simplicity and its language-independence. Even though *middle-split* relies on syntactic similarity between target and source languages, this condition is met by a large number of language pairs, in the order of millions, as detailed in Appendix A.

7 Conclusions

Multi-encoder models are a broad family of context-aware NMT models. In this work we have discussed the difficulty of training contextual parameters due to the sparsity of the words in need of context, and their relevant context. We have proposed a pre-training approach called *divide and rule*, based on splitting the training sentences, with four variants. After having analysed the implications of splitting on discourse phenomena, we have shown that *d&r* allows to learn contextual parameters better and faster than by simply adding training data. We have also shown that the simplest and language independent splitting variant, *middle-split*, is a strong baseline that can be easily applied for pre-training any multi-encoder NMT model.

by antecedent distance, and an ablation study in which we test models on ContraPro with inconsistent context.

[13] A detailed comparison between single and multi-encoder models is beyond the scope of this work.

[14] More sophisticated *synt-split* methods could be devised, targeting other discourse phenomena, or several of them at the same time, with different degrees of priority.

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.

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.

Matthew S. Dryer. 2013a. Order of Adjective and Noun. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Matthew S. Dryer. 2013b. Order of Adverbial Subordinator and Clause. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Matthew S. Dryer. 2013c. Order of subject, object and verb. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Matthew S. Dryer and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fenning. 2021. Ethnologue: Languages of the World. Library Catalog: www.ethnologue.com.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. 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.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? *arXiv:1704.05135 [cs, stat]*. 00039 arXiv: 1704.05135.

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, 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.

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

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

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. *arXiv:1511.06114 [cs, stat]*. ArXiv: 1511.06114.

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.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

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. 2019a. 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.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019b. A Survey on Document-level Machine Translation: Methods and Evaluation. *arXiv:1912.08494 [cs]*. 00000 arXiv: 1912.08494.

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.

Andrei Popescu-Belis. 2019. Context in Neural Machine Translation: A Review of Models and Evaluations. *arXiv:1901.09115 [cs]*. 00010 arXiv: 1901.09115.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in*

Machine Translation (DiscoMT 2019), pages 51–61, Hong Kong, China. Association for Computational Linguistics.

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

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

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

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.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient Transformers: A Survey. *arXiv:2009.06732 [cs]*. 00008 arXiv: 2009.06732.

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.

Russell S. Tomlin. 2014. *Basic Word Order (RLE Linguistics B: Grammar): Functional Principles*. Routledge. Google-Books-ID: OlPIAgAAQBAJ.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Long Beach, California, USA. Curran Associates Inc.

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

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

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

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.

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-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization.

Coreferences - original data			
d	#tokens	Occurrences All	Occurrences Pronouns
0	21.01	67,864 (3230)	50,556 (2406)
1	42.02	68,703 (1635)	43,220 (1029)
2	63.03	35,780 (568)	21,234 (337)
3	84.04	25,533 (304)	14,284 (170)
Coreferences - split data			
d	#tokens	Occurrences All	Occurrences Pronouns
0	10.51	32,190 (3063)	24,328 (2315)
1	21.02	54,424 (2589)	37,966 (1806)
2	31.53	37,837 (1200)	23,732 (753)
3	42.04	22,529 (536)	14,035 (334)

Dependency trees	
Split dependency	Occurrences
subj or obj	41,065
complement	21,726
modifier	21,144
any	147,066

Table 4: Number of coreference antecedents at a given distance d from the mention in the current sentence, for both original and split En→Fr IWSLT17. In brackets, the same figure normalized by the average number of tokens that the model has to attend to resolve the coreference (#tokens). At the bottom, the number of sentences for which at least one syntactic dependency is split in two segments when using the split data.

Coreferences - original data			
d	#tokens	Occurrences All	Occurrences Pronouns
0	8.32	36,628 (4402)	27,179 (3267)
1	16.64	60,204 (3618)	41,652 (2503)
2	24.96	26,397 (1058)	16,142 (647)
3	33.28	11,571 (348)	6,654 (200)
Coreferences - split data			
d	#tokens	Occurrences All	Occurrences Pronouns
0	4.16	13,322 (3202)	9,134 (2196)
1	8.32	46,227 (5556)	34,104 (4099)
2	12.48	33,566 (2690)	22,676 (1817)
3	16.64	18,961 (1139)	12,248 (736)

Table 5: Same as in table 4 for the Low Resource subset of En→Ru OpenSubtitles2018.

A Splitting methods

We provide here some extra details on the splitting methods that have been proposed and tested. For full details, we refer to our implementation.

Aligned-split. As already mentioned, we use $\mathrm{wheresplit} = \mathrm{alignedsplit}(S^i, T^i, A^i)$, which takes as input the word alignments A^i:

$$A^i = \{(j,k) | S^i_j \text{ and } T^i_k \text{ are aligned}\},$$

where $j = 1, ..., |S^i|$ and $k = 1, ..., |T^i|$ are the indices of the words belonging to S^i and T^i, respectively. $\mathrm{alignedsplit}$ initially takes $m_S = \lfloor len(S^i)/2 \rfloor$ and $m_T = \max\{k\colon (j,k) \in A^i, j \leq m_S\}$. Then, it checks whether this choice is not breaking apart two aligned words. Formally, it checks that:

$$S^i_j \in S^{i,1} \wedge T^i_k \in T^{i,1} \text{ or } S^i_j \in S^{i,2} \wedge T^i_k \in T^{i,2}. \quad (1)$$

If this condition is not encountered, it tries to split the sentence pairs in the neighbouring distance, where condition (1) is met. If the condition can not be met (e.g., because one of the two segments would be too short (<3 tokens)), $\mathrm{alignedsplit}$ falls back on $\mathrm{middlesplit}$.

Synt-split. In our implementation, the function $\mathrm{wheresplit} = \mathrm{syntsplit}(S^i, T^i, L^i)$ takes as input the coreference relation L^i detected by CoreNLP on the source sentence i. If L^i is not empty, it means that a relevant intra-sentential relation is present (in our experiments, we look at pronominal coreferences). In this case, the algorithm checks whether splitting in the middle ($m_S = \lfloor len(S^i)/2 \rfloor$) allows to break L^i, i.e., to separate the two related tokens in different segments. If *middle-split* does not achieve this goal, m_S is set to the closest index from the middle that breaks the relation, except for the case in which breaking the relation would mean generating a too short segment (<3 tokens). In this case, the algorithm falls back to *middle-split*.

A.1 On the scope of middle-split

Even though *middle-split* relies on syntactic similarity between source and target languages, this condition is met by a large number of language pairs, in the order of millions. In fact, there are around 4,000 written languages in the world (Eberhard et al., 2021), and most of them can be grouped in a few types with similar word orders, as shown by the ample literature on word order typologies (Tomlin, 2014; Dryer and Haspelmath, 2013). The primary order of interest is the *constituent order*, concerning the relative order of subject (S), object (O) and verb (V) in a clause. There are seven possible language types with respect to the constituent order (Dryer, 2013c): SOV, SVO, VSO, VOS, OVS, OSV, NDO (non-dominant or-

	Total	$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d > 3$
K0	46.37	**83.3**	32.4	44.8	48.9	71.9
K1	47.05	82.5	33.9	45.3	48.0	69.9
K3	46.48	82.4	32.8	45.0	48.9	71.7
K1-d&r	**60.21**	81.1	**56.5**	44.9	48.7	**73.3**
K3-d&r	56.22	81.7	46.8	**55.2**	**56.2**	72.4
Sample Size	12000	2400	7075	1510	573	442
Relative Size	100.0%	20.0%	59.0%	12.6%	4.8%	3.7%

Table 6: Accuracy(%) of Low Res models on ContraPro En→De by pronoun antecedent distance. The first column represents the weighted average, calculated on the basis of the sample size of each group.

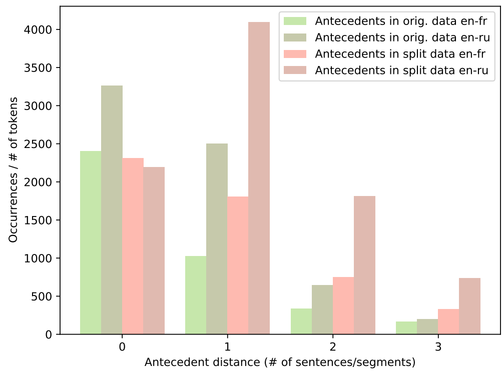

Figure 4: En-Fr IWSLT vs Low Res En-Ru OpenSubtitles2018: comparison of the number of antecedents of anaphoric pronouns at a given distance in terms of sentences/segments, normalized by the number of tokens that the model needs to attend for resolving the coreference. Since sentences are much shorter in the En-Ru data (8.32 vs. 21.02 tokens on average), the density of discourse phenomena within the sentence is much higher.

der). Tomlin (2014) estimates that more than 40% of the world languages belong to the SOV type (languages adopting the SOV order), another 40% belong to the SVO type, while almost 10% of languages adopt VSO order. The other types are rarer. In the previous section, we have shown that the middle-split method is beneficial both in the case of language pairs of the same type, that deploy the same constituent order, like En-Fr/Ru, which all adopt SVO order, as well as for languages that belong to different types, as for En-De, where English is SVO and German is NDO, deploying both SOV and SVO according to the use cases (Dryer, 2013c). Similar observations also apply when we look at other word order categories. For instance, when looking at the order of modifiers or adverbials, languages can be clustered in a few types, where the wide majority of languages belong to the biggest or second biggest type (Dryer, 2013b,a). Therefore, we believe that our method could be beneficial for millions of language pairs, including many low resource languages belonging not only to same word order types, but also slightly different ones (as in the case of SOV and SVO).

B Impact of splitting

In Table 4, we provide details on the syntactic features and the impact of splitting (with *middle-splip*) for En→Fr IWSLT17, while Table 5 show the equivalent figures for the Low Resource subset of En→Ru OpenSubtitles2018. A visual comparison of the two datasets is presented in Figure 4. This complementary information confirms that the *middle-split* method is an effective way to strengthen the contextual training signal and to facilitate its exploitation by context-aware NMT systems, in different text domains.

C Experimental Setup

C.1 Data recap

We recap in Table 7 the datasets that we use at each stage of training and test. The sentence-level training concerns the baselines, whose parameters are also used to initialize the sentence-level encoder and decoder of the context-aware models (Θ_S). Concerning En→Ru, Voita et al. (2019b) released two datasets extracted from OpenSubtitles2018: a document-level dataset of 1.5M sentences with context (document boundaries are available), and a sentence-level dataset of 6M sentences, which includes the sentences of the document-level dataset.

C.2 Data preprocessing

The Opensubtitles2018 release by Voita et al. (2019b) has been already pre-processed. Therefore, we only apply byte pair encoding (Sennrich et al., 2016) using 32k merge operations jointly for source and target languages.

	En→Ru		En→De		En→Fr	
	Low Res	**Hig Res**	**Low Res**	**Hig Res**	**Low Res**	**Hig Res**
Sentence-level training	OpenSubs2018	OpenSubs2018	WMT17	WMT17	WMT14	WMT14
Context-aware training	1/10th of OpenSubs2018	OpenSubs2018	IWSLT17	News-v12 Europarl-v7 IWSLT17	IWSLT17	News-v9 Europarl-v7 IWSLT17
Fine-tuning	-	-	-	IWSLT17	-	IWSLT17
Test (BLEU)	OpenSubs2018	OpenSubs2018	IWSLT17	IWSLT17	IWSLT17	IWSLT17
Contrastive test	EllipsisVP	EllipsisVP	ContraPro	ContraPro	ContraPro	ContraPro

Table 7: Summary of the datasets used at each stage of training and evaluation of the models.

Model	**Setting**	**En→Ru**	**En→De**	**En→Fr**
Concat2to1	Low Res	3.624	3.628	3.207
Concat2to1	High Res	3.659	3.734	3.228
K0	-	3.626	3.629	3.230
K1	Low Res	3.599	3.617	3.216
K3	Low Res	3.605	3.618	3.215
K1	High Res	3.596	3.617	3.210
K3	High Res	3.597	3.617	3.211
K1-d&r	Low Res	3.595	3.617	3.213
K3-d&r	Low Res	3.595	3.616	3.212
K1-d&r	High Res	3.593	3.616	3.211
K3-d&r	High Res	3.592	3.615	3.211

Table 8: Corresponding loss on development set for each reported test result with *middle-split*.

$\mathbf{P}_{\text{len}}$	**En→Ru**	**En→De**	**En→Fr**
0.6	**31.76**	**32.80**	44.47
0.7	31.58	32.76	44.48
0.8	31.47	32.72	44.50
0.9	31.33	32.65	44.53
1	31.23	32.64	**44.59**
1.1	31.12	32.60	**44.59**
1.2	31.06	32.57	44.58

Table 9: Performance (BLEU) of *K0* on the development set according to different values of length penalty.

The other datasets are tokenized with the Moses toolkit (Koehn et al., 2007), further cleaned by removing long sentences, and byte pair encoded using 32k merge operations jointly for source and target languages. While IWSLT provides document boundaries for TED subtitles, the WMT releases of New-Commentary and Europarl do not provide them. Therefore, a small fraction of sentences in the High Resource setting will be paired with wrong context. However, we found the models to be robust against occasional random context (see also Voita et al. (2018) and Müller et al. (2018)). In order to make the models correctly learn how to translate headlines (the first line in a document), we need to have headlines in the training set. As such, we set artificial document boundaries in News-Commentary and Europarl, following the average document length of TED talks.

C.3 Training and evaluation

All models are implemented in *fairseq* (Ott et al., 2019). After having pre-trained the baseline on 4 Tesla V100 for 200k steps, we train all models on a single Quadro RTX 6000, with a fixed batch size of approximately 16k tokens,[15] 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 5 consecutive non-improving validation steps (in terms of loss on dev). Corresponding validation performance for each reported test result with *middle-split* are reported in Table 8. We train models with the optimizer configuration and learning rate (LR) schedule described in Vaswani et al. (2017). The maximum LR is 0.0007 for baselines on En→Ru/De, 0.001 for models on En→De/Fr low resource settings, and 0.0005 for all the others. In the En→De/Fr High Resource setting, contextual-parameters are finetuned on IWSLT17 with an initial LR of 0.0002 that shrinks by a factor of 0.99 at every epoch. We use label smoothing with an epsilon value of 0.1 (Pereyra et al., 2017) for all settings. Since the sentence-level parameters are pre-trained on a large amount of parallel data (WMT), the models are pretty robust to generalization, and dropout can be set to 0.1, which gave the best results for the non-contextual baseline *K0*. At inference time, we use beam search with a beam of 4 for all models. We adopt a length penalty (P_{len}) of 0.6 for

[15]The optimizer update is delayed to simulate the 16k tokens.

Model	En→De BLEU	En→De ContraPro	En→Fr BLEU	En→Fr ContraPro
K0	32.97 (+0.00)	46.37 (0.00)	41.44 (-0.00)	79.46 (0.00)
K1	33.06 (+0.06)	46.7 (-0.35)	41.75 (-0.12)	79.05 (-0.19)
K3	32.73 (-0.13)	46.21 (-0.27)	41.47 (+0.15)	79.24 (-1.29)
K1-d&r	33.1 (-0.34)	47.6 (-12.61)	41.64 (-0.14)	78.94 (-5.12)
K3-d&r	33.05 (-0.31)	47.96 (-8.26)	41.55 (-0.13)	79.05 (-6.45)

Table 10: BLEU and accuracy results on ContraPro (and their changes) when the context provided to the model is inconsistent. All models are trained on the Low Resource setting.

all models ($P_{len} < 1$ favors shorter sentences), with the exception of En→Fr models, to which we assign $P_{len} = 1$. The LR for training was searched in $\{0.001, 0.0007, 0.0005, 0.0002\}$). The LR achieving the best loss on the validation set after convergence was selected. P_{len} was searched in $\{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$ for *K0* only (see Table 9). The length penalty resulting in the best BLEU score on the validation set was then used for all models within the same language pair. The other hyperparameters were set according to the relevant literature (Vaswani et al., 2017; Popel and Bojar, 2018; Voita et al., 2019b; Lopes et al., 2020).

D Results Analysis

D.1 Accuracy by antecedent distance

Here, we want to investigate more in details the performance of the proposed approach on the accuracy of ambiguous pronoun translation. We report in Table 6 the accuracy on En→De ContraPro, detailed by varying antecedent distance. We notice that all the improvements achieved by *-d&r* models are related to those pronouns whose antecedent is in the context ($d \geq 1$), which is in line with the expectations of context-aware models exploiting context for disambiguation. *K1-d&r* is very strong in translating pronouns with antecedent distance $d = 1$, surpassing *K0* and *K1* baselines by 22+ points of accuracy. Similarly, *K3-d&r* surpasses baselines by a large margin on $0 \leq d \leq 3$, beating all the other models on $d = 2, 3$, as expected. We notice however that *K3-d&r* lacks behind *K1-d&r* on $d = 1$. On one side, this could be explained by the fact that *K1-d&r* is more specialized at modeling a single past sentence. On the other side, we also notice that the hierarchical context-encoding architecture by Miculicich et al. (2018), at the core of *K3*, is not aware of the distance of the context sentences that are encoded. Hence, we believe that *K3-d&r* might perform worse on $d = 1$ than *K1-d&r* because it gives the same importance to further away context ($d = 2, 3$). Since pronouns with antecedent distance $d = 1$ are the most frequent in the test set, *K1-d&r* has the highest average result (reported in "Total"). It has to be noticed also that *K3* is more affected by the challenge of sparsity than *K1*, since it has to spot relevant context among 3x more tokens. This might be the reason why *K3* starts beating *K1* only when the training setting is the most favorable to context-aware learning: with *d&r* pre-training plus high resources.

D.2 Ablation: shuffling context

We want to verify that the proposed approach improves learning by making the context-aware model to rely on its modeling of the context. Table 10 shows the performance of models trained on Low Res, when the evaluation is undertaken by randomly shuffling the context of every sentence with other sentences from the same dataset (c.f. Scherrer et al. (2019)). In brackets, the delta w.r.t. the results with consistent context presented in the main table of the paper. A random context is inconsistent with the current sentence in many cases, and thus misleading for a context-aware system. Indeed, *-d&r* models display a significant drop in accuracy when they are evaluated with inconsistent context, which confirms that they rely on context information to achieve the improvement in pronoun translations. Nonetheless, the same models prove to be robust against being shown a random context as they obtain a similar performance to *K0*. In other words, the splitting method does not produce models that are over reliant on context. This robustness is confirmed by BLEU: the average translation quality is very slightly affected by the shuffling. The changes are so small that are probably negligible. This results also show once again that BLEU is ill-equipped to measuring improvements in document-level translation.