What Works When in Context-aware Neural Machine Translation?

Harritxu Gete[1,2] **Thierry Etchegoyhen**[1] **Gorka Labaka**[2,3]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
[2]University of the Basque Country UPV/EHU
[3]HiTZ Basque Center for Language Technologies - Ixa
{hgete,tetchegoyhen}@vicomtech.org, gorka.labaka@ehu.eus

Abstract

Document-level Machine Translation has emerged as a promising means to enhance automated translation quality, but it is currently unclear how effectively context-aware models use the available context during translation. This paper aims to provide insight into the current state of models based on input concatenation, with an in-depth evaluation on English–German and English–French standard datasets. We notably evaluate the impact of data bias, antecedent part-of-speech, context complexity, and the syntactic function of the elements involved in discourse phenomena. Our experimental results indicate that the selected models do improve the overall translation in context, with varying sensitivity to the different factors we examined. We notably show that the selected context-aware models operate markedly better on regular syntactic configurations involving subject antecedents and pronouns, with degraded performance as the configurations become more dissimilar.

1 Introduction

Neural Machine Translation (NMT) systems (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) have traditionally translated sentences in isolation without considering relations between discourse elements. This leads to translations lacking crucial textual properties 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).

Properly handling discourse-related phenomena requires extending the scope of the translation model beyond the sentence level. As a result, many methods have been developed to extend the modeling window beyond isolated sentences. These approaches range from extending the input of standard NMT models (Tiedemann and Scherrer, 2017) to architectural variants (Tu et al., 2018; Miculicich et al., 2018; Li et al., 2020).

Despite the promising results achieved by context-aware NMT, determining the precise use of context remains a significant challenge, leading to contradictory findings, including studies suggesting that context-aware models do not improve intersentential phenomena, but rather act as mere regularisers (Kim et al., 2019; Li et al., 2020; Rauf and Yvon, 2020). Standard translation metrics have limitations to measure document-level phenomena, whereas contrastive evaluations provide more precise measures but do not delve into how context information is actually used or ignored. We believe that in-depth analyses of context usage by context-aware models could help better understand their current strengths and limitations.

In this paper, we analyse the performance of various approaches based on context concatenation, a strong baseline for document-level NMT, examining variations in the use of source and target context. We provide an in-depth analysis of the results achieved by the selected NMT models in terms of data bias, context complexity, as well as part of speech and syntactic functions of the relevant elements in contextual translation. We focus our study on pronoun translation for English–German and English–French, for which there are publicly available annotated datasets.

© 2023 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

2 Related work

Using contextual information to improve machine translation has been a topic of interest in the community for decades (Mitkov, 1999; Tiedemann and Scherrer, 2017). Research within the NMT paradigm, where contextual information may be accessed over extended input windows, has led to a number of new approaches to incorporate intersentential context for more accurate translation.

A variety of studies have explored context-aware NMT approaches, analysing the improvements that these models can provide over non-contextual baselines (Li et al., 2020; Lopes et al., 2020; Ma et al., 2020; Fernandes et al., 2021). One of the first proposed methods is the concatenation of context sentences to the sentence to be translated (Tiedemann and Scherrer, 2017). This simple approach is still efficient, achieving comparable or superior performance to more complex approaches (Lopes et al., 2020; Sun et al., 2022). Other methods involve refining context-agnostic translations (Voita et al., 2019a), or modelling context information with specific NMT architectures (Jean et al., 2017; Li et al., 2020). Some of these models only use source language context (Wang et al., 2017; Zhang et al., 2018), while others include target language context as well (Voita et al., 2019a).

Context-aware models have shown to be effective in the translation of context-dependent phenomena (Müller et al., 2018) and several test sets have been created to specifically evaluate the ability of models to accurately translate pronouns within their 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).

Stojanovski et al. (2020) show that inserting small amounts of distracting information is enough to strongly decrease scores in contrastive tests, and Kim et al. (2019) found that only a few sentences are really useful to improve translation quality. A deeper and more thorough analysis is thus still required to draw firm conclusions about the strengths and weaknesses of context-aware models.

	EN-DE	EN-FR	
	DOC-LEVEL	SENT-LEVEL	DOC-LEVEL
TRAIN	5,852,458	11,221,790	234,738
DEV	2,999	4,992	5,818
TEST	6,002	-	1,210

Table 1: Parallel corpora statistics (number of sentences)

3 Experimental setup

3.1 Data

All selected datasets described below were normalised, tokenised and truecased identically to WMT2017 data, using Moses (Koehn et al., 2007) scripts. The data were segmented with BPE (Sennrich et al., 2016), using 32,000 operations. For the experiments in Sections 4.4 and 4.5, syntactic tags were obtained with Stanza (Qi et al., 2020).

Parallel Data For English–German, we followed Müller et al. (2018) and used the data from the WMT 2017 news translation task, using newstest2017 and newstest2018 as test sets, and the union of newstest2014, newstest2015 and newstest2016 for validation. Both sentence-level and context-aware models use the same data in this language pair. For English–French, we use parallel data from publicly available resources to train baseline models, namely Europarl v7, News-Commentary v10, CommonCrawl, UN, Giga from WMT 2017 and the IWSLT17 TED Talks. Following Lopes et al. (2020), we then fine-tuned context-aware models on IWSLT17, using the test sets 2011-2014 as dev sets, and 2015 as test sets. Table 1 summarises parallel corpora statistics.

Test Data To evaluate the models, we selected the task of pronoun translation, for which document-level evaluation suites exist.

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 for 80% of them to select the correct translation. Table 2 summarises the numbers of instances in this set by pronominal category and by distance from the antecedent.

For English–French, we used the large-scale contrastive pronoun test set (hereafter, LSCP) (Lopes et al., 2020), which is similar to ContraPro but includes the translation of *they* as *elles* or *ils*, in addition to *it* as *elle* or *il*. This corpus was also prepared from OpenSubtitles2018 data and, as shown in Table 2, consists of 3,500 examples for each type of pronoun, totaling 14,000. Slightly less than 60% of the examples need contextual information

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

	EN-DE				EN-FR				
	it→es	*it→er*	*it→sie*	TOTAL	*it→elle*	*it→il*	*they→elles*	*they→ils*	TOTAL
0	872	736	792	2,400	1,658	1,628	1,535	1,165	5,986
1	1,892	2,577	2,606	7,075	1,144	1,094	1,148	1,180	4,566
>1	1,236	687	602	2,525	698	778	817	1,155	3,448
TOTAL	4,000	4,000	4,000	12,000	3,500	3,500	3,500	3,500	14,000

Table 2: Distribution of pronouns according to distance in sentences from the antecedent. English–German ContraPro (left) and English–French LSCP (right).

	EN-DE				EN-FR		
	BLEU			ACC	BLEU		ACC
	wmt2017	wmt2018	ContraPro	ContraPro	iwslt2017	LSCP	LSCP
SENT-LEVEL	**27.7**	41.1	22.7	49%	41.2	27.7	80%
2TO1-SRC	26.8†	40.7†	23.4†	58%	**42.6**†	28.7†	84%
2TO1-TGT	27.3†	40.7	**25.1**†	69%	**42.7**†	28.9†	87%
2TO2	**27.6**	**41.6**†	24.5†	73%	**42.5**†	**29.2**†	91%

Table 3: BLEU and contrastive accuracy (ACC) results for English–German and English–French. † indicates statistically significant BLEU results against the sentence-level baseline, for $p < 0.05$; best performing systems, without statistically significant differences between them, are shown in bold.

to make the correct translation choice. This test has less variety than ContraPro, as it is restricted to subject pronouns and noun antecedents.

3.2 Models

We trained sentence-level baselines and different variants of context-aware models. 2to1 models extend the input by concatenating the previous sentence to the current one, and included either the source language context (2to1-src) or the target language context (2to1-tgt). The extended input includes an additional sentence break token between the context and the current sentence. We also trained 2to2 models, which not only extended the input, but also the output; at inference time, the translated context was discarded. These approaches were selected as, despite their simplicity, they obtained competitive results without modifying the architecture (Tiedemann and Scherrer, 2017; Lopes et al., 2020; Majumde et al., 2022).

All models followed the Transformer-base architecture (Vaswani et al., 2017) and were trained with the MarianNMT toolkit (Junczys-Dowmunt et al., 2018). The embeddings for source, target and output layers were tied and optimisation was performed with Adam (Kingma and Ba, 2015). Context-aware models were initialised with the weights of the baseline models. For English–German, training was restarted resetting the learning rate, while for English–French, due to the limited data available, the baseline model was fine-tuned with the document-level data.

4 Results and Analysis

4.1 Metrics Results

We first evaluated the sentence- and context-level models in terms of BLEU and contrastive accuracy, with the results shown in Table 3. The scores were computed with the SacreBLEU toolkit (Post, 2018) and statistical significance was computed via paired bootstrap resampling (Koehn, 2004). Note that we evaluate target-dependent models using the reference target context, in order to assess the capability of these model with an ideal context.

For English–German, context-aware models achieved degraded BLEU results on wmt2017 and wmt2018, except for the 2to2 model, which improved over the sentence-level baseline on the latter test set. On ContraPro, all models markedly improved over the baseline, with better accuracy for models that include target context, the 2to2 model achieving the best scores overall.

In English–French, context proved beneficial for all tests and models, with no significant differences in terms of BLEU amongst context-aware models. The use of context substantially improved accuracy in the contrastive test set, and, in this language pair as well, with better results for models relying on the target context, notably the 2to2 model.

The relatively strong performance of the English–French sentence-level model is notewor-

	EN-DE			EN-FR			
	es	*er*	*sie*	*elle*	*il*	*elles*	*ils*
SENT-LEVEL	90%	11%	28%	59%	84%	35%	97%
2TO1-SRC	93%	37%	41%	71%	89%	59%	98%
2TO1-TGT	93%	55%	60%	77%	90%	66%	98%
2TO2	94%	65%	66%	90%	94%	83%	99%

Table 4: Accuracy results on the contrastive test sets for English–German and English–French (dist=1)

	EN-DE			EN-FR			
	es	*er*	*sie*	*elle*	*il*	*elles*	*ils*
SENT-LEVEL	34%	55%	50%	79%	66%	92%	60%
2TO1-SRC	38%	84%	82%	87%	74%	97%	71%
2TO1-TGT	47%	91%	90%	89%	79%	98%	75%
2TO2	52%	95%	93%	94%	90%	98%	85%

Table 5: Precision results on the contrastive test sets for English–German and English–French (dist=1)

thy. This could be partly attributed to the large number of instances of the test where contextual information is not required to achieve proper translation. Although such cases may be interesting to measure the impact of context in intra-sentential cases, they are not relevant to evaluate the use of extra-sentential information. To ensure a precise evaluation of the latter in our experiments, in what follows we only considered cases where the antecedent is in the immediately preceding sentence (dist=1). This discarded cases where no contextual information is required, as well as cases where the distance between the antecedent and the pronoun is greater than one sentence, which are beyond the scope of the selected models.

4.2 Data Bias

Table 4 shows the accuracy results per pronominal category with dist=1. The English–German model exhibits a clear inclination towards selecting the pronominal category *es*. This is likely due to the distribution in the training data, with a 33% probability of occurrence of the neuter pronoun, making it challenging for the model to learn to translate *er* and *sie*, with probabilities of 8% and 6%, respectively (Müller et al., 2018). Similarly, as shown in Table 5, the English–French model tends to favor the masculine pronouns *il* and *ils* over the feminine pronouns *elle* and *elles*. While this bias is more prominent in sentence-level models, the tendency is still notable in context-aware models, especially for English–German, as illustrated by the low precision results for *es*. Breaking down the results into more specific categories is thus important, as it provides more insight than relying on a single accuracy value, as is often the case.

It is worth noting that context-aware models improve both accuracy and precision across all categories. Although the improvements are more noticeable for categories negatively affected by bias, context also improves those that initially achieved high scores, such as the pronominal category *ils*, which improves from 97% to 99% of accuracy.

4.3 Part of Speech

We now turn to evaluating the impact of the part of speech (POS) of the antecedent on context-aware accuracy, focusing on cases where the antecedent is not expected to help contextual pronoun translation. This analysis was only conducted for English–German, as the English–French corpus exclusively contains nominal antecedents.

Overall, 79.5% of the antecedents in ContraPro are of a nominal (non-pronominal) type with POS NN (72.76%), NNP (5.64%) or NNS (1.10%).[2]. In all such cases, barring an erroneous identification of the actual antecedent, it is expected that the models can use the nominal antecedent to perform contextual translation. The remaining 20.5% of the cases feature POS categories that should not provide a relevant context for the translation of pronouns. We selected the most representative of those cases, namely personal pronoun *it/itself* (PRP: 14.22%), determiner (DT: 4.50%), and cardinal number (CD: 0.48%), discarding cases such as adjectives (JJ: 0.69%), which appeared along actual nominal antecedents in several cases.

[2]This information is included in the test set itself.

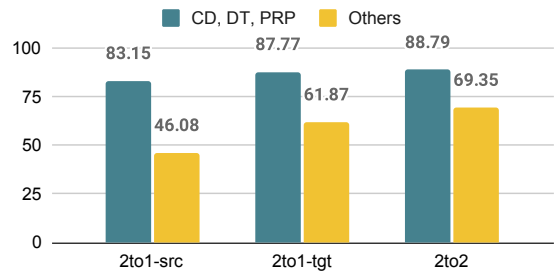

Figure 1: Accuracy results on EN-DE contrastive sets depending on antecedent POS : non-informative (CD,DT,PRP) vs informative (Others)

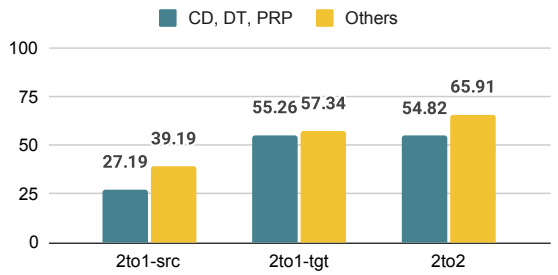

Figure 2: Accuracy results on EN-DE contrastive sets depending on antecedent POS , filtering the biased pronoun *es*: non-informative (CD,DT,PRP) vs informative (Others)

Figure 1 shows accuracy results on the selected POS categories, contrasted with all remaining ones. Surprisingly, all models performed better in the non-informative POS cases, including the 2to1-src model, which only uses source information. This may be due to the fact that a large percentage of these cases (83%) involve the pronoun *es*, which is often a default translation, as previously noted. As shown in Figure 2, when only the pronouns *er* and *sie* are considered, the models commit more errors with the uninformative POS antecedent, as might be expected, particularly the models that use source context.

The models that use source context show differences of more than 10 percentage points between the two analysed groups, whereas the model that only uses the target (2to1-tgt) achieves a more balanced result, which may be attributed to the use of target context information in the latter case. A chi-square test of independence (95% CI) showed that the results of the 2to1-src and 2to2 models depend on whether the antecedent is informative or not, which is not the case for the 2to1-tgt model.

Regarding uninformative antecedents, the 2to1-tgt and 2to2 models, which exploit target context, achieve similar results with an accuracy that is almost 30 percentage points higher than that of the source-context model. Cases in the test sets where the source context is uninformative may thus be compensated significantly by the use of the target context for correct gender selection.

Taking into account the above results regarding the translation of biased categories, in what follows we restrict our analyses to cases where the target pronoun is *er* or *sie* in English–German, and *elle* or *elles* in English–French.

4.4 Context complexity

We first set to analyse the impact of context complexity in terms of context length. Intuitively, it would seem that shorter contexts should be easier to handle, as they contain less information to discriminate, as well as less potential noise. We divided the selected cases within each test set (dist=1 and non-biased categories) into three groups based on context length: those with a length within the interquartile range Q1–Q3 (6–12 subwords for English–German and 7–14 subwords for English–French), those below this range, and those above it. Note that this analysis was performed using only source context length data, even though some models use only target context. This approach was chosen to ensure a fair comparison of results across all models and because source and target context lengths were found to be strongly correlated in the tests, with Pearson values of 0.87 for English–German and 0.89 for English–French.

Accuracy scores on the contrastive test sets for each group are shown in Figure 3. For English–German, shorter contexts did result in higher scores for all models, as per the initial intuition. Moreover, according to a chi-square test (95% CI), the results for the 2to1-src and 2to2 models were dependent on the length of the context, although this was not the case for 2to1-tgt. In contrast, in English–French only the results of the 2to2 model were dependent on context length, with the best results obtained for cases where the context length was closer to the median.

Besides length, context complexity could also be viewed as a factor of the number of potential nominal antecedents. To evaluate this aspect, we divided the test sets into simple and complex categories: cases where the context contained more than one subject or contained an object or a nominal oblique, in addition to the subject, were classi-

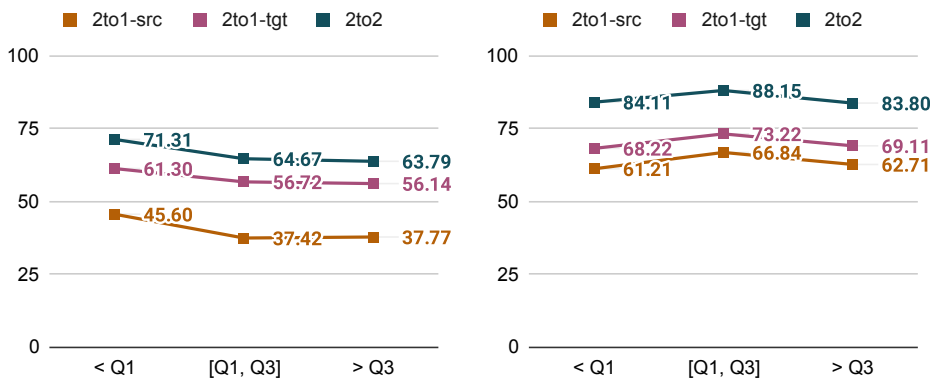

Figure 3: Accuracy results on contrastive sets depending on context length (#tokens) in English–German (left) and English–French (right). For English–German: Q1=6, Q3=12; for English–French: Q1=7, Q3=14.

fied as complex, all other contexts were considered simple. Note that we chose these three defining cases as they were the most common among antecedents in the test sets.

The results for both languages can be seen in Figure 4. According to a chi-square test of independence (95% CI), simple contexts generally performed better for all models in English–German For English–French, there were no statistically significant differences in results between complex and simple contexts, although absolute values were higher for the 2to1-src and 2to2 models on the complex dataset.

Overall, although shorter and simpler contexts tend to result in better performance for English–German, this was not the case for English–French, and the relation between context complexity and accuracy may thus vary depending on model architecture and language pair. We leave further analyses of these differences for future research.

Figure 4: Accuracy results depending on context complexity for English–German (top) and English–French (bottom).

4.5 Syntactic Function

We also investigated whether the syntactic functions of the pronoun and its antecedent influenced translation results. More specifically, we aimed to evaluate the accuracy of context-aware translation according to two variables: the actual syntactic functions of a pronoun and its antecedent, and whether the two differed in function.

We first analysed the accuracy of the models on the main combinations of syntactic tags listed in Table 6, which accounted for more than 85% of the cases. The results are shown in Figure 5.

In English–German, *nsubj–nsubj* was the most successful combination, followed by *obj–obj* and *root–nsubj*. The same trend was observed for all models, with some minor differences for the relative ranks of the worst configurations, although the worst cases overall consistently involved the *obj–nsubj*, *nmod–nsubj*, and *obl–nsubj* combinations.

In English–French, the results were less marked, particularly for the 2to2 model, which obtained similar results across all combinations, all above 80%. 2to1 models maintain the same trend as in English–German, except for the *obl–nsubj* case. When considering the two most common combinations, *nsubj–nsubj* and *obj–nsubj*, which covered about 70% of the cases, *nsubj–nsubj* consistently

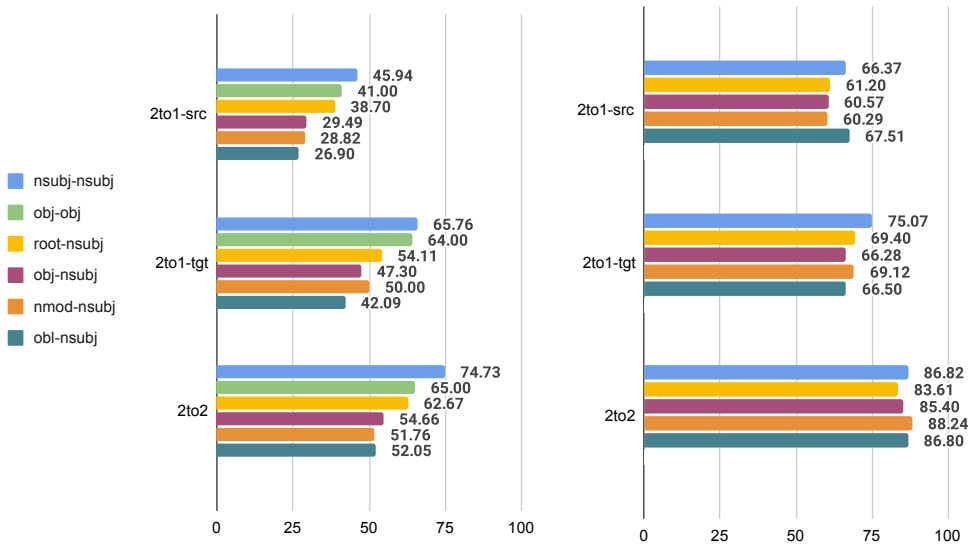

Figure 5: English–German (left) and English–French (right) accuracy results depending on syntactic functions

	EN-DE	EN-FR
NSUBJ–NSUBJ	46,19%	48,65%
OBJ–NSUBJ	23,62%	26,00%
OBL–NSUBJ	6,12%	8,60%
ROOT–NSUBJ	5,63%	7,98%
NMOD–NSUBJ	3,28%	2,97%
OBJ–OBJ	1,93%	-

Table 6: Distribution of antecedent-pronoun syntactic tags in the contrastive test sets

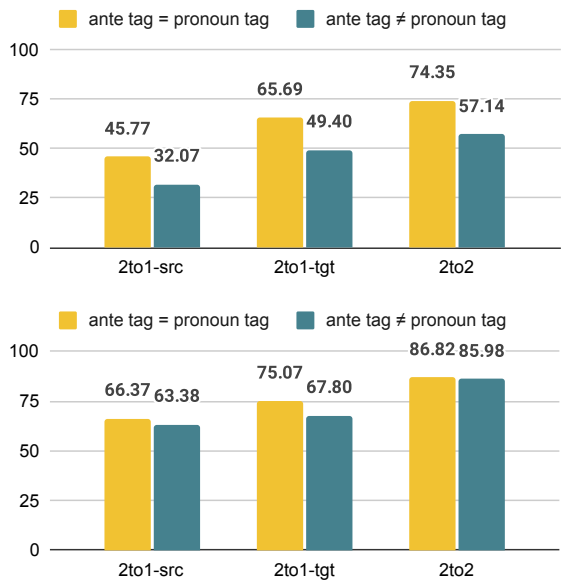

Figure 6: Accuracy results depending on syntactic tag identity in English–German (top) and English–French (bottom)

performed better in both language pairs.

Overall, the concatenation models thus seem to perform markedly better for the *nsubj–nsubj* configuration in both language pairs, followed by *obj–obj* in English-German. It might thus be the case that, more than the actual combination of syntactic tags for the pronoun and its antecedent, it is the fact that they share the same tag which leads to the best results with concatenated models.

To test whether this is actually the case, we evaluated the accuracy of the models in terms of tag identity between pronoun and antecedent, with the results shown in Figure 6. In English–German models, markedly better results were obtained with all models when the antecedent and the pronoun had the same syntactic function, which was confirmed by a chi-square test of independence. A similar result was obtained in English–French, but in this case, the chi-square test indicated significance only for the 2to1-tgt model. Overall, these findings suggest that syntactic function identity between pronoun and antecedent might be a determining factor for current concatenation models.

The results so far are still somewhat unclear, though, as the determining factors for optimal results might be either having a subject antecedent or identical tags for the pronoun and its antecedent. The results are further obscured by the fact that the LSCP dataset only contains subject pronouns, whereas ContraPro features more variety but still contains subject pronouns in 93% of the cases. Furthermore, in the latter test set, of the 2,495 cases with identical tags between pronoun and antecedent, 96% are cases where the antecedent is a

			2TO1-SRC	2TO1-TGT	2TO2
EN-DE	ante tag=pronoun tag	ante tag=nsubj	54.05%	65.75%	74.73%
	ante tag=pronoun tag	ante tag≠nsubj	41.58%	64.36%	65.35%
	ante tag≠pronoun tag	ante tag=nsubj	37.63%	60.22%	65.05%
	ante tag≠pronoun tag	ante tag≠nsubj	31.65%	48.60%	56.55%
EN-FR	ante tag=pronoun tag	ante tag=nsubj	66.37%	75.07%	86.82%
	ante tag≠pronoun tag	ante tag≠nsubj	63.38%	67.80%	85.98%

Table 7: Accuracy results as a function of tag identity and antecedent tag type in English-German and English-French

subject. And of the 2,580 cases with a subject antecedent, 93% also have a subject pronoun, resulting in shared syntactic functions that overly represent subjects.

This raises the question of whether it is one of the two conditions, tag identity or subject antecedents, that truly lead to improved results, or if their substantial overlap makes the findings difficult to interpret. To address this issue, we analysed separately the results for each of these subsets as well as for cases that did not meet either condition, across all models and language pairs. The results in Table 7 seem to provide a more consistent picture in both language pairs and across models. Function identity involving subjects is optimal across the board, followed by identity irrespective of the subject function, with the worst results when pronoun and antecedent have different syntactic functions and the antecedent is not a subject. This seems to indicate that concatenation models of the kind explored in this work are currently limited to specific regular configurations to properly handle context information. However, new contrastive test sets with more varied configurations would be needed in the future to further assess the observed limitations.

5 Conclusions

In this paper, we presented a systematic analysis of various concatenation-based context-aware models to help gain a clearer view of their current strengths and limitations. We compared the performance of three different approaches, using a limited context window of one sentence from the source and/or the target context, in English-German and English-French using the standard ContraPro and Large-scale Contrastive Pronoun test, respectively. Our experiments focused on several dimensions of analysis: (i) metric results on sentence-level and contrastive sets in terms of BLEU and accuracy, (ii) data distribution bias, (iii) part-of-speech of the antecedents, (iv) context complexity in terms of length and number of potential antecedents, and (v) syntactic functions of the pronoun and the antecedent.

Our results confirm the ability of context-aware models based on concatenation approaches to improve the accuracy of neural machine translation, particularly for pronominal categories affected by bias. Integrating target information was shown to be particularly beneficial across experiments, with 2to2 models achieving the best results overall.

The part of speech of the antecedent in source sentences was shown to be impactful, once translation bias towards the most frequent pronouns was accounted for. Models that made use of source context were thus shown to perform better when the tag of the antecedent was of a nominal type, as opposed to uninformative antecedents, in contrast with models relying on the target context.

Context complexity, in terms of either length or number of potential antecedents, was shown to be impactful for English-German, but less conclusively so for English-French. Further analyses on other datasets would be needed to properly assess the impact of context complexity.

We also found that the syntactic function of pronouns and antecedents was a determining factor for all models, with a similar tendency across models and language pairs for context information to be better exploited when both elements shared the same syntactic tag and the antecedent was the subject of the context sentence. Function identity with non-subject antecedents performed as a distant second overall, followed by different tags with subject antecedents, and finally by dissimilar tags and non-subject antecedents.

These results highlight current limitations of concatenated context-aware models, which seem to mainly capture the most regular and simpler configurations. It might be worth developing new contrastive test sets with higher variability to more

precisely assess the strengths and limitations of context-aware models.

It is worth noting that our analysis was conducted using the reference target context in our evaluations with target-dependent models, and further analyses would be necessary to determine the impact of using translated target context sentences instead of references. Additionally, a more detailed analysis based on different pronominal categories could also be helpful, although this was beyond the scope of this work. We also leave for future work further explorations of the differences between models that use source information and those that use target information, as well as including other types of models that are not based on input concatenation for context modelling.

References

Bahdanau, Dzmitry, 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*.

Bawden, Rachel, 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.

Fernandes, Patrick, 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.

Gete, Harritxu, 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.

Guillou, Liane 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.

Guillou, Liane, 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.

Jean, Sebastien, 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.

Junczys-Dowmunt, Marcin, 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.

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

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

Koehn, Philipp, 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.

Koehn, Philipp. 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, July. Association for Computational Linguistics.

Läubli, Samuel, 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.

Li, Bei, 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.

Lison, Pierre, 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.

Lopes, António, 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.

Ma, Shuming, 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.

Majumde, Suvodeep, 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*.

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

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

Müller, Mathias, 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.

Post, Matt. 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, October. Association for Computational Linguistics.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online.

Rauf, Sadaf Abdul and François Yvon. 2020. *Document level contexts for neural machine translation*. Ph.D. thesis, LIMSI-CNRS.

Sennrich, Rico, 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.

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

Sun, Zewei, 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.

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

Tiedemann, Jörg 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.

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

Vaswani, Ashish, 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.

Voita, Elena, 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.

Voita, Elena, 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.

Wang, Longyue, Zhaopeng Tu, Andy Way, and Qun Liu. 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.

Zhang, Jiacheng, 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.