HAL open science

Divide and Rule: Training Context-Aware Multi-Encoder Translation Models with Little Resources

Lorenzo Lupo, Marco Dinarelli, Laurent Besacier

To cite this version:

Lorenzo Lupo, Marco Dinarelli, Laurent Besacier. Divide and Rule: Training Context-Aware Multi-Encoder Translation Models with Little Resources. 2021. hal-03455113

HAL Id: hal-03455113

https://hal.science/hal-03455113v1

Preprint submitted on 29 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Divide and Rule: Training Context-Aware Multi-Encoder Translation Models with Little Resources

Lorenzo Lupo[1] **Marco Dinarelli**[1] **Laurent Besacier**[1,2]
[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

arXiv:2103.17151v1 [cs.CL] 31 Mar 2021

Abstract

Multi-encoder models are a broad family of context-aware Neural Machine Translation (NMT) systems that aim 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 show that training these parameters takes large amount of data, since the contextual training signal is sparse. We propose an efficient alternative, based on splitting sentence pairs, that allows to enrich the training signal of a set of parallel sentences by breaking intra-sentential syntactic links, and thus frequently pushing the model to search the context for disambiguating clues. We evaluate our approach with BLEU and contrastive test sets, showing that it allows multi-encoder models to achieve comparable performances to a setting where they are trained with $\times 10$ document-level data. We also show that our approach is a viable option to context-aware NMT for language pairs with zero document-level parallel data.

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 been even 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 multi-encoder models learn effectively when given enough *contextual training signal*, i.e. the signal that pushes systems to translate in a context-aware fashion. We show that the simplest way to increase *contextual* training signal effectively 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, this solution is not always ideal because it requires more time and computation. We therefore introduce a simple and yet powerful technique to augment the contextual training signal that can be extracted from a small parallel training set: splitting parallel sentences in segments (see Figure 1). Feeding a context-aware model with a sequence of incomplete, consecutive segments, forces it to look for context (i.e., surrounding segments) more frequently, which makes the model learn faster to translate in a context-aware fashion. We train multi-encoder models on split datasets and evaluate them in two ways: BLEU score, and a contrastive evaluation of anaphoric pronoun translation. 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; (ii) we propose a method to augment the contextual training signal by splitting sentences in two segments, with two variants; (iii) we

$S^{i,1}$	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**).

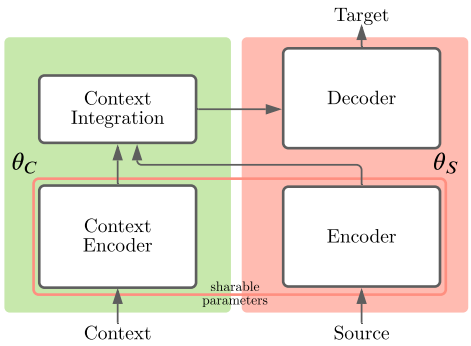

Figure 2: Multi-encoder approach integrating context outside the decoder.

support this method with a statistical analysis of its impact on the distribution of discourse phenomena requiring context; (iv) we demonstrate that this method is both effective and efficient, as it allows multi-encoder models to learn better and faster than by simply increasing the training data; (v) we provide a proof of concept for the application of our approach to train context-aware models in a setting where document-level data are not available.

2 Background

2.1 Single-encoder approaches

The most straightforward approach to context-aware NMT consists in concatenating the contextual sentences to the current source 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 make translation models distinguish the boundaries between sentences. Generation can then follow two strategies: the *many-to-many* strategy consists in decoding as many sentences in the target language as those input in the source language, and then discarding all the contextual sentences; the *many-to-one* strategy consists in decoding 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 that encode and integrate the context of the current sentence, either on source side, target side, or both. We refer to this set of 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 in which 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 the parameters (some or all of them) of the current-sentence encoder and the context encoder (Voita et al., 2018; Li et al., 2020). In this way, the amount of contextual parameters to learn, $|\theta_C|$, is drastically reduced, as well as the computation required to encode a sentence. 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. Both single and multi-encoder models can process one or multiple context sentences, from the past or future. 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 $\mathcal{C}_S$. Secondarly, 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). As a consequence, 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. Contrastive 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 the correct translation higher than the incorrect one. In many cases, the correct translation can only be identified by looking at context, which is provided for both source and target sides. Therefore, the ranking accuracy reflects the context-modeling ability of the translation system under evaluation.

3 A sparse training signal

Some works have criticized multi-encoder methods (Kim et al., 2019; Li et al., 2020), arguing that they do not improve over 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 to be attributed 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 (see Section 6.4). Indeed, many proponents of multi-encoder models only show BLEU improvements, without providing any kind of targeted evaluation. This raises the suspicion that their context-modeling capability is negligible. We posit that training contextual parameters is non-trivial because of the sparsity of the training signal, i.e. the sparsity of words that need context to be correctly translated. As such, missing the right experimental setting can bring to unsuccessful training and unconvincing results.

More data. A straightforward way to offset the sparsity of training signal 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 (see Table 1). With a set of systematic experiments (see Section 6), we 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. We also give an empirical confirmation to the intuition that increasing the volume of training data is a solution to this difficulty.

However, the need of a large training set is not ideal, since it slows down the *speed of iteration*, i.e. how long it takes to test a new model, which translates in slower progress. Moreover, large document-level training sets may not be available in many languages. In the following section, we propose an alternative solution that allows to drastically reduce the needed amount of training data.

4 Proposed Approach

One way to magnify the training signal for contextual parameters is to break intra-sentential syntactic links by splitting the sentence in two segments. In this way, a word that was disambiguated by the system by looking at its neighbours within the sentence, now requires contextual information from the other segment in order to be correctly translated. Within the framework of MT, if we split

Algorithm 1: Split parallel corpus

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

the source sentence, its corresponding reference has to be split too. The proposed approach consists in 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 than 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_1 for translating S_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 3rd-person neuter in French). The antecedent “a project", which belongs to 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 wheresplit function returns the token indices m_S and m_T of S^i and T^i, respectively, where we need to split. In this work, we propose two possible definitions of this function.

Middle-split. The most simple strategy is to split both the source and the target sentence in the middle. In this case, we can take $\text{wheresplit} = \text{middlesplit}(S^i, T^i)$, which simply returns $m_S = \lfloor len(S^i)/2 \rfloor$ and $m_T = \lfloor len(T^i)/2 \rfloor$.

Aligned-splits. With the *middle-split* 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

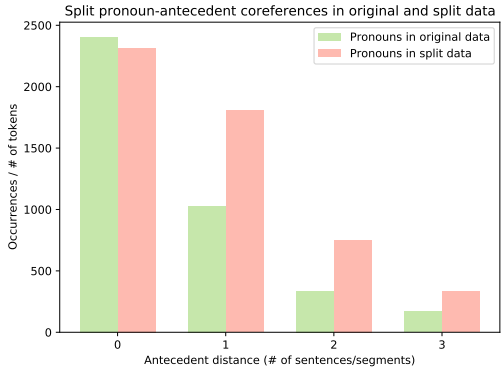

Figure 3: The number of antecedents of anaphoric pronouns at a given distance in terms of sentences/segments, normalized by the number of tokens that the model needs to attend for resolving the coreference.

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 ends up in the other segment. As a solution, we can calculate word alignments A^i and use them to inform our splitting strategy by setting $\text{wheresplit} = \text{alignedsplit}(S^i, T^i, A^i)$, where alignedsplit consists in splitting each sentence close to the middle, while trying to avoid to separate aligned words in different segments. Clearly, the proposed approach 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 $j = 1, 2$. For more details, we refer to the appendix and to our implementation.[1]

4.2 Impact of splitting on training signal

In order to give an explicit picture of how and why contextual training signal augments when splitting sentences, 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 the sentences of a document augments the contextual training signal 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

[1] Which will be soon open-sourced.

an idea of the magnitude of this effect, we calculated the percentage of the sentences where the splitting method breaks the syntactic link 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 for those of punctuation, we find that 84.8% of the sentences contain a dependency that will be broken. These cases add up to the many other cases of broken dependencies, 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 pronomial anaphora more dense in the set of 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.

4.3 Zero-resource context-aware learning

Another application of the split approach is that of training context-aware models for language pairs where document-level parallel data are not available. We call this setting *zero-resource*. If we only have access to a sentence-level corpus C, we can split it so that $\mathcal{C}_D$ is a document-level corpus where each "document" D_i is comprised of two parallel segments $D_i = \{(S^{i,1}, T^{i,1}), (S^{i,2}, T^{i,2})\}$. In other words, the splitting turns each sentence in a standalone document of two segments, with which we can train a context-aware model that looks at one segment back (or ahead). Such model could be used to translate documents in a language where sentence-level MT was the only alternative. We provide a proof of concept of this in Section 6.2.1.

	News	Europarl	TED	Total
IWSLT	x	x	0.20/0.23	0.20/0.23
NEI	0.26/0.17	1.82/1.92	0.20/0.23	2.28/2.32

Table 1: Millions of En→De/Fr sentence pairs used for training context-aware models.

5 Experimental setup

5.1 Data

We conduct experiments for two language pairs: English→German and English→French. 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 from the WMT17[2] En→De news translation shared task (∼5.2M sentences), and WMT2014[3] En→Fr (∼35.8M sentences). Then, sentence-level baselines are fine-tuned on TED talks subtitles from IWSLT17 (Cettolo et al., 2012), in both languages. We train context-aware models with two document-level data settings:

Low resource setting - IWSLT: consisting of TED subtitles from IWSLT17, for both En→De/Fr;

High resource setting - NEI: consisting of three document-level corpora (i) News-Commentary-v12 for En→De and News-Commentary-v9 for En→Fr; (ii) Europarl-v7, for both En→De/Fr; (iii) IWSLT17's TED talks, for both En→De/Fr.

The resulting size of training sets after pre-processing is reported in Table 1. For all models, we use IWSLT17's test sets 2011-2014 as development set, and 2015 for evaluation. Pre-processing details can be found in the appendix.

5.2 Evaluation

Besides evaluating models on the translation of TED talks subtitles with case-sensitive detokenized BLEU (Papineni et al., 2002),[4] we employ two *out-of-domain* contrastive test suites for the evaluation of anaphoric pronoun translation.

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

[2]http://www.statmt.org/wmt17/translation-task.html
[3]http://www.statmt.org/wmt14/translation-task.html
[4]Moses' `multi-bleu-detok` (Koehn et al., 2007).

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 (Lison et al., 2018), completely analogous to the previous one but focused on the translation of two English pronouns: *it* and *they*. The test set 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 Transformer baseline, pre-trained on WMT data and then finetuned on IWSLT. The model architecture follows the *Transformer-base* introduced 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.[5]

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). Context-aware models are trained following *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_S) is trained (4.7M over a total of 65.2M). Details can be found in the appendix.

6 Results and Analysis

6.1 Training with more data

In this section we provide evidence about the inefficacy of training contextual parameters on small document-level data, while larger datasets do the job. In the first block of Table 2 we report performance of context-aware models trained on IWSLT (*-orig* models), and compare them with the same models trained on NEI ($\times 10$ training size). When trained on IWSLT alone, contextual models *K1-orig* and *K3-orig* display good BLEU on the test set, generally without degradation with respect to the baseline *K0*, and with some improvements in the case of *K1-orig*. However, such marginal improvements in BLEU are difficult to interpret, as they do not necessarily correspond to better translation (Freitag et al., 2020). Similarly, accuracy on the contrastive test sets increases marginally over baseline, if at all. *K1* even show a degradation of performance over the 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). However, *K1* and *K3* show a substantial improvement in their context-modeling capabilities when trained on larger data (*-fromnei*), gaining 3.5+ accuracy points each over the baseline (+11.38 for *K1*), in both language pairs. On the other hand, BLEU improves of a few decimal points only, showing its limits to measure improvements in context-aware translation. These results would suggest that multi-encoder models need to be trained with more data before comparing them with alternative approaches. However, this is not ideal because it requires more time and computation for every experiment. Indeed, the last column of Table 2 shows that a model trained on NEI requires almost $\times 10$ the time needed by a model trained on IWSLT.

6.2 Splitting data

Here we show that the proposed splitting approach is a more convenient alternative than using large training data. The second block of Table 2 displays performance of models that have been trained on the IWSLT data, after having split sentences in two segments following the *middle-split* method with $l_{min} = 7$. For this training setting, we also split the validation set, while the test set is always kept in its original form. Beside the sentence-level pre-training on WMT, *-split* models are trained on IWSLTsplit alone prior to evaluation, while *-fromsplit* models are also finetuned on the original, non-split data, after training on IWSLTsplit. The *middle-split* method is very effective, as all models

[5] 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.

Model	Contextual Training	En→De			En→Fr			
		BLEU	acc	Δacc	BLEU	acc	Δacc	Hours
K0	-	32.97	46.37	-	41.44	79.46	-	0
K1-orig	IWSLT	33.00	47.08	-	41.87	79.24	-	2.4
K3-orig	IWSLT	32.86	46.48	-	41.32	80.53	-	2.8
K1-fromnei	NEI	**33.16**	**57.75**	-	41.49	**84.32**	-	19.1
K3-fromnei	NEI	33.10	51.14	-	**41.73**	82.94	-	21.5
K1-split	IWSLTsplit	33.07	51.12	+1.42	41.54	81.44	+0.49	4.2
K3-split	IWSLTsplit	33.25	53.98	-0.68	**41.94**	83.44	+1.48	4.5
K1-fromsplit	IWSLTsplit+IWSLT	**33.44**	**60.21**	+0.69	41.78	84.06	+0.27	6.5
K3-fromsplit	IWSLTsplit+IWSLT	33.36	56.22	-1.38	41.68	**85.50**	+0.20	6.8
K1-zerores	IWSLTsplit as if sentence-level	31.41	50.89	-	39.62	82.35	-	2.4

Table 2: BLEU scores on the in-domain IWSLT test set and accuracies (in %) on the ContraPro test sets. The last column reports the training time spent on the document-level MT task by En→Fr models. All differences in accuracy between any model and the baselines (*K0* and *-orig*) are statistically significant with $p < 0.01$.[6] Δacc measures the change in accuracy when IWSLT is split with the *aligned-split* method instead of *middle-split*.

show strong improvements over the baseline in terms of accuracy on the test suites. The *-fromsplit* approach proves to be the most advantageous. This was expected, since providing the model with both original and split sentences (instead of split alone) increases the number of cases that require context. The result of *-fromsplit* models is striking, because they manage to outperform models trained on NEI with one-tenth of the data (the sole exception is *K1-fromsplit* on En→Fr). In addition, we see from last column of Table 2, that the proposed approach is efficient, as it cuts training time by a factor of 3 when training *-fromsplit* instead of *-fromnei*.

6.2.1 Proof of Concept: Zero-Resource

In this section we consider IWSLT as if it was a sentence-level parallel corpus, where each sentence is a stand-alone document. It follows that, after splitting, *K1* can look at its past context only when translating the second segment of each split sentence, $S^{i,2}$, but not for $S^{i,1}$. The last row of Table 2 shows that a *K1* model trained in this fashion can indeed learn to model context more effectively than both *K0* and *-orig* baselines. However, it also displays a degradation in terms of BLEU. This is probably related to the substantial difference in document-length between the training and the evaluation setting, which could be offset, for example, by adding to the sentence-level split corpus some (longer) back-translated monolingual documents (Sugiyama and Yoshinaga, 2019).

[6] According to the paired McNemar test (McNemar, 1947).

6.2.2 Impact of the splitting method

Following Section 4.1, we study whether a better alignment of source and reference segments is advantageous to our objective of learning contextual parameters. In our experiments, we use the open-source tool *fast_align* (Dyer et al., 2013), which learns token alignments in an unsupervised fashion. After applying the *aligned-split* method (always with $l_{min} = 7$) to IWSLT, we train context-aware multi-encoder models on it and measure their accuracy on ContraPro. We report in Table 2 the improvements Δacc achieved with this method compared to the *middle-split* one. Alignments prove to be useful in most cases, although we measure a degradation of performance for *K3* over the En-De language pair. In general, the improvements over the simpler *middle-split* method are marginal, and thus may be worth the added complexity and the additional computational cost for alignment only for some specific language pairs and context-aware models. In fact, there are at least two factors to take into account while making this choice. First, the accuracy of alignment tools is language-dependent. E.g., *fast_align* is stronger on En-Fr than En-De (Ho and Yvon, 2019). Second, if we favor the alignment between source and target when deciding where to split, we might be doing it at the cost of breaking less syntactic links into separate segments, which is harmful with respect to the objective of the splitting method.

6.3 Ablation: shuffling context

As an ablation study, we want to verify that the proposed approach improves learning by making the context-aware model to rely on its modeling

Model	Contextual Training	En→De BLEU	En→De acc	En→Fr BLEU	En→Fr acc
K0	-	32.08	45.00	40.92	79.70
Zhang2018	IWSLT	31.03	42.60	**40.95**	59.00
*Tu2018**	IWSLT	**32.10**	45.20	40.91	79.70
Concat21	IWSLT	31.84	48.00	40.67	80.90
*Concat22**	IWSLT	30.89	**70.80**	40.57	**83.20**
K1-fromsplit	IWSLTsplit+IWSLT	**33.44**	**60.21**	41.78	84.06
K3-fromsplit	IWSLTsplit+IWSLT	33.36	56.22	41.68	**85.50**
K1-fromneisplit	NEIsplit+IWSLT	32.82	**61.09**	41.81	84.17
K3-fromneisplit	NEIsplit+IWSLT	**33.07**	59.56	**41.91**	**85.66**

Table 3: Comparison between Lopes et al. (2020)'s evaluation of context-aware models (1st block), our models trained on split IWSLT (2nd), and split NEI (3rd). We report BLEU scores on IWSLT and accuracies (in %) on ContraPro. * : these models have the advantage of accessing target-side context, hence comparison is unfair.

Model	En→De acc (Δacc)	En→Fr acc (Δacc)
K0	46.37 (-0.00)	79.46 (-0.00)
K1-orig	46.70 (-0.35)	79.05 (-0.19)
K3-orig	46.21 (-0.27)	79.24 (-1.29)
K1-fromnei	47.72 (-10.03)	78.75 (-5.57)
K3-fromnei	46.85 (-4.29)	78.86 (-4.08)
K1-fromsplit	47.60 (-12.61)	78.94 (-5.12)
K3-fromsplit	47.96 (-8.26)	79.05 (-6.45)

Table 4: Accuracy results (and their changes) on contrastive examples when using inconsistent context.

of the context. Table 4 shows performances on ContraPro after having shuffled the context of every contrastive example in the test set (c.f. Scherrer et al. (2019)), and the delta w.r.t. the results with consistent context presented in Table 2. A random context is incoherent with the current sentence in many cases, and thus misleading for a context-aware system. Indeed, *-fromsplit* models display a significant drop in performance when they are evaluated with inconsistent context, as well as *-fromnei* models, 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 obtains a similar performance to *K0*. In other words, the splitting method does not produce models that are over reliant on context.

6.4 Comparison with the literature

In Table 3 we compare our results with those reported by Lopes et al. (2020), who evaluated various context-aware approaches on the same IWSLT setting as ours. They also adopt a large sentence-level pre-train, a two-step training strategy with frozen *Transformer-base* parameters for multi-encoder models, and the exact same evaluation setting as ours. We report their results for a baseline ***K0*** like ours; a multi-encoder ***Zhang2018*** with both *inside integration* and *outside integration*, that encodes three past source sentences (Zhang et al., 2018); a multi-encoder ***Tu2018*** with *inside integration*, that encodes all past context, both source-side and target-side; a *many-to-one* (2-to-1) single-encoder approach ***Concat21***; a *many-to-many* (2-to-2) single-encoder approach ***Concat22***.

Lopes et al. (2020)'s multi-encoder models perform poorly or even lag behind *K0*, as it happens for *-orig* models when trained on IWSLT original data alone. Instead, concatenation approaches are very strong, because they do not have extra contextual parameters to train and simply finetune the same sentence-level Transformer architecture on the context-aware task. Our splitting strategy proves to be very effective, since *-fromsplit* models outperform *Concat21* using the same amount of training data. Moreover, *-fromsplit* multi-encoder models beat the strong *Concat22* on En→Fr, which has the non-negligible advantage of using also the target-side context, on top of the source-side. Finally, we try to train the contextual parameters of *K1* and *K3* on the much larger NEIsplit, where sentences are split according to the *middle-split* method, and then fine-tune on IWSLT for comparison. Results reported in Table 3 show further improvements on the anaphoric pronoun translation set, with no BLEU degradation with respect to the baseline. Nonetheless, considering the large increase in training data, these improvements seem to be marginal. Diminishing returns with the increase of training data could be related to the fact that context-aware models used for the experiments are limited to the source-side context, while target-

side context is sometimes key for solving some examples in contrastive test suites.

7 Conclusions

Multi-encoder models are a broad family of context-aware NMT models. In this work we have shown that their contextual parameters are difficult to train because the training signal is sparse. We have proposed an approach based on splitting the training sentences, which augments the training signal by breaking syntactic links. After having analysed the implications of our approach on discourse phenomena, we have shown empirically that it allows to learn contextual parameters very effectively, and more efficiently than by simply increasing the amount of training data. Our approach also proves to be a viable option to zero-resource context-aware machine translation.

We leave to future work experiments with a syntactically motivated splitting method, which would use the syntactic information to be sure to break a syntactic link every time it is possible.

8 Acknowledgements

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.

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.

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.

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.

Anh Khoa Ngo Ho and François Yvon. 2019. Neural Baselines for Word Alignments. Proceedings of 16th International Workshop on Spoken Language Translation (IWSLT).

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.

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, Łukasz 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.

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, Łukasz 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.

A Aligned-split method

In this section, we provide more details on the *aligned-split* method. As already mentioned, we use $\text{wheresplit} = \text{alignedsplit}(S^i, T^i, A^i)$, 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. 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, alignedsplit falls back to middlesplit. We refer to our open-source implementation for the full details.

B Experimental Setup

B.1 Data preprocessing

All datasets are tokenized with the Moses toolkit (Koehn et al., 2007), further cleaned by removing long sentences, and byte pair encoded (Sennrich et al., 2016) 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 NEI 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.

B.2 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,[7] 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. We train models with the optimizer configuration and learning rate schedule described in Vaswani et al. (2017). The maximum learning rate is 0.0005 for models trained on NEI, 0.003 for all the others. We use label smoothing with an epsilon value of 0.1 (Pereyra et al., 2017). Since the models are pre-trained on a large amount of parallel data (WMT), the dropout is 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. The only hyperparameter that changes across training settings is length penalty (P_{len}). In particular, we adopt a length penalty of 0.6 for all models trained on En→De data ($P_{len} < 1$ favors shorter sentences). For En→Fr models, the length penalty changes according to the training data employed: $P_{len} = 0.6$ for models trained on original, non-split data, i.e., *-orig* and *-fromnei* models; $P_{len} = 1.3$ for models trained on split-data alone, i.e., *-split* models ($P_{len} > 1$ favors longer sentences); and $P_{len} = 1.0$ for models that have been trained on both split and non-split data, i.e., *-fromsplit* models. This choice is intuitive considering that split corpora have a shorter average sentence length.

C Results Analysis

C.1 Accuracy by antecedent distance

In this section, we want to investigate more in details the performance of the proposed approach on the accuracy of ambiguous pronoun translation. We report in Table 5 the accuracy on En→De ContraPro, detailed by varying antecedent distance. We notice that all the improvements achieved by *-fromsplit* 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-fromsplit* is very strong in translating pronouns with antecedent distance $d = 1$, surpassing *K0* and *-orig*

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

	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-orig	47.05	**82.5**	33.9	45.3	48.0	69.9
K3-orig	46.48	82.4	32.8	45.0	48.9	71.7
K1-fromnei	**57.75**	82.1	**52.1**	44.7	48.7	**72.9**
K3-fromnei	51.14	81.0	39.9	**50.1**	**53.2**	70.1
K1-fromsplit	**60.21**	81.1	**56.5**	**44.9**	48.7	**73.3**
K3-fromsplit	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 5: Accuracy(%) of 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.

Model	**En→De ΔBLEU**	**En→Fr ΔBLEU**
K0	0.00	0.00
K1-orig	+0.06	-0.12
K3-orig	-0.13	+0.15
K1-fromnei	-0.17	-0.22
K3-fromnei	-0.30	-0.37
K1-fromsplit	-0.34	-0.14
K3-fromsplit	-0.31	-0.13

Table 6: Changes in BLEU results when the test set is shuffled, making context incoherent.

baselines by 22+ points of accuracy. Similarly, *K3-fromsplit* surpass 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-fromsplit* lacks behind *K1-fromsplit* on $d = 1$. On one side, this could be explained by the fact that *K1-fromsplit* 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-fromsplit* might perform worse on $d = 1$ than *K1-fromsplit* 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-fromsplit* has the highest average result (reported in "Total").

C.2 Ablation: shuffling context

We have shown that well trained context-aware models display a significant drop in accuracy of ambiguous pronoun translation when they are evaluated with inconsistent context, which suggests that they are heavily relying on context for solving ambiguous translations. We complement these findings with the equivalent experiment on the IWSLT test set. We shuffle the test set in order to make context incoherent for most of the sentences in the documents, and then we evaluate the translation output of context-aware models with BLEU. Results reported in Table 6 display a small degradation in average quality performance as measured by BLEU. However, the changes are so small that are probably negligible, showing once again that BLEU is ill-equipped to measuring improvements in document-level translation.