Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3026–3037
Marseille, 20-25 June 2022

TANDO: A Corpus for Document-level Machine Translation

Harritxu Gete[1*], Thierry Etchegoyhen[1*], David Ponce[1], Gorka Labaka[2], Nora Aranberri[2], Ander Corral[3], Xabier Saralegi[3], Igor Ellakuria Santos[4], Maite Martin[5]

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]IXA taldea, University of the Basque Country, [3]Elhuyar, [4]ISEA, [5]Ametzagaiña
[1]{hgete, tetchegoyhen, adponce}@vicomtech.org, [2]{gorka.labaka, nora.aranberri}@ehu.eus,
[3]{x.saralegi, a.corral}@elhuyar.eus, [4]iellakuria@isea.eus, [5]maite@adur.com

Abstract

Document-level Neural Machine Translation aims to increase the quality of neural translation models by taking into account contextual information. Properly modelling information beyond the sentence level can result in improved machine translation output in terms of coherence, cohesion and consistency. Suitable corpora for context-level modelling are necessary to both train and evaluate context-aware systems, but are still relatively scarce. In this work we describe TANDO, a document-level corpus for the under-resourced Basque-Spanish language pair, which we share with the scientific community. The corpus is composed of parallel data from three different domains and has been prepared with context-level information. Additionally, the corpus includes contrastive test sets for fine-grained evaluations of gender and register contextual phenomena on both source and target language sides. To establish the usefulness of the corpus, we trained and evaluated baseline Transformer models and context-aware variants based on context concatenation. Our results indicate that the corpus is suitable for fine-grained evaluation of document-level machine translation systems.

Keywords: Document-level Machine Translation, Parallel Corpus, Contrastive tests, Basque, Spanish

1. Introduction

Neural machine translation (NMT) typically performs translation by considering sentences in isolation, ignoring discursive phenomena (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017). Despite the improvements achieved at the sentence level, inconsistencies due to the lack of extra-sentential information raise important issues in terms of coherence, cohesion and consistency (Läubli et al., 2018; Toral et al., 2018; Voita et al., 2019b). Context-aware NMT is currently an active area of research and has notably been included among the shared tasks at the WMT conference series in recent years (Barrault et al., 2019; Barrault et al., 2020).

Despite initial advances in context-aware models that tackle the generation of consistent and coherent translations in context, advances in this field are hampered by two main issues. First, parallel corpora that include contextual information are relatively scarce (Liu and Zhang, 2020). Second, standard machine translation (MT) metrics, such as BLEU (Papineni et al., 2002), are usually not sensitive to improvements at the extra-sentential level (Wong and Kit, 2012; Sennrich, 2017).

This work aims to address both of the aforementioned deficiencies. Regarding the first issue, we prepared a novel parallel dataset that includes contextual information for the Basque-Spanish language pair, covering different domains such as literature, subtitles and varied news. To address the second issue, we prepared several contrastive test sets aimed at evaluating the ability of NMT models to handle common contextual errors in Basque-Spanish translation, namely gender and register selection. Our contrastive test sets include cases where the relevant contextual information is either in the source or in the target language, to address the need for contextual coherence in both cases. The TANDO corpus, which includes both parallel and contrastive datasets, is shared with the community for research purposes.[1].

To assess the validity of the corpora, and to provide reference results for future studies, we trained and evaluated context-aware model baselines on the corpora. For the purposes of this study, we selected a simple yet efficient approach, namely context concatenation (Tiedemann and Scherrer, 2017), and explored several variants that exploit source and target contextual information.

Our contributions can thus be summarised as follows:

- A multi-domain corpus for Basque-Spanish, suitable for context-aware NMT.
- Novel contrastive datasets for fine-grained evaluations of contextual phenomena.
- Baseline evaluation results, with context-aware models trained and evaluated on the prepared datasets.

The remainder of the paper is organised as follows: Section 2 presents related work; Section 3 describes the corpora and preparation methodology; Section 4 presents the different NMT models trained and evaluated on the corpora; Section 5 presents experimental results; Section 6 draws conclusions from this work.

[*]These authors contributed equally to this work

[1]The corpus is available under a Creative Commons CC-BY-NC-SA 4.0 license and can be downloaded at the following address: https://github.com/Vicomtech/tando

2. Related Work

A variety of studies have tackled document-level approaches within the framework of statistical machine translation (Hardmeier and Federico, 2010; Tiedemann, 2010; Gong et al., 2011; Xiao et al., 2011; Webber, 2014). Within NMT, context modelling has received increasing attention in recent years, with a significant number of context-aware models reporting improvements over non-contextual baselines (Jean et al., 2017a; Wang et al., 2017; Tiedemann and Scherrer, 2017; Miculicich et al., 2018; Maruf and Haffari, 2018; Zhang et al., 2018; Junczys-Dowmunt, 2019; Voita et al., 2019a; Maruf et al., 2019; Xu et al., 2020; Ma et al., 2020; Jauregi Unanue et al., 2020; Mansimov et al., 2021).

Determining the actual impact of context modelling in NMT is not straightforward, in particular, because improvements in translation metrics may be attributed to context-driven regularisation, which can act as a noise generator, especially with small-scale data (Kim et al., 2019; Li et al., 2020). However, other recent studies have shown that contextual information can indeed help capture discursive phenomena that cannot be modelled at the sentence level (Liu and Zhang, 2020; Rikters and Nakazawa, 2021; Xu et al., 2021; Mansimov et al., 2021).

The simplest method to perform translation at the document level is by concatenating context sentences to the sentence to be translated (Agrawal et al., 2018; Tiedemann and Scherrer, 2017). This simple technique achieves performances comparable to that of more sophisticated approaches, in particular, in high-resource scenarios (Lopes et al., 2020). An alternative approach is to take advantage of sentence-level models by refining their translations using reinforcement learning (Xiong et al., 2019; Mansimov et al., 2021) or by adding a component that learns to post-edit errors produced by the context-agnostic system (Voita et al., 2019a). Finally, several studies centre on modelling contextual information by modifying the NMT architecture. These include multi-encoder approaches, which encode context sentences separately via dedicated encoders (Jean et al., 2017b; Zhang et al., 2018; Li et al., 2020). Along these lines, hierarchical architectures have been explored by Wang et al. (2017) and Tan et al. (2019). Dynamic memory components, which store information from previous translations, have also been proposed for context modelling in NMT (Tu et al., 2018; Kuang et al., 2018; Maruf and Haffari, 2018). More recently, Xu et al. (2021) represent a complete document as a graph connecting relevant contexts and Morishita et al. (2021) proposed to include contextual information in a mini-batch.

Most existing models establish a fixed contextual window of preceding or following sentences (Zhang et al., 2018; Voita et al., 2018; Voita et al., 2019b; Yang et al., 2019; Rikters and Nakazawa, 2021), as long-distance context can be challenging to model (Junczys-Dowmunt, 2019; Tan et al., 2019; Zheng et al., 2020; Sun et al., 2020). Additionally, experiments in Kim et al. (2019) and Fernandes et al. (2021) show that only a limited portion of the context may be useful, with long-range context resulting in degraded translation quality. Several studies have noted that only a portion of the context is relevant at a given time, and focused on selecting relevant contextual data irrespective of distance to the sentence to be translated (Jean and Cho, 2019; Kimura et al., 2019; Maruf et al., 2019; Xu et al., 2021). Along similar lines, Kang et al. (2020) explicitly select a variable number of context sentences for each sentence.

Another relevant issue is the use of source or target context. While Yang et al. (2019) and Zhang et al. (2018) only use source-side context, for instance, other approaches also include target hypotheses (Bawden et al., 2018; Agrawal et al., 2018; Maruf and Haffari, 2018; Yamagishi and Komachi, 2019; Xu et al., 2021).

To properly assess the usefulness of contextual information for translation, specific types of test sets have been introduced, in addition to standard reference metrics. Contrastive test sets (Sennrich, 2017) thus measure the accuracy of models on a ranking task between correct and incorrect translations on data that feature elements sensitive to contextual information. Several contrastive tests focus on pronoun translation, for instance, ContraPro (Müller et al., 2018) for English-German, and the datasets described in Lopes et al. (2020) and Bawden et al. (2018) for English-French. Other test benchmarks target phenomena such as politeness, consistency, ellipsis and lexical cohesion, in different language pairs: Bawden et al. (2018) for English-French, Voita et al. (2019b) for English-Russian, Nagata and Morishita (2020) for Japanese-English and Rios Gonzales et al. (2017) for German-English, for instance.

3. Corpora

Our document-level corpora centre on the Basque-Spanish language pair and include data from different domains: literature, media content subtitles, proceedings of the Basque Parliament, and news.

We first created train, test and development partitions from the available data in the aforementioned domains, as described in Section 3.1. Additionally, we prepared constrastive datasets focusing on specific contextual phenomena involving gender and register (Section 3.2), to support the evaluation of context-aware models.

3.1. Base Corpora

For the base corpora, three domains were selected as they contained sufficient data to prepare datasets with contextual information, namely EhuHAC, a collection of literary documents; EiTB, a collection of comparable news; and OpenSubs, a collection of subtitles.

The detailed characteristics of each domain are provided in the next sections, along with domain-specific

processes. For all three domains, the goal was to extract blocks of contiguous aligned sentences that could form part of a large multi-domain corpus, suitable to train and evaluate context-aware NMT models. The overall process was similar across domains and consisted of the following steps:

- Document alignment, via either metadata information or content-based alignment processes.
- For each aligned document pair:
  - Sentence splitting and text normalisation (tokenisation and truecasing), performed with scripts from the Moses toolkit (Koehn et al., 2007).
  - Sentence alignment, with the relevant tools for the considered data.
  - Filtering of sentence pairs under a specific alignment threshold.
- Context block preparation:
  - A valid context block consists of a sequence of n aligned sentence pairs, $n > 1$.
  - Each discarded sentence-aligned pair resets a new contextual block, i.e. no context block could contain non-contiguous sequences of aligned sentence pairs.

Each corpus prepared from data in the three specified domains was then split into train, development and test partitions via uniform sampling, taking into account the specifics of each domain, as described in the following sections.

3.1.1. EhuHAC

The EHUHAC corpus is made up of a collection of 137 classic books and includes their translations into Basque, English, French and Spanish[2]. The collection includes classic works of philosophy ("Metaphysics" by Aristotle), literature ("The Lady of the Camellias" by Alexandre Dumas), and the Bible[3]. The corpus was compiled and aligned at sentence level by the Basque Language Institute of the University of the Basque Country in 2015 (Sarasola et al., 2015). Based on this alignment, the parallel Spanish-Basque corpus was processed to adapt it to the needs of contextual machine translation.

The original corpus maintains context information, considering the book in its entirety as a single document, but this range of context was excessively broad for our purposes. Since the books consist of chapters, sections and subsections, the original documents were divided using heuristics that searched for section headings. Specifically, a document division was included each time a sentence was found that began with the word "chapter", "section", "part" or "volume" (or its equivalents in Basque and Spanish), followed by a number.

Additionally, it was necessary to apply a filter to eliminate misaligned parallel sentences. For this purpose, the sentence pairs that met one of the following heuristics were filtered out: (1) pairs for which neither the source nor the target sentence existed or (2) pairs for which the length ratio between sentences exceeded 1:3.

After this preprocessing, the sentences were divided into training, development and test sets. For this purpose, complete documents consisting of a minimum of 10 sentences and a maximum of 50 sentences were randomly selected until a minimum of 1000 sentences were collected for development and 2000 for testing. The remaining data were kept in the training set.

3.1.2. EiTB

The EITB corpus is composed of news independently produced in Basque and Spanish by the Basque public broadcaster EiTB[4]. The corpus is strongly comparable and has been exploited as a source of parallel data for the under-resourced Basque-Spanish language pair (Etchegoyhen et al., 2016).

The original documents were first aligned with DOCAL (Etchegoyhen and Azpeitia, 2016a), a lightweight content-based document aligner with high accuracy across domains and language pairs (Azpeitia and Etchegoyhen, 2019).

Since the data are comparable in nature, sentence alignment was performed on each document pair with a dedicated tool, namely the STACC aligner (Etchegoyhen and Azpeitia, 2016b), in its version with lexical weighting (Azpeitia et al., 2017), using an alignment threshold of 0.15.

News documents in the corpus were short on average, which led to discarding significant portions of the original data as a consequence of constraining contextual blocks to contain only contiguous alignments.

Test and development sets were prepared with additional constraints, to maximise their utility. First, all alignments containing less than 50% of alphanumeric characters were discarded to remove pairs such as sports results. Additionally, only contextual blocks of at least 5 contiguous aligned sentence pairs were selected for these datasets to ensure minimal context representation at validation and testing time.

3.1.3. OpenSubtitles

The OpenSubtitles platform[5] is a free collaborative platform for creating and sharing series and movies subtitles. It is a growing platform and constitutes one of the largest freely available subtitle databases, with more than 3 million subtitles in more than 60 languages. From a linguistic perspective, subtitles cover a large number of genres (series, movies, documentaries,

[2] https://www.ehu.eus/ehg/hac/

[3] Full list of books in https://www.ehu.eus/ehg/hac/liburua

[4] Euskal Irrati Telebista: https://www.eitb.eus

[5] https://www.opensubtitles.org/

DOMAIN	TRAIN (MIN/MAX/AVG)	DEV (MIN/MAX/AVG)	TEST (MIN/MAX/AVG)
EHUHAC	513,613 (2/160/9)	1009 (10/49/20)	2024 (10/49/19)
EITB	472,963 (2/198/3)	1027 (5/14/6)	2017 (5/14/6)
OPENSUBS	785,478 (10/50/42)	1037 (25/50/46)	2085 (10/50/42)
MERGED	1,753,726 (2/198/8)	3051 (5/50/12)	6078 (5/50/13)

Table 1: Corpora statistics (#sentence pairs). MIN, MAX and AVG indicate the minimum, maximum and average context sizes, respectively

shows...) with colloquial, informal, formal or narrative language. Due to these characteristics, subtitle data may feature interesting phenomena for context-aware machine translation, such as deixis and ellipsis.

The OpenSubtitles2016 parallel corpus (Lison and Tiedemann, 2016), has been generated from the subtitles of the OpenSubtitles platform by preprocessing and aligning subtitles for different language pairs. For our purposes, we used the parallel corpus corresponding to the Basque-Spanish language pair in the OPUS repository[6] (Tiedemann, 2012), and used the information related to the original documents to establish the context of the sentence pairs.

A further analysis of the dataset indicated noise in the original alignments and we therefore applied several filtering rules to avoid spurious sentence pairs, thus removing pairs where either the source or the target was missing, or the length ratio between sentences exceeded 1:3, or the sentences consisted only of punctuation symbols.

After applying these filtering rules, contextual flow was broken in some of the documents, and therefore, they were then divided into sub-documents of smaller context. All contexts below 10 pairs were discarded and those above 50 sentences were split. For the dev and test sets, randomly selected context blocks were merged until 1000 and 2000 sentences were obtained, respectively. The rest of the documents were used as training corpus.

3.1.4. Merged Corpus

The train, test and dev partitions from each of the three domains were merged via simple concatenation to obtain a joint corpus. Table 1 summarises the final data in terms of parallel sentences.

With over 1.7 million Basque-Spanish parallel sentences overall, the corpus provides a solid basis to train and compare NMT models with context information on this language pair. The merged test sets, which contain approximately 6000 sentence pairs, enable the computation of reference metrics over significant samples that represent the three selected domains.

3.2. Contrastive Corpora

To support a fine-grained evaluation of context-aware models, in addition to the test sets described in the previous section, we prepared contrastive datasets aimed at evaluating the following phenomena for Basque to Spanish translation:

- Pronouns are marked for gender in Spanish, but not in Basque (e.g., zuek → vosotras (fem.)/vosotros (masc.)). Context information is therefore necessary to determine pronoun references in Basque and generate correct translations. A similar phenomenon occurs with some adjectives (polita → bonita (fem.)/bonito (masc.)) and nouns (ikasle → alumna (fem.)/alumno (masc.)).
- In Basque, although there are different forms for formal and informal register, the use of formal expressions is widespread for both registers. This can lead to the use of an incorrect register when translating into Spanish, where both forms are clearly marked.

Given the lack of sufficient data manifesting the above phenomena in the test sets described in Section 3.1, additional corpora were mined to prepare the contrastive datasets, namely monolingual data from collected books, TED talks, and proceedings of the Basque Parliament. Once context block samples had been collected, along the lines described above, the monolingual datasets were machine-translated with high-quality generic in-house NMT models, professionally post-edited to ensure the final quality of the translations, and manually revised to verify the accuracy of contextual information in the final datasets.

In the literature domain, we mined books collected from the Gutenberg and Elejandria repositories[7], in Basque or Spanish for the former, and Spanish for the latter. For TED talks, we used the Basque-Spanish 2020 v1 dataset (Reimers and Gurevych, 2020) from the OPUS repository, and mined relevant examples in each language separately. Finally, parliamentary speech transcriptions were collected from the Basque Parliament plenary sessions.[8] PDF files were crawled to find relevant examples and text normalisation was performed as described for the base corpora, on all collected data.

The contrastive datasets reflecting the aforementioned phenomena consist of two main separate sets for

[6]https://opus.nlpl.eu/

[7]Respectively located at: https://www.gutenberg.org/ and https://www.elejandria.com/

[8]Available online at: https://www.legebiltzarra.eus/

Basque to Spanish translation, each consisting of context blocks from different domains that include contrastive pairs and previous sentences with context information.

In the first set (hereafter, GDR-SRC+TGT), the disambiguating information for the contrastive pairs is present in both the source and target languages, as illustrated in Example 1 of Table 2. For this dataset, only examples relevant for gender selection were included, due to the difficulties in obtaining examples where the informal register is marked in the source.

The second set (hereafter, COH-TGT) was created to evaluate cases where, despite the absence in the source language of the necessary information to make a correct selection of gender or register, the translation must be contextually coherent in the target language (Example 2 in Table 2). For this dataset, both gender and register examples were included.

Each context block in either contrastive set consists of (1) a sentence in Basque to be translated that contains an ambiguous word in terms of gender or register; (2) a context of up to 5 preceding sentences, which provides relevant information to predict the gender or register of the ambiguous word; (3) a reference translation in Spanish; and (4) a contrastive translation, created by switching the gender or register of the target word. Note that, since the contrastive translations must be grammatically correct, the necessary changes were manually made to ensure agreement in terms of gender and/or register within the context.

Context blocks that satisfied the established constraints were identified in the monolingual corpora, in Basque or Spanish, adhering to document boundaries. To find appropriate gender-based contrastive examples, commonly used Spanish pronouns, nouns and adjectives with dual gender forms were mined. In the case of register, pronouns such as *tú* (informal *you*) or *usted* (formal *you*) were mined in Spanish, and verbs in second person in Basque. After manual inspection of the collected data, incorrect matches were removed until 100 examples were obtained for each phenomenon.

In most cases, context blocks feature full contexts, although the final corpora also contain blocks with fewer than 5 context sentences. These cases were left as is in the corpus, as they also represent typical situations where contextual information is reduced, as is the case for the first sentence of a text, for instance.

Antecedent distance was not artificially balanced in order to maintain a relative variety among context blocks and the actual representation in the corpus. Domain representation, as well as feminine/masculine and formal/informal alternations were, however, balanced to mitigate possible gender or register biases in the machine-translated data.

Overall, there were thus 200 context blocks for each of the three domains, 300 for the GDR-SRC+TGT test set, split into 150 masculine and 150 feminine cases, and 300 for the COH-TGT test set, split into 75 for each gender in the target gender test, and 75 for each register, formal and informal, in the target register set.

4. Models

To evaluate the previously described corpora and provide reference results, we trained baseline NMT models and context-aware variants. We describe them in turn in the following sections.

4.1. Baseline models

As baselines, for both translation directions, we trained Transformer-base models (Vaswani et al., 2017) with 6-layer encoders and decoders, feed-forward networks of 2048 units, embeddings vectors of dimension 512, 8 attention heads and a dropout rate between layers of 0.1. All datasets were segmented with BPE (Sennrich et al., 2016), with 30,000 operations, using the fastBPE toolkit[9]. Sentences larger than 100 tokens were filtered from the training set.

The baselines were trained with the MarianNMT toolkit (Junczys-Dowmunt et al., 2018) on 2 GPUs with 11GB of RAM each. Optimisation was performed with Adam (Kingma and Ba, 2015), with $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate was set to increase linearly for the first 16,000 training steps and decrease afterwards proportionally to the inverse square root of the corresponding step. We used a working memory of 6000MB and automatically chose the largest mini-batch that fit the specified memory. The validation data was evaluated every 3500 steps, and the training process ended if there was no improvement in the perplexity of 10 consecutive checkpoints. Embeddings for source, target and output layer were tied.

4.2. Context-aware Models

Our context-aware core approach consisted in an extension of the input, via sentence concatenation, without any change to the architecture of the model (Tiedemann and Scherrer, 2017). This approach was selected for its simplicity and efficiency, as it obtained competitive results against more sophisticated approaches (Lopes et al., 2020).

The extended input includes the context of the previous n sentences, with an additional sentence break token between the context and the current sentence. Different variants were trained by using either source or target language contexts, and either 1 or 5 context sentences.

As in the extended context model of Tiedemann and Scherrer (2017), the output is a single sentence, regardless of the number of input sentences. We discarded variants that allow more than one output sentence to be generated, since they could not be directly evaluated with our contrastive test sets.

The model parameters were initialised with those of the trained baseline models, as this improved results and

[9]https://github.com/glample/fastBPE

EXAMPLE 1	
SOURCE (EU)	Hori nire **arreba**$_{fem}$ da.
	Berak$_{[?]}$ zaindu zituen nire argazkiak.
TARGET (ES)	Esa es mi **hermana**$_{fem}$.
	Ella$_{fem}$ cuidó mis fotos.
EN	That's my **sister**.
	She took care of my photos.
EXAMPLE 2	
SOURCE (EU)	—Begira, Joaquin, **haiek**$_{[?]}$ bezala ezkonduta gaude...
	—**Haiek**$_{[?]}$ bezala, ez, Antonia, haiek bezala, ez!
TARGET (ES)	—Mira, Joaquín, que estamos casados como **ellos**$_{masc}$...
	—¡Como **ellos**$_{masc}$ no, Antonia, como **ellos**$_{masc}$, no!
EN	—Look, Joaquín, we are married like **them**...
	—Not like **them**, Antonia, not like **them** !

Table 2: Examples of discursive phenomena in Basque to Spanish translation

reduced training times, according to preliminary experiments. Hyper-parameters were identical to the ones described in Section 4.1.

5. Results

In this section, we describe and discuss the results obtained with the different NMT models on the parallel test sets (Section 5.1) and on the contrastive test sets (Section 5.2).

5.1. Metrics Results

Metrics results were computed with the SacreBLEU toolkit (Post, 2018) on cased detokenised output, in terms of BLEU (Papineni et al., 2002) and chrF (Popović, 2015), in their default configurations. Statistical significance was computed with paired bootstrap resampling (Koehn, 2004). Results on the domain-specific and merged datasets are shown in Table 3 and Table 4 for Spanish to Basque and Basque to Spanish translation, respectively.[10]

Overall, results were similar in both translation directions. Across domains, the best performing variants used a single previous sentence as context, either on the source side or on the target side. In the latter case, statistically significant improvements were obtained only when using the reference context sentence; the more realistic use of machine-translated output as target context performed on a par with the baseline overall.

Using a larger context of 5 sentences resulted in degraded performance across the board. Considering the better results obtained when using only the previous sentence as context, this degraded performance may be attributed to the fact that the relevant contextual information tends to occur in the previous sentence, with additional contextual information acting as noise. Determining whether this is actually the case would require manual examination of each context block in the parallel test sets, which was beyond the scope of this work. We performed this type of manual analysis on the contrastive sets, as described in the next section.

In terms of specific domains, the largest improvements with the 1:SRC variants were obtained on the EiTB test set, while those obtained with 1:TGT:RF were similar across domains. These results indicate that the characteristics of contextual information may vary depending on the domain, making it necessary to evaluate context-aware models on datasets that allow for separate domain-specific evaluation.

5.2. Contrastive Results

In Table 5, we indicate the percentage of cases where each system selected the correct answer in the source-target gender contrastive test sets. On a par with the metrics results previously described, the use of context information improved markedly over the baseline on this task, in all cases.

Using a single context sentence was also more beneficial overall than using the 5 previous context sentences, as was the case in terms of metrics results, although with smaller differences between context-aware variants than between variants and the baseline. Using the previous sentence in either the source or the target side resulted in minor differences, indicating that sufficient information was available on either side to determine the correct answer. This result was expected for this

[10] We indicate context-aware models with the following naming convention: *context-size:context-source*, where *context-size* denotes the number of context sentences used for the evaluation, and *context-source* indicates either SRC, when the context originates from the source language, and TGT when it originates from the target language. Thus, 5:SRC would denote a model trained and evaluated over blocks of 5 source context sentences. Additionally, for models that use target context, we append the notation RF when the target context used at inference time is composed of the reference translations, and MT when the target context is composed of target sentences as translated by the model; the latter option is meant to evaluate a more realistic inference-time scenario, where target context references are not available.

MODEL	MERGED		EITB		EHUHAC		OPENSUBS	
	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
BASELINE	22.7	54.4	27.5	59.8	15.6	48.0	21.8	50.5
1:SRC	**23.3**†	**54.8**†	**28.2**†	**60.3**†	15.9	**48.4**†	22.3	**51.0**†
1:TGT:RF	**23.1**†	**54.8**†	27.9	60.1	**16.1**†	**48.6**†	**22.6**†	50.9
1:TGT:MT	22.9	54.3	27.9	59.7	15.6	47.9	21.9	50.2
5:SRC	21.6†	52.9†	25.9†	58.1†	15.0†	46.8†	21.4	49.6†
5:TGT:RF	22.2†	53.5†	26.5†	58.7†	15.5	47.3†	22.5	50.5
5:TGT:MT	22.0†	53.5†	26.3†	58.7†	15.3	47.3†	21.7	50.1

Table 3: Metrics results for Spanish to Basque translation; † indicates statistically significant results against the baseline, for $p < 0.05$; best performing systems, significantly better than the baseline and without statistically significant differences between them, are shown in bold

MODEL	MERGED		EITB		EHUHAC		OPENSUBS	
	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
BASELINE	31.2	54.6	38.5	61.7	22.7	47.1	25.5	47.2
1:SRC	**31.7**†	**55.0**†	**39.2**†	**62.2**†	23.1	47.5†	25.4	47.4
1:TGT:RF	**31.9**†	**55.2**†	**39.2**†	**62.2**†	**23.4**†	**47.8**†	**26.1**†	**47.7**†
1:TGT:MT	31.5†	54.8	**38.9**†	61.8	22.9	47.4	25.1	47.0
5:SRC	29.9†	53.4†	36.8†	60.2†	21.9†	46.1†	24.3†	46.1†
5:TGT:RF	29.4†	52.9†	36.0†	59.6†	21.6†	45.7†	24.5†	46.0†
5:TGT:MT	29.1†	52.6†	35.8†	59.3†	21.4†	45.4†	23.7†	45.6†

Table 4: Metrics results for Basque to Spanish translation; † indicates statistically significant results against the baseline, for $p < 0.05$; best performing systems, significantly better than the baseline and without statistically significant differences between them, are shown in bold

test set, where antecedent information is present in both languages.

The higher scores obtained overall when the correct answer was masculine, a marked tendency in these results for the baseline and all model variants, can be attributed to the predominance of this gender in the training sets. Selecting the feminine gender as the correct option proved more difficult for all models, showing some of the limits of the use of context to properly disambiguate gender in the implemented approach.

As shown in Table 6, where the relevant context information is present only in the target side, context-aware models also outperformed the baseline in most cases, although source-context models were worse than the baselines on the Parliament feminine sets, the TED masculine set with a single context sentence, and the Literature feminine set with 5 context sentences. Here too, selected translations on masculine sets were more accurate than on feminine ones. Models based on target information were also more accurate than source-based ones with similar context size, which was expected considering that context information is located on the target side in this test set.

One notable difference with the results obtained on the source-target gender test sets is that the use of a larger context was beneficial overall on this task in most cases; this was the case using either source or target side information. To measure whether the differences between selection results on the GDR-SRC+TGT and COH-TGT sets were due to a difference in the location of the relevant contextual information, we manually analysed the datasets in both cases to locate said information.

The distribution was similar overall, with the following data for the source and target context respectively: 64.67% and 62.00% of cases where the disambiguating information is in the first context sentence; 20.67% and 19.33% in the second sentence; 9.33% and 10.00% in the third; 2.00% and 7.33% in the fourth; 3.33% and 1.33% in the fifth.

This preponderance of disambiguating elements in the first context sentence may have contributed to the slightly better results obtained on the GDR-SRC+TGT test set with models limited to one previous sentence as context, with the remaining context acting as noise on models operating on larger contexts. However, with close to 40% of cases where the relevant information is located beyond the first sentence, the impact of larger contexts is expected to be significant as well. Differences in impact of larger contexts would benefit from further analysis, which we leave for future work.

Finally, the results on the contrastive target register test

MODEL	MERGED	LITERATURE		PARLIAMENT		TED	
		MASC	FEM	MASC	FEM	MASC	FEM
BASELINE	53.67	68.00	32.00	84.00	36.00	76.00	26.00
1:SRC	71.00	80.00	**64.00**	**94.00**	**56.00**	88.00	44.00
1:TGT	**71.33**	**82.00**	**64.00**	92.00	54.00	82.00	**54.00**
5:SRC	69.67	78.00	60.00	90.00	50.00	**90.00**	50.00
5:TGT	66.00	76.00	52.00	90.00	50,00	80.00	48.00

Table 5: Percentage of correct *gender* answers on the GDR-SRC+TGT test set for Basque to Spanish translation

MODEL	MERGED	LITERATURE		PARLIAMENT		TED	
		MASC	FEM	MASC	FEM	MASC	FEM
BASELINE	51.33	76.00	28.00	92.00	12.00	72.00	28.00
1:SRC	52.67	88.00	28.00	**100.00**	4.00	56.00	40.00
1:TGT	60.67	88.00	44.00	96.00	**28.00**	68.00	44.00
5:SRC	56.00	76.00	20.00	**100.00**	4.00	80.00	**56.00**
5:TGT	**64.67**	**100.00**	**48.00**	96.00	12.00	**84.00**	48.00

Table 6: Percentage of correct *gender* answers on the COH-TGT test sets for Basque to Spanish translation

MODEL	MERGED	LITERATURE		PARLIAMENT		TED	
		FORMAL	INFORMAL	FORMAL	INFORMAL	FORMAL	INFORMAL
BASELINE	56.67	24.00	88.00	56.00	64.00	48.00	60.00
1:SRC	62.00	32.00	84.00	76.00	60.00	56.00	64.00
1:TGT	75.33	64.00	88.00	68.00	**84.00**	64.00	84.00
5:SRC	50.00	24.00	60.00	68.00	40.00	56.00	52.00
5:TGT	**84.00**	**80.00**	**92.00**	**88.00**	**84.00**	**68.00**	**92.00**

Table 7: Percentage of correct *register* answers on the COH-TGT test sets for Basque to Spanish translation

sets are shown in Table 7. Models trained on target context outperformed the baseline in all cases, and the variants based on source context in most, as expected in this case as well – the latter were even outperformed by the baselines in 5 cases out of 12, and performed similarly in 1 other case.

Even more markedly than with target gender results, using larger context information was beneficial overall for target-based models on the register test set, indicating that these models could exploit the relevant information beyond the closest context.

6. Conclusion

We described TANDO, a multi-domain document-level corpus for the under-resourced Basque-Spanish language pair, shared for research purposes. The corpus is composed of parallel data from three different domains, covering literature, news and subtitles, and has been prepared with context-level information. Additionally, we prepared contrastive test sets for targeted evaluations of gender and register contextual phenomena, on both source and target language sides.

To establish the usefulness of the corpus, we trained and evaluated baseline Transformer models, and context-aware variants based on context concatenation that exploited the context of either the source or the target language. Our results indicate that the corpus is suitable for fine-grained evaluations of document-level machine translation, with context-aware variants outperforming the sentence-level baselines in most scenarios, on both parallel and contrastive test sets.

Overall, model variants relying on a single context sentence performed slightly better than those based on larger contexts on the parallel test sets and the gender contrastive sets with relevant information in both source or target sides. On target contrastive sets, for both gender and register, models operating on larger contexts obtained markedly better results. Gender biases in the training data were reflected in the results for the contrastive sets, with lower accuracy obtained across the board on contrastive feminine gender.

In future work, we will further analyse the TANDO data, in particular, the location of relevant sources of contextual information, and evaluate additional context-aware modelling alternatives on the corpus.

7. Acknowledgements

This work was partially supported by the Department of Economic Development of the Basque Government, via project TANDO (KK-2020/00074). We wish to thank the Basque public broadcaster EiTB for their support and the anonymous LREC reviewers for their helpful comments.

8. Bibliographical References

Agrawal, R., Turchi, M., and Negri, M. (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 (EAMT)*.

Azpeitia, A. and Etchegoyhen, T. (2019). Efficient document alignment across scenarios. *Machine Translation*, 33:205–237.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora*, pages 41–45, Vancouver, Canada.

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

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

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

Etchegoyhen, T. and Azpeitia, A. (2016a). A portable method for parallel and comparable document alignment. *Baltic Journal of Modern Computing*, 4(2):243–255. *Special Issue: Proceedings of EAMT 2016*.

Etchegoyhen, T. and Azpeitia, A. (2016b). Set-theoretic alignment for comparable corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2009–2018, Berlin, Germany.

Etchegoyhen, T., Azpeitia, A., and Pérez, N. (2016). Exploiting a large strongly comparable corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Fernandes, P., Yin, K., Neubig, G., and Martins, A. F. T. (2021). Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online, August. Association for Computational Linguistics.

Gong, Z., Zhang, M., and Zhou, G. (2011). Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 283–289, Paris, France, December 2-3.

Jauregi Unanue, I., Esmaili, N., Haffari, G., and Piccardi, M. (2020). Leveraging discourse rewards for document-level neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4467–4482, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jean, S. and Cho, K. (2019). Context-aware learning for neural machine translation. *CoRR*, abs/1903.04715.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017a). Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017b). Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 54–57, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

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

Kang, X., Zhao, Y., Zhang, J., and Zong, C. (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, November. Association for Computational Linguistics.

Kim, Y., Tran, D. T., and Ney, H. (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, November. Association for Computational Linguistics.

Kimura, R., Iida, S., Cui, H., Hung, P.-H., Utsuro, T., and Nagata, M. (2019). Selecting informative context sentence by forced back-translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 162–171, Dublin, Ireland, August. European Association for Machine Translation.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Kuang, S., Xiong, D., Luo, W., and Zhou, G. (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, August. Association for Computational Linguistics.

Läubli, S., Sennrich, R., and Volk, M. (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, October-November. Association for Computational Linguistics.

Li, B., Liu, H., Wang, Z., Jiang, Y., Xiao, T., Zhu, J., Liu, T., and Li, C. (2020). Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online, July. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2015: Extracting large parallel corpora from movie and tv subtitles. In *International Conference on Language Resources and Evaluation*.

Liu, S. and Zhang, X. (2020). Corpora for document-level neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France, May. European Language Resources Association.

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

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Ma, S., Zhang, D., and Zhou, M. (2020). A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online, July. Association for Computational Linguistics.

Mansimov, E., Melis, G., and Yu, L. (2021). Capturing document context inside sentence-level neural machine translation models with self-training. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 143–153, Punta Cana, Dominican Republic and Online, November. Association for Computational Linguistics.

Maruf, S. and Haffari, G. (2018). Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia, July. Association for Computational Linguistics.

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

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (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, October-November. Association for Computational Linguistics.

Morishita, M., Suzuki, J., Iwata, T., and Nagata, M. (2021). Context-aware neural machine translation with mini-batch embedding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2513–2521, Online, April. Association for Computational Linguistics.

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

Nagata, M. and Morishita, M. (2020). A test set for discourse translation from Japanese to English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3704–3709, Marseille, France, May. European Language Resources Association.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Rikters, M. and Nakazawa, T. (2021). Revisiting context choices for context-aware machine translation. *CoRR*, abs/2109.02995.

Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sarasola, I., Salaburu, P., and Landa, J. (2015). Hizkuntzen arteko corpusa (hac).

Sennrich, R., Haddow, B., and Birch, A. (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, August. Association for Computational Linguistics.

Sennrich, R. (2017). How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April. Association for Computational Linguistics.

Sun, Z., Wang, M., Zhou, H., Zhao, C., Huang, S., Chen, J., and Li, L. (2020). Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *CoRR*, abs/2010.08961.

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

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

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden, July. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218, Istanbul, Turkey.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. (2018). Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin,

I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (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, July. Association for Computational Linguistics.

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

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

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

Webber, B. (2014). Discourse for machine translation. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 27–27, Phuket,Thailand, December. Department of Linguistics, Chulalongkorn University.

Wong, B. T. M. and Kit, C. (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July. Association for Computational Linguistics.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23.

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

Xu, H., Xiong, D., van Genabith, J., and Liu, Q. (2020). Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3933–3940. International Joint Conferences on Artificial Intelligence Organization, 7. Main track.

Xu, M., Li, L., Wong, D. F., Liu, Q., and Chao, L. S. (2021). Document graph for neural machine translation. In Marie-Francine Moens, et al., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8435–8448. Association for Computational Linguistics.

Yamagishi, H. and Komachi, M. (2019). Improving context-aware neural machine translation with target-side context. *CoRR*, abs/1909.00531.

Yang, Z., Zhang, J., Meng, F., Gu, S., Feng, Y., and Zhou, J. (2019). Enhancing context modeling with a query-guided capsule network for document-level 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 1527–1537, Hong Kong, China, November. Association for Computational Linguistics.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zheng, Z., Yue, X., Huang, S., Chen, J., and Birch, A. (2020). Towards making the most of context in neural machine translation.