An Evaluation of Source Factors in Concatenation-based Context-aware Neural Machine Translation

Harritxu Gete[1,2], **Thierry Etchegoyhen**[1]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]University of the Basque Country UPV/EHU
{hgete,tetchegoyhen}@vicomtech.org

Abstract

We explore the use of source factors in context-aware neural machine translation, specifically concatenation-based models, to improve the translation quality of inter-sentential phenomena. Context sentences are typically concatenated to the sentence to be translated, with string-based markers to separate the latter from the former. Although previous studies have measured the impact of prefixes to identify and mark context information, the use of learnable factors has only been marginally explored. In this study, we evaluate the impact of single and multiple source context factors in English-German and Basque-Spanish contextual translation. We show that this type of factors can significantly enhance translation accuracy for phenomena such as gender and register coherence in Basque-Spanish, while also improving BLEU results in some scenarios. These results demonstrate the potential of factor-based context identification as a research path in context-aware machine translation.

1 Introduction

Machine translation typically operates at the sentence level, leaving aside larger context information. This mode of operation remains dominant within the Neural Machine Translation (NMT) framework (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017), although it limits accurate translation for linguistic phenomena that depend on context information, such as cohesion, discourse coherence or intersentential anaphora resolution (Bawden et al., 2018; Läubli et al., 2018; Voita et al., 2019b; Lopes et al., 2020; Post and Junczys-Dowmunt, 2023).

Addressing discourse-related phenomena in translation requires extending the scope of the translation models to address the relevant information present in the context sentences, in addition to that of the sentence to be translated. Several approaches have been proposed within NMT to extend the modelling window beyond isolated sentences, extending the input by including context sentences (Tiedemann and Scherrer, 2017) or modifying the NMT architecture to model context information (Jean et al., 2017; Zhang et al., 2018; Voita et al., 2019b; Li et al., 2020).

Despite the marked improvements achievable with the aforementioned approaches, the identification of the relevant contextual information to improve the translation of a given sentence is still an open research topic. Within concatenation-based approaches (Tiedemann and Scherrer, 2017), a simple yet strong document-level NMT baseline, context sentences are typically prepended to the sentence to be translated, and separated from it by a simple marker. Further identification of what belongs to the context or to the sentence to be translated is typically discarded, following in part initial results by Tiedemann and Scherrer (2017) where the use of prefixes to identify context tokens led to degraded results at best. An alternative method that may provide better context identification is the utilization of factors as context markers. Factors are learnable embeddings associated to input tokens that provide supplementary information about the token. Different approaches, such as addition or concatenation, can be employed to combine token embeddings with factor embeddings. Within the context identification process, this supplementary information may serve to indicate whether the token belongs to the context or not. To our knowledge, the use of these markers for context aware NMT has only been partially explored, and the results obtained so far have been inconclusive (Rikters et al., 2020; Lupo et al., 2023).

In this work, we present extended results on the use of factors for context-aware NMT, centred on using source factors and measuring their impact on both standard and contrastive datasets. We re-

Proceedings of Recent Advances in Natural Language Processing, pages 399–407
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_045

port results on English-German pronoun translation using the ContraPro test set (Müller et al., 2018), and on Basque-Spanish gender selection and register coherence with the TANDO test sets (Gete et al., 2022). We show that source factors can significantly enhance translation accuracy for phenomena such as gender and register coherence in Basque-Spanish, while also improving BLEU results in some cases. These results demonstrate the potential of factor-based context identification as a research path to improve context-aware machine translation.

2 Related Work

The inclusion of contextual information to improve machine translation is a long-standing topic of interest in the field (Mitkov, 1999; Tiedemann and Scherrer, 2017). Within the NMT paradigm in particular, an increasing number of studies have centred on context-aware NMT approaches and the improvements that these models may provide over non-contextual baselines (Li et al., 2020; Ma et al., 2020; Lopes et al., 2020; Fernandes et al., 2021; Majumde et al., 2022; Sun et al., 2022).

One of the first methods proposed for the task is the concatenation of context sentences to the sentence to be translated (Tiedemann and Scherrer, 2017), a simple approach which provides a robust baseline that often matches or outperforms more sophisticated methods (Lopes et al., 2020; Sun et al., 2022; Post and Junczys-Dowmunt, 2023). Variants of this approach include discounting the loss generated by the context (Lupo et al., 2022), extending model capacity (Majumder et al., 2022; Post and Junczys-Dowmunt, 2023) or encoding the specific position of the context sentences (Lupo et al., 2023). The latter in particular includes the use of learned embeddings for each sentence position, for which they report mixed results with improvements in English-Russian and a negative impact in English-German, using three context sentences. We include a variant of this approach in the form of separate factors for each context sentence, without discounting context loss and applying it to a larger context on English-German and Basque-Spanish datasets.

Alternative approaches to input extension notably include refining context-agnostic translations (Voita et al., 2019a) and modelling context information with specific NMT architectures (Jean et al., 2017; Li et al., 2020).

Since context-aware models are particularly suited to improve the translation of phenomena that directly depend on context information, several challenge test sets have been created specifically to evaluate the ability of models to adequately translate these phenomena in context (Guillou and Hardmeier, 2016; Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Lopes et al., 2020; Gete et al., 2022).

The use of factors was introduced in Statistical Machine Translation as a means to incorporate additional linguistic information (Koehn and Hoang, 2007). For NMT, the concurrent work of Sennrich and Haddow (2016) and Hoang et al. (2016) explored how sentence-level NMT models could benefit from incorporating additional linguistic information via factors in the source language. They thus added morphological features, part-of-speech tags, and syntactic dependency labels as input features, obtaining promising results in terms of perplexity reduction and higher BLEU (Papineni et al., 2002) scores.

Source factors have only been partially explored for context-aware NMT. In addition to the previously cited work of Lupo et al. (2023) on learnable context sentence position embeddings, Rikters et al. (2020) also employ factors to identify tokens as pertaining to the context or to the sentence to be translated. In their experimental results on Japanese–English translation, using one context sentence, the use of factors provided only minimal absolute improvements in terms of BLEU over simple input concatenation. Our work differs from theirs in several respects: we used larger contexts of 5 sentences, evaluated them on two language pairs, used contrastive evaluations on context phenomena in addition to BLEU scores, and measured the impact of both unique and multiple context factors.

3 Experimental Setup

3.1 Data

We describe in turn below the parallel and contrastive data used to train and test our NMT models in Basque-Spanish and English-German.

Parallel Data For Basque–Spanish, we selected the TANDO corpus (Gete et al., 2022), which contains parallel data from subtitles, news and literary documents, and includes validation and test sets. For English–German, we followed the approach of Müller et al. (2018) and the data was obtained from the WMT 2017 news translation task, using newstest2017 and newstest2018 as test sets, and the union of newstest2014, newstest2015 and newstest2016 for validation. Table 1 summarises parallel corpora statistics.

	EU-ES	EN-DE
TRAIN	1,753,726	5,852,458
DEV	3,051	2,999
TEST	6,078	6,002

Table 1: Parallel corpora statistics (number of sentences)

Contrastive Test Data For Basque–Spanish, we used the contrastive test set included in TANDO, a set created from collected books, TED talks, and proceedings of the Basque Parliament. It is designed to assess a model's ability to select the correct translation in terms of the choice of gender (feminine or masculine) or register (formal or informal) of certain words and it is composed of 600 instances, divided into two subsets: GDR-SRC+TGT, where the disambiguating information to predict the gender is present in both the source and target languages and COH-TGT, which evaluates the contextual coherence of the translation despite the absence of necessary information in the source language to make a correct selection of gender or register. All instances require contextual knowledge to select the correct translation.

For English–German, we used ContraPro (Müller et al., 2018) a contrastive test created from OpenSubtitles2018[1] (Lison et al., 2018) excerpts aiming to test the ability of a model to identify the correct German translation of the English anaphoric pronoun *it* as *es*, *sie* or *er*. It contains 12,000 instances, 4,000 per category, and requires knowledge of the context in 80% of the cases to select the correct translation.

All selected datasets were normalised, tokenised and truecased using Moses scripts (Koehn et al., 2007) and segmented with BPE (Sennrich et al., 2016), using 32,000 operations.

3.2 Models

We trained sentence-level baselines and concatenation-based context-aware models, which extend the input by concatenating the previous sentences to the current one to be translated (Tiedemann and Scherrer, 2017). This approach was selected for its simplicity and robustness, as it typically obtains competitive results without any modification of the NMT architecture (Tiedemann and Scherrer, 2017; Lopes et al., 2020; Majumde et al., 2022). We opted to use 5 context sentences, since for the two selected contrastive tests, the disambiguation information is always found within this context window.

Gete et al. (2022) noted that, although they provide marked improvements in terms of contrastive evaluations, models trained on concatenated context can worsen translation quality in terms of BLEU, especially with longer contexts. This might be due to increasing difficulties in identifying which parts of the information provided to the model are actually relevant to properly translate the current sentence. For larger contexts in particular, factors may help discriminate the different parts of the input provided to the model, at least in terms of separating context tokens from those of the sentence to be translated.[2]

To explore this hypothesis, we trained three variants of concatenation-based models, along with a sentence-level baseline, based on the Transformer-base architecture (Vaswani et al., 2017):

- SENTENCE-LEVEL: a standard Transformer-base model without input context.
- CONTEXT-AWARE: a standard Transformer-base model with concatenated input context, separated from the input sentence with a BREAK marker.
- CONTEXT-AWARE+FACTOR: a concatenation-based model that includes source factors with two different values to differentiate the sentence to be translated (S) from the context sentences (C). The factors are added at the token level and we eliminate the BREAK marker, as the factors serve to delimit which tokens are part of the context.
- CONTEXT-AWARE+MULTIFACTOR: This approach is similar to the previous one, but uses different values for the factor of each sentence in the context (C1, ..., C5). This approach is

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

[2] Note that this differs from the use of prefixes attached to context subwords, as in Tiedemann and Scherrer (2017). In preliminary experiments, we also experimented with inline annotations to indicate if an input token pertained to the context. This method was discarded as it resulted in losses in terms of both BLEU scores and accuracy on the contrastive test sets.

CONTEXT-AWARE	
Text: I think we work on the m_ ou_ sta_ che first . give him a little s_ no_ op . this side 's too long . give him a little s_ no_ op this side . now this side is too short . [BREAK] it 's too short .	
CONTEXT-AWARE+FACTOR	
Text: I think we work on the m_ ou_ sta_ che first . give him a little s_ no_ op . this side 's too long . give him a little s_ no_ op this side . now this side is too short . it 's too short . **Factors:** C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C S S S S S	
CONTEXT-AWARE+MULTIFACTOR	
Text: I think we work on the m_ ou_ sta_ che first . give him a little s_ no_ op . this side 's too long . give him a little s_ no_ op this side . now this side is too short . it 's too short . **Factors:** C5 C5 C5 C5 C5 C5 C5 C5 C5 C4 C4 C4 C4 C4 C4 C3 C3 C3 C3 C3 C3 C2 C2 C2 C2 C2 C2 C2 C2 C2 C2 C1 C1 C1 C1 C1 C1 C1 S S S S S	

Table 2: Examples of input for context-aware models. C denotes context, Ci context provided by the i-th preceding sentence, and S the sentence to be translated.

	EU-ES		EN-DE		
	parallel	*contrastive*	*wmt2017*	*wmt2018*	*ContraPro*
SENTENCE-LEVEL	31.1	35.6	28.0	41.1	22.4
CONTEXT-AWARE	**32.0**	**38.3**	28.4	**42.0**	24.4
CONTEXT-AWARE+FACTOR	**32.0**	**39.3**	28.4	**42.1**	**25.2**
CONTEXT-AWARE+MULTIFACTOR	**31.8**	**39.1**	**28.8**	**42.4**	**25.2**

Table 3: BLEU results for Basque–Spanish and English–German. Best performing systems, without statistically significant differences between them ($p < 0.05$), are shown in bold.

similar to the learned sentence position embeddings of Lupo et al. (2023), although we removed the context separation token and did not use context loss discarding.[3].

Factor and token embeddings can be combined using addition or concatenation. We opted for addition since this approach maintains the dimension of the original embeddings, whereas concatenation leads to larger embeddings overall. We left an exploration of the concatenation approach for future work.

An example of input data for each of the context-aware methods is shown in Table 2. Factors were only used on the source language side in this study. The target side includes a context separation BREAK marker between context sentences and the translated source sentence. All 5 source context sentences are translated along with the non-context source sentence, and all translated context target sentences that occur before the target break marker are discarded.

Factor embeddings were added for each source token and summed to the token embeddings, as is typically done with positional encodings in Transformer models. Thus, each token vector contains information about the token itself, its position in the input, and its belonging or not to the context.[4]

The embeddings for source, target and output layers were tied and 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 then decrease proportionally to the inverse square root of the corresponding step. Validation data were evaluated every 5,000 training steps, and the process ended if there was no improvement in the perplexity of 10 consecutive checkpoints. All models were trained with the MarianNMT toolkit (Junczys-Dowmunt et al., 2018) and context-aware models were initialised with the weights of the baseline sentence-level models.

[3] Our experimental setup also differs, notably in terms of training corpora.

[4] An alternative approach would have involved concatenating the factor embeddings instead of summing them. We left variants of this type for future experiments.

	TOTAL	GDR-SRC+TGT	COH-TGT GDR	COH-TGT REG
SENTENCE-LEVEL	54%	55%	48%	58%
CONTEXT-AWARE	71%	**78%**	61%	69%
CONTEXT-AWARE+FACTOR	74%	**78%**	63%	74%
CONTEXT-AWARE+MULTIFACTOR	**78%**	77%	**71%**	**86%**

Table 4: Accuracy results on the contrastive test sets for Basque–Spanish. Best results are shown in bold.

	TOTAL	*es*	*er*	*sie*
SENTENCE-LEVEL	49%	88%	23%	35%
CONTEXT-AWARE	74%	**93%**	63%	67%
CONTEXT-AWARE+FACTOR	**77%**	92%	**69%**	**71%**
CONTEXT-AWARE+MULTIFACTOR	**77%**	**93%**	68%	69%

Table 5: Accuracy results on the contrastive test sets for English–German. Best results are shown in bold.

4 Results and Analysis

4.1 BLEU Results

We first assessed the sentence- and context-level models in terms of BLEU (Papineni et al., 2002) using the SacreBLEU toolkit (Post, 2018)[5] on cased detokenised output. To determine whether differences in scores between models actually reflect differences in overall quality, we determined the statistical significance of our findings using paired bootstrap resampling (Koehn, 2004).

The results are presented in Table 3. In both language pairs, context-aware models obtained higher scores than the sentence-level baselines, which is not always the case with context-aware models on the BLEU metric (Gete et al., 2022). Turning to factor-based models, in Basque-Spanish the use of factors resulted in higher absolute values but none of these apparent improvements were statistically significant. In English-German similar results were obtained on the wmt2018 test set. However, both factored models obtained significantly better results than the context-aware baseline on the ContraPro test set. Additionally, the multi-factor variant also improved over the alternatives on the wmt2017 test set.

Overall, the improvements that had statistical significance ranged from .4 to .8 BLEU points. Although relatively minor, these gains indicate that the use of source factors has the potential to enhance translation outcomes in certain scenarios, and did not worsen them in any of the cases in our experiments.

[5]signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

4.2 Contrastive Results

Accuracy results for the contrastive test sets described above are shown in Tables 4 and 5, for Basque–Spanish and English-German, respectively.

Regarding coherence, the use of factors clearly enhanced the performance of Basque-Spanish translation models for both gender and register tests. Notably, models that incorporate multiple context factors exhibited marked improvements, with gains of 10 and 17 percentage points on gender and register, respectively. For the GDR-SRC+TGT test, however, the outcomes remained practically unchanged with respect to those of the non-factored model.

In the case of English-German models, the use of factors led to lesser differences, with an overall increased accuracy of only 3 percentage points for both single and multiple factors. Looking at the different pronominal categories, the improvements were mostly based on increased accuracy for the translation of pronouns *er* and *sie*, with improvements of 6 and 4 percentage points, respectively, when using single factors in the first case and multiple factors in the second case. This is not totally unexpected considering the already high accuracy for the translation of *es* by all models, including the sentence-level baseline.

For both language pairs, it is worth noting that the most substantial improvements are observed in cases with initially lower results, while those with high initial scores (GDR-SRC+TGT for Basque-Spanish and the subset corresponding to *es* in English-German) remain similar overall.

	EN-DE	EU-ES		
	TOTAL	GDR-SRC+TGT	COH-TGT GDR	COH-TGT REG
CONTEXT-AWARE+FACTOR	15%	17%	15%	29%
CONTEXT-AWARE+MULTIFACTOR	14%	17%	26%	33%

Table 6: Difference in predictions compared to the model without factors, for English-German and Basque-Spanish factored models.

	EN-DE	EU-ES		
	TOTAL	GDR-SRC+TGT	COH-TGT GDR	COH-TGT REG
CONTEXT-AWARE	1.14	1.67	1.97	1.65
CONTEXT-AWARE+FACTOR	1.18	1.66	1.87	1.49
CONTEXT-AWARE+MULTIFACTOR	1.13	1.71	2.14	1.71

Table 7: Average distance in number of sentences (from the current sentence to the disambiguating information) of the test cases that cannot be solved by the models.

4.3 Impact of Factors Beyond Metrics

To complement the results in terms of BLEU and accuracy on contrastive test sets, we examined two different aspects regarding the use of factors.

First, we aimed to evaluate the extent to which the use of factors impacted translation results, even when the final score remained almost identical. To gain further understanding on this question, we computed the percentage of predictions that differed in each contrastive test between factored models and baseline context-aware models. The results in Table 6 indicate that, for Basque-Spanish, even for models where results were identical, as between the context-aware baseline and the single factor model (78% in this case), or almost identical as with the multi-factor model (77%), the predictions between models differed by 17%. A similar figure was obtained for English-German, where the difference amounted to 15% for the single factor model, and 14% when using multiple factors. The latter model featured the largest differences on the two coherence test sets in Basque-Spanish, which is in line with the larger metrics improvements obtained for the gender and register coherence contextual categories. Determining the specific conditions where the use of factors resulted in accuracy loss, thus negatively balancing the cases where factors resulted in gains, would require a more specific analysis which we leave for future work.

Additionally, we measured the average distance to the context sentence in all cases where the models made an incorrect contrastive prediction, with the results shown in Table 7. In English-German, the differences were minor overall, in line with the relatively close results in terms of metrics described in the previous sections. In Basque-Spanish, the model with the largest improvements, using multi-factors, was associated with increased distances, i.e. an extended context window over which the model could provide more accurate results. In this case as well, a more precise analysis of the contrastive predictions would be needed to further establish the strengths and weaknesses in the use of context factors.

5 Conclusions

In this work, we explored the use of factors in context-aware neural machine translation to improve the translation quality of inter-sentential phenomena. Specifically, we evaluated the impact of source factors in concatenation-based models, using both single factors for all context sentences, and multi-factors, where separate factors are assigned for each context sentence.

We conducted our experiments on parallel and contrastive test sets in English-German and Basque-Spanish, using larger contexts than in previous related studies, and targeting different phenomena such as pronoun translation, gender selection, and coherence in both register and gender.

Overall, both of the evaluated factor-based approaches improved over the concatenation-based baseline. In terms of BLEU, these approaches either matched or improved over the baseline, although the gains were relatively minor and only statistically significant on two test sets in English-German. On the contrastive sets, the largest gains were obtained in Basque-Spanish on the coherence-related tests, achieving gains of 10 and 17 percentage points in accuracy. On the gender selection test, no improvements were observed in this language pair, however. In English-German, the factor approach improved over the baseline overall, but with comparatively smaller gains.

The multi-factor approach provided the most consistent benefits across metrics, with additional results showing its increased accuracy in context-based predictions at a larger distance than the baseline and the single factor approach. This approach might thus be worth exploring further in different contexts or in combination with other approaches.

Our study mainly aimed to measure the potential of context factors in NMT, on a diverse set of test sets with relatively large contexts. In future work, we will further investigate factor-based context-aware NMT variants, notably by measuring the impact of target-side factors, evaluating the use of factors in combination with other context identification markers, and extending the analyses to more language pairs and contextual phenomena.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

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.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 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.

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. 2022. TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14, Melbourne, Australia.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 54–57, Copenhagen, Denmark.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia.

Diederick P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

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

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, 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.

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.

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.

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.

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.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid).

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 33–44, Dubrovnik, Croatia.

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.

Suvodeep Majumde, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: Pushing the limits of multi-segment models for context-aware translation. *arXiv preprint arXiv:2210.10906v2*.

Suvodeep Majumder, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: Pushing the limits of multi-segment models for context-aware translation.

Ruslan Mitkov. 1999. Introduction: Special issue on anaphora resolution in machine translation and multilingual NLP. *Machine translation*, pages 159–161.

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.

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.

Matt Post. 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.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv:2304.12959v1*.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2020. Document-aligned Japanese-English conversation parallel corpus. In *Proceedings of the Fifth Conference on Machine Translation*, pages 639–645, Online.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany.

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.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

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.

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.

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.