On Measuring Context Utilization in Document-Level MT Systems

Anonymous ACL submission

Abstract

Document-level translation models are usually evaluated using general metrics such as BLEU, which are not informative about the benefits of context. Current work on context-aware evaluation, such as contrastive methods, only measure translation accuracy on words that need context for disambiguation. Such measures cannot reveal whether the translation model uses the correct supporting context. We propose to complement accuracy-based evaluation with measures of context utilization. We find that perturbation-based analysis (comparing models' performance when provided with correct versus random context) is an effective measure of overall context utilization. For a finer-grained phenomenon-specific evaluation, we propose to measure how much the supporting context contributes to handling the phenomena. We show that automatically-annotated supporting context gives similar conclusions to human-annotated context and can be used as alternative for cases where human annotations are not available. Finally, we highlight the importance of using discourse-rich datasets in any such endeavor.[1]

1 Introduction

Documents are one of the primary ways in which we produce and consume text. While for some languages, sentences provide a base unit of meaning, there are many sentences that contain discourse phenomena that are difficult to disambiguate at sentence level (Figure 1). Despite the vital need for document-level translation systems in order to handle those context-dependent phenomena, most of the current works on machine translation focus on sentence-level translation. Post and Junczys-Dowmunt (2023) listed the problem of evaluation as one of the reasons for the inability to move beyond sentence-level translation. In this work, we focus on this problem of evaluation. In particular, we focus on evaluating document-level translation models based on how well they utilize the inter-sentential information provided when translating at the document level.

The research on document-level translation evaluation has progressed significantly. Early works used general metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) which proved to be inadequate for capturing improvements in discourse phenomena. Subsequent research introduced phenomena-specific automatic metrics and contrastive test suites. Maruf et al.'s (2022) survey includes a comprehensive list of works in this direction. While these metrics provide an accuracy measure of models' performance on phenomena, they do not account for correct context utilization. Unlike prior studies, we adopt an interpretable approach to context utilization evaluation. We evaluate models based on their ability to use the correct context, and not only the ability to generate the correct translation without necessarily utilizing the context.

To assess the models' correct context utilization, we perform a perturbation-based analysis. Previous studies in perturbation analysis, such as those conducted by Voita et al. (2021), Li et al. (2020), and Rikters and Nakazawa (2021), were limited to specific architectures, evaluated on particular metrics, and perturbed only the source context. In our more comprehensive study, we analyze performance across various document-level architectures using multiple metrics including BLEU and CXMI (Fernandes et al., 2021). Additionally, our analysis involves perturbing both source and target contexts to examine the influence of both sides.

For more fine-grained analysis at the level of a specific discourse phenomenon, Yin et al. (2021) collected annotations of supporting context words from expert translators for the pronoun resolution phenomenon. However, they propose using such annotations as supervision to guide models' attention.

[1]Code will be released in the paper's final version.

In contrast, we focus on evaluating context-aware models' performance on the phenomenon. We evaluate models based on the attribution scores of supporting context. To obtain attribution scores, we use one of the state-of-the-art interpretability methods for transformer models: ALTI+ (Ferrando et al., 2022). Moreover, we use automatically annotated (using coreference resolution models) supporting context as an alternative to human annotated context and show that it gives similar conclusions. Using automatic annotations allowed us to scale to different languages and has the potential to extend to other discourse phenomena.

As an accuracy measure on discourse phenomena, Fernandes et al. (2023) proposed a novel systematic approach to tag words in a corpus with specific discourse phenomena and evaluate models' performance using the F1 measure. However, they mention that context-aware models make only marginal improvements over context-agnostic models. Our analysis reveals that this in fact depends on the richness of the dataset with phenomena, and that challenge sets curated to target context-dependent discourse phenomena are better in distinguishing the differences between models in handling the phenomena.

Our contributions are the following:

- We perform a perturbation-based analysis on different document-level MT models and find that the single-encoder concatenation model is able to make use of the correct context vs. a random context.
- We propose the use of attribution scores of *supporting context* to evaluate correct context utilization. We analyze the pronoun resolution phenomenon as a case study and find that sentence-level models and single-encoder context-aware models are better than multi-encoder models in the amount of attribution that the pronoun's antecedents have to generating the pronoun.
- We propose the use of automatically annotated *supporting context* as an alternative to human-annotated context to perform the attribution evaluation. We show that, despite noise in automatic annotation, results are consistent with the human-annotated context, paving the way towards efficient use of linguistic expertise in document-level translation evaluation.

[EN] One of the Chinese worked in an amusement park. **It** was closed for the season.

[DE] Ein Chinese arbeitete in einem Vergnügungspark. **Er** war gerade geschlossen.

Figure 1: An example illustrating the pronoun resolution phenomena which can not be disambiguated at sentence level. The pronoun **It** is ambiguous and its translation depends on the antecedent.[2]

- We highlight the importance of using a discourse rich dataset when evaluating the ability of models to handle context-dependent discourse phenomena.

2 Background

Sentence-level MT models treat sentences in a document as separate units. They only consider intra-sentential dependencies. In contrast, document-level models take into account intra-sentential as well as inter-sentential dependencies. Formally, if we consider a document $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, the probability of translating sentence x_i into y_i using a sentence-level model is

$$P(y_i|x_i) = \prod_{t=1}^{T} P(y_{i,t}|y_{i,<t}, x_i),$$

while the probability using a document-level translation model with context C_i is:

$$P(y_i|x_i, C_i) = \prod_{t=1}^{T} P(y_{i,t}|y_{i,<t}, x_i, C_i),$$

where T is the number of tokens in the target sentence y_i.

There are several ways to design neural architectures for document-level MT. The main architectures developed so far can be broadly classified into two categories based on how they combine the context and current sentence representations: single-encoder and multi-encoder approaches.

2.1 Single-Encoder Approaches

The single-encoder approach to document level MT works by concatenating previous sentences to the current sentence separated by a special token. It is commonly deployed under two setups: a 2-to-2

[2]Example is drawn from ContraPro dataset https://github.com/ZurichNLP/ContraPro

setup in which the previous and current source sentences are translated together, the translation of the current source sentence is then obtained by extracting tokens after the special concatenation token on the target side, and a 2-to-1 setup where the concatenation happens only in the source side, the target in this case is only the current sentence translation (Tiedemann and Scherrer, 2017; Bawden et al., 2018).

2.2 Multi-Encoder Approaches

The multi-encoder approach uses extra encoders for source and target contexts. The encoded representations of the context and current sentences are combined together before being passed to the decoder. There are different ways to combine the context and current sentence representations. Methods in the literature include concatenation, hierarchical attention, and attention gating (Libovický and Helcl, 2017; Zoph and Knight, 2016; Wang et al., 2017; Bawden et al., 2018).

3 Experimental details

3.1 Data

We train our models on IWSLT2017 TED data (Cettolo et al., 2012). We consider two language pairs in our experiments, namely EN → DE and EN → FR. For EN → DE, we use the same splits used by Maruf et al. (2019); we combine `tst2016-2017` into the test set and the rest are used for development. For EN → FR, we use the same splits as Fernandes et al. (2021); we use the test sets `tst2011-2014` as validation sets and `tst2015` as the test set.

3.2 Models

For both language pairs, we consider an encoder-decoder transformer architecture as our base model (Vaswani et al., 2017). Similar to Fernandes et al. (2021), we train a transformer small model (hidden size of 512, feedforward size of 1024, 6 layers, 8 attention heads). All models are trained on top of Fairseq (Ott et al., 2019). We use the same hyper-parameters as Fernandes et al. (2021), we train using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and use an inverse square root learning rate scheduler with an initial value of 5×10^{-4} and with a linear warm-up in the first 4000 steps. We train the models with early stopping on the validation perplexity. For models that use context, we train the models using a dynamic context size of 0–5 previous source and target sentences to ensure robustness against varying context size, as recommended by Sun et al. (2022). We develop three models for our evaluation experiments:

- **A sentence-level model:** As in Figure 2a, we train an encoder-decoder model on sentence-level data. This model has two evaluation setups: a sentence-level and a document-level setup. When evaluating at the sentence level, we refer to this model as the sentence-level model. To perform document-level evaluation, context and current sentences are concatenated with a special separator token in between them; this is referred to as the sentence-level* model in the rest of the paper.

- **A single-encoder concatenation model:** As seen in Figure 2b, we use the 2-to-2 setup (§2.1) with a sliding window across sentences in each document, allowing us to consider both source and target contexts. We will refer to this model as the concatenation model in the rest of the paper.

- **A multi-encoder concatenation model:** As in Figure 2c, we add two extra encoders to represent source and target contexts. The outputs of the three encoders are concatenated before being passed to the decoder. We will refer to this model as the multi-encoder model in the rest of the paper. Per §2.2, there are other methods to combine the outputs of multiple encoders beyond concatenation. However, we opt for concatenation due to its simplicity and its comparable BLEU performance to other architectures, as presented in Bawden et al. (2018).

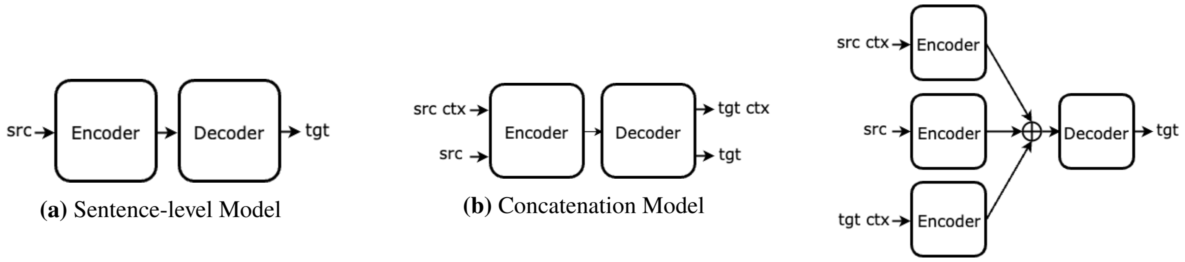
4 Method

Our goal is to build interpretable metrics to measure the extent of context utilization in context-aware MT. To this end, we propose two methods: a perturbation analysis and an attribution analysis.

4.1 Perturbation-Based Analysis

We look at the difference in performance when passing the correct versus random tokens as context. The correct context is the previous 5 sentences on source side, and the previous 5 generated translations on the target side.[3] To generate random context, we sample random tokens from the model's

[3]We avoid using the gold target context at inference time to eliminate exposure bias.



(a) Sentence-level Model **(b)** Concatenation Model **(c)** Multi-encoder Model

Figure 2: Model architectures for different settings. src & tgt refer to the current source and target sentence pair. src ctx & tgt ctx refer to the previous source and target sentence pairs used as context. In the concatenation model, the context and current sentences are concatenated together with a special separator token in between them. In the multi-encoder model, the ⊕ symbol refers to a concatenation operation.

vocabulary with a size similar to the correct context size. We compare models across BLEU and CXMI (conditional cross-mutual information (Fernandes et al., 2021)) metrics. CXMI is used to measure context usage by comparing the model distributions over a dataset with and without context . It should be noted that the numerical CXMI value cannot be compared across models since the multi-encoder model has a different number of parameters which will affect the probability distribution learned by the model. Therefore, we mainly focus on the sign of the CXMI value for the comparison. A positive CXMI value means that introducing context increases the probabilities assigned by the model to output tokens, and a negative CXMI means that the context is reducing them. Formally, for a source–target pair (x, y) and a context C, it reads:

$$\mathrm{CXMI}(C \rightarrow y|x) = H_{q_{\mathrm{MT}_A}}(y|x) - H_{q_{\mathrm{MT}_C}}(y|x, C),$$

where $H_{q_{MT_A}}$ is the entropy of the context agnostic model and $H_{q_{MT_C}}$ is the entropy of the context-aware model. In our analysis, we evaluate the same model with and without context, i.e., $q_{\mathrm{MT}_A} = q_{\mathrm{MT}_C} = q_{\mathrm{MT}}$

As for the BLEU score, we directly compare the numerical value of the BLEU score in the correct vs. random context setup. We compute:

$$\Delta = \mathrm{BLEU}(\mathrm{correct}) - \mathrm{BLEU}(\mathrm{random}).$$

The higher the Δ, the better the model at utilizing the correct context.

4.2 Attribution Analysis

In this experiment, we measure the attribution of supporting context words to model predictions. By *supporting context words*, we mean the words that are necessary to resolve context-dependent phenomena. For example, in case of pronoun resolution, the supporting context words are the pronoun's antecedents.

We look at the percentage of attribution of pronoun antecedents to generating a pronoun against the attribution of the entire input. We make use of the ContraPro contrastive evaluation dataset for the analysis. For EN → DE, the dataset considers the translation of the English pronoun *it* to the three German pronouns *er*, *sie* or *es*. It consists of 4K examples per pronoun (Müller et al., 2018). For EN → FR, the dataset concerns the translation of the English pronouns *it*, *they* to their French correspondents *il*, *elle*, *ils*, and *elles*. It includes 14K samples evenly split across the pronouns (Lopes et al., 2020). In particular, we use a subset of the data that has an antecedent distance between 1–5 since we are using 5 previous sentences as context.[4]

The attribution method we used is the ALTI+ (Aggregation of Layer-wise Token-to-token Interactions) method (Ferrando et al., 2022), which has been shown to be effective in explaining model behaviors (e.g. detecting hallucinations (Dale et al., 2023)). ALTI+ is designed to work for any encoder-decoder transformer model, so it can be readily applied for both the sentence-level and concatenation models. However, further consideration is needed to apply it in the multi-encoder setup. In the multi-encoder model, the input consists of separate source context, source, and target context sequences $x = [x_{sc}, x_s, x_{tc}]$. Each sequence is encoded separately by a different encoder giving ALTI contribution matrices $C^{enc_{sc}}_{e_{sc} \leftarrow x_{sc}}$, $C^{enc_s}_{e_s \leftarrow x_s}$ and $C^{enc_{tc}}_{e_{tc} \leftarrow x_{tc}}$,

[4]For EN→DE, we exclude 2400 examples with antecedent distance 0, and 118 examples with a distance greater than 5. for EN→FR, 5986 examples with antecedent distance 0 are excluded.

	antecedents	context	current
ContraPro DE			
sentence-level	0.00	0.00	100
sentence-level*	1.69	89.71	10.29
concatenation	2.86	78.09	21.91
multi-encoder	0.07	2.36	97.64
ContraPro FR			
sentence-level	0.00	0.00	100
sentence-level*	3.57	84.38	15.62
concatenation	2.59	76.19	23.81
multi-encoder	0.25	3.07	96.93

Table 1: The percentage of attribution of pronouns' antecedents, the entire context words, and current sentence words to generating the ambiguous pronoun in ContraPro dataset.

respectively. Since we concatenate the output of each encoder giving $e = [e_{sc}, e_s, e_{tc}]$, the overall encoder contribution matrix is block diagonal:

$$C^{enc}_{e\leftarrow x} = \begin{bmatrix} C^{enc_{sc}}_{e_{sc}\leftarrow x_{sc}} & 0 & 0 \\ 0 & C^{enc_s}_{e_s\leftarrow x_s} & 0 \\ 0 & 0 & C^{enc_{tc}}_{e_{tc}\leftarrow x_{tc}} \end{bmatrix}$$

The rest of ALTI+ algorithm proceeds unchanged.

We obtain word-level attributions scores and then compute the percentage of the sum of attributions of source and target antecedent words against the total attribution of the entire input.[5]

5 Results and Discussion

5.1 Are Models Sensitive To The Correct Context?

Results of the perturbation analysis are shown in Table 2. For both language pairs, the concatenation model is making use of correct context tokens, and presenting random context tokens to the model results in worse BLEU performance and a negative CXMI value. Even though the sentence-level model has a high BLEU score, its performance drops significantly when evaluated at the document level (sentence-level*). This is expected; since the model has not been trained on longer contexts. Regarding the multi-encoder model, even though it has the best BLEU performance, the consistent performance of the model with correct and random context suggests that it is not utilizing the correct context, as can also be confirmed by the low or negative CXMI values. This analysis highlights the importance of looking beyond the BLEU score when evaluating context utilization of document-level MT models.

5.2 Are Models Paying "Attention" To The Supporting Context?

We obtain the attribution scores of the *supporting context* provided in the ContraPro pronoun resolution dataset. The *supporting context* is automatically generated using coreference resolution tools. Looking at Table 1, we can see that the sentence-level* model and the concatenation model have higher attribution scores compared to the multi-encoder model. This can also be confirmed by the low overall context attribution compared to the current sentence attribution in the multi-encoder model. It should be noted that our implementation of the multi-encoder model depends on simple concatenation of the encoders' outputs before being fed to the decoder. More complicated multi-encoder setups (*e. g.*, using gating mechanisms or hierarchical attention) might have better context attribution. Moreover, for German pronouns, looking at the total context contributions, we observe that despite the fact that the sentence-level* model has the highest context attributions, it is not the best at utilizing the *supporting context*. This highlights the importance of focusing on important parts of the context when evaluating context utilization.

5.3 Does automatically Annotated Supporting Context Align With Human Annotated Supporting Context?

We investigate whether the automatically annotated *supporting context* aligns with the way humans utilize context for pronoun disambiguation. We use the SCAT (Supporting Context for Ambiguous Translations) data provided by Yin et al. (2021) which contains human annotations of *supporting context* for pronoun resolution on the French ContraPro data. We filter the data for instances that has an antecedent outside the current sentence and end up with 5961 instances for evaluation. We calculate the attribution scores of human context for the models we built for EN→FR translation. Comparing the attribution percentages in Table 3 to the attributions on ContraPro FR data in Table 1, we observe similar trends across models. The sentence-level* and concatenation models have comparable attribution scores and are higher than the multi-encoder model. This shows that automatically annotated context can be a good alternative to human annotations which are expensive to obtain at scale.

[5]We compute the scores for the first occurrence of the antecedent. This might penalize a model that pays attention to another occurrence of the antecedent. This is rare: the average number of antecedents is 1.09 for DE and 1.18 for FR.

context setup	BLEU			CXMI		
	random	correct	Δ	no-context	random	correct
EN→DE						
sentence-level	–	–	–	23.22	–	–
sentence-level*	2.49	3.54	1.05	–	–2.98	–2.10
concatenation	20.24	23.32	**3.08**	23.39	–0.32	0.014
multi-encoder	23.72	**23.72**	0.00	23.71	–0.002	–0.002
EN→FR						
sentence-level	–	–	–	36.17	–	–
sentence-level*	5.61	9.36	3.75	–	–2.95	–1.84
concatenation	27.90	35.57	**7.67**	35.82	–0.32	0.006
multi-encoder	36.85	**36.85**	0.00	36.63	0.002	0.002

Table 2: BLEU and CXMI scores of correct vs. random context on IWSLT2017 test data. The best BLEU score in a correct setup (with context for the concatenation and multi-encoder models and without context for the sentence-level model) is bolded. Δ represents the difference in BLEU scores between the correct and random context setups, the higher the difference, the better the model in utilizing the correct context. The best Δ is highlighted in bold. A positive CXMI value means that the probabilities of output tokens are increased with context while a negative CXMI value means that context is reducing them.

model	antecedents	context	current
sentence-level	0.00	0.00	100
sentence-level*	1.25	87.12	12.88
concatenation	1.03	74.23	25.77
multi-encoder	0.53	2.49	97.5

Table 3: Attribution percentages of human annotated antecedents, the entire context words, and current sentence words to generating the ambiguous pronoun in SCAT dataset.

5.4 Are Models Able To Handle Context-Dependent Phenomena?

The ultimate goal of context-aware MT is being able to model context-dependent phenomena. Hence, we evaluate models on their ability to address these phenomena. We use the Multilingual Discourse Aware benchmark (MuDA) to automatically tag datasets with context-dependent phenomena (Fernandes et al., 2023). We considered 4 linguistic discourse phenomena in our analysis: lexical cohesion, formality, pronoun resolution and verb form. **Lexical cohesion** refers to consistently translating an entity in the same way throughout a document. **Formality** is the phenomenon where the second-person pronoun that the speaker uses depends on their relationship the the person being addressed. **Pronoun resolution** denotes the phenomenon in languages that use gendered pronouns for pronouns other than the third-person singular, or assign gender based on formal rules instead of semantic ones. **Verb form** denotes the phenomenon in languages with a fine-grained verb morphology, where the translation of the verb should reflect the tone, mood and cohesion of the document.

We use the IWSLT2017 test set as well as ContraPro data (including context sentences) in the analysis. Table 6 presents the statistics of discourse phenomena in these datasets. We then evaluate models using the F1 measure based on whether a word tagged in the reference exists and is also tagged in the hypothesis. As can be seen in Table 6, for both language pairs, ContraPro dataset has a higher percentage of tokens tagged with pronouns (since the dataset targets this phenomena). Looking at the F1 measure of models on this dataset in Table 5, we can see that the concatenation model has a higher score compared to other models which is reflected in the ContraPro accuracy as well (Table 4). On the other hand, the lower percentages of phenomena in the IWSLT data results in similar performance across models on this data. We highlight the importance of using a discourse rich dataset to benchmark models' performance on handling context-dependent phenomena. Evaluation on other discourse phenomena, which neither of the datasets targeted, resulted in no distinction between the models as seen in Tables 7 and 8. The low F1 measure of the sentence-level* model across phenomena on the IWSLT data can be linked to its low translation performance as presented in §5.1. Surprisingly on the other hand, for the more challenging ContraPro data, its performance is comparable to other models.

Moreover, we show that *supporting context* attri-

	EN→DE		EN→FR	
Context size	**0**	**5**	**0**	**5**
sentence-level	42	–	76	–
sentence-level*	–	47	–	81
concatenation	45	58	76	85
multi-encoder	43	43	76	75

Table 4: ContraPro contrastive accuracy (%) for different context sizes. The accuracy is calculated based on the percentage of time a model correctly scores a positive example above its incorrect variant.

	EN→DE		EN→FR	
Model	**IWSLT**	**CPro**	**IWSLT**	**CPro**
sentence-level	0.62	0.39	0.70	0.44
sentence-level*	0.38	0.45	0.53	0.48
concatenation	0.60	0.48	0.67	0.49
multi-encoder	0.61	0.40	0.70	0.44

Table 5: F1 measure of models on pronoun resolution phenomena on IWSLT and ContraPro data. The F1 measure is evaluated based on if a word tagged with a discourse phenomena in the reference exists and is also tagged in the hypothesis

bution should be considered as a separate evaluation dimension from translation quality using Pareto-style plots: Figure 3 shows the Pareto plot of two evaluation methods for EN→FR pronoun resolution: the F1 measure and the supporting context attribution percentage. It can be noticed that the multi-encoder model is sub-optimal on both dimensions, while the sentence-level* and concatenation methods present a trade-off. furthermore, despite the comparable F1 measure of the sentence-level to the multi-encoder model, it has zero attribution.

5.5 Discussion

Previous sections outlined different evaluation techniques for assessing context utilization of document-level MT models. These evaluations are complementary to each other and equally important. We start with a perturbation analysis to confirm whether the model is utilizing the correct context and it is not just acting as regularization. furthermore, we show that utilizing the correct context is not enough to handle context dependent phenomena; since not all context is important. Therefore, for a more fine-grained evaluation, we assess models in how well they utilize the parts in the context that are necessary to handle the phenomena. For this purpose, We use attribution scores supported with an accuracy evaluation (F1 measure) on the phenomena.

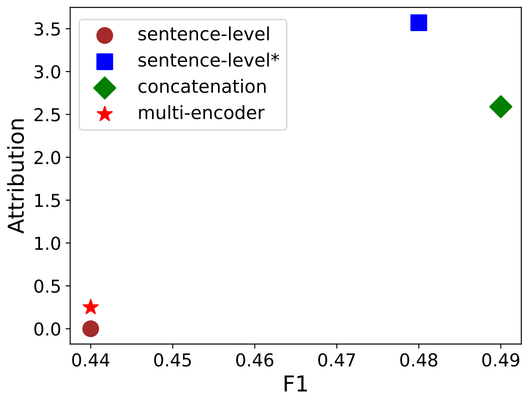


Figure 3: Pareto plot for EN→FR pronouns. The plot shows that attribution evaluations and accuracy based evaluations are complementary to each other. In particular, there is a trade-off between the sentence-level* and concatenation models, while the multi-encoder and sentence-level models are dominated.

Overall, our study highlights the important aspects to consider when evaluating context utilization: the use of correct context, the utilization of the correct parts of the context, the accuracy performance on the discourse phenomena, in addition to the general translation performance of course.

6 Related Work

Previous studies on evaluating context influence on MT performance often examined specific context-aware architectures or particular discourse phenomena. Nayak et al. (2022) explored context effects on the hierarchical attention context-aware MT model, showing that the improved performance on general metrics is due to a context-sensitive class of sentences. Bawden et al. (2018) improved the multi-encoder model by encoding the source and context sentences separately while concatenating the current and previous target sentences on the decoder side, demonstrating the importance of target-side context. In contrast, we offer a generalizable approach applicable to any context-aware MT model. While we focus on pronoun resolution, our tools can extend to various linguistic phenomena given appropriate rules for annotating *supporting context*.

In comparing various document-level models, Huo et al. (2020) found performance variation based on tasks, with no universally superior model. They also highlight back-translation's benefit to document-level systems, noting their robustness against to sentence-level noise. Unlike their general metric approach, we enhance the analysis using

Dataset	pronouns	cohesion	formality	verb form	no. sentences	no. tokens
EN→ DE						
IWSLT	180 (0.4)	569 (1.4)	641 (1.5)	–	2,271	40,877
ContraPro	14,477 (2.4)	87 (0.01)	9,710 (1.6)	–	70,718	599,197
EN→FR						
IWSLT	311 (1.2)	150 (0.6)	329 (1.3)	787 (3.1)	1,210	25,638
ContraPro	22,810 (2.6)	195 (0.02)	10,505 (1.2)	16,211 (1.8)	81,989	865,890

Table 6: Discourse phenomena statistics in different datasets along with the total number of the sentences and tokens in each dataset. The numbers between brackets represent the percentage of tokens tagged against the total number of tokens.

Model	cohesion	formality
IWSLT		
sentence-level	0.68	0.67
sentence-level*	0.20	0.29
concatenation	0.67	0.68
multi-encoder	0.66	0.67
ContraPro		
sentence-level	0.29	0.31
sentence-level*	0.24	0.33
concatenation	0.27	0.35
multi-encoder	0.31	0.33

Table 7: F1 measure of models on lexical cohesion and formality phenomena on ContraPro and IWSLT datasets for EN→DE

Model	cohesion	formal	vb. form
IWSLT			
sentence-level	0.81	0.71	0.42
sentence-level*	0.36	0.45	0.13
concatenation	0.81	0.75	0.42
multi-encoder	0.82	0.74	0.43
ContraPro			
sentence-level	0.58	0.32	0.28
sentence-level*	0.53	0.31	0.26
concatenation	0.56	0.32	0.28
multi-encoder	0.58	0.33	0.29

Table 8: F1 measure of models on lexical-cohesion, formality and verb-form phenomena on ContraPro and IWSLT datasets for EN→FR

perturbation methods and attribution evaluation.

In interpreting context's benefits, Kim et al. (2019) quantified the causes of improvements of context-aware models on general test sets using attention scores. They found that context usually acts as a regularization and is rarely utilized in an interpretable way. Our work differs in that we use ALTI+ attribution scores instead of attention scores to interpret models' behaviors.

In a concurrent work, Sarti et al. (2023) introduced an end-to-end interpretability pipeline for analyzing context reliance in context-aware models. In contrast, we we rely on linguistic rules instead of attention weights or gradient norms to extract contextual cues, which we show to align with human annotated cues. Additionally, we use attribution scores to compare different MT models, including single– and multi-encoder ones.

7 Conclusion

In this work, we shed light on multiple angles to look from when evaluating context utilization in document-level MT. We use a perturbation-based analysis to investigate correct context utilization. Additionally, for phenomena-specific evaluation, we propose using attribution scores as measure context utilization. We suggest calculating the attributions of only the supporting context that is necessary for handling context-dependent phenomena. Moreover, we show that automatically annotated supporting context is inline with human annotated supporting context and can be used as an alternative. Finally, we highlight the importance of using discourse-rich data in evaluation.

Based on our proposed analysis and evaluation tools, we argue that the single encoder approaches to document-level MT demonstrate a priori better context use while also scoring high for translation quality, suggesting that multi-encoder models need more careful design or tuning as highlighted by Rikters and Nakazawa (2021).

For future work, we aim to extend attribution evaluation to other discourse phenomena, by designing rules for automatic annotation of supporting context for the phenomena with the aid of linguistic expertise. We would also like to apply our evaluation tools and setups to different document-level architectures to provide a solid benchmark of context utilization by context-aware models.

Limitations

One limitation is that our conclusions regarding the multi-encoder model are considering only one instance of the multi-encoder approaches to document-level MT. We do not claim that all multi-encoder approaches to document-level MT will have low degrees of context utilization. We leave to to future work to investigate the context utilization of other multi-encoder approaches.

Due to the lack of *supporting context* annotations for discourse phenomena, we focused only on the pronoun resolution phenomena on two language pairs: EN→DE and EN→FR. However, we hope that this study opens the door towards works on automatic *supporting context* annotations for all identified discourse phenomena.

References

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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation, EAMT 2012, Trento, Italy, May 28-30, 2012*, pages 261–268. European Association for Machine Translation.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 36–50. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? A data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 606–626. Association for Computational Linguistics.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6467–6478. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8756–8769. Association for Computational Linguistics.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 604–616. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 24–34. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? A case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3512–3518. Association for Computational Linguistics.

Jindrich Libovický and Jindrich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202. Association for Computational Linguistics.

António V. 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, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 225–234. European Association for Machine Translation.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long *and Short Papers)*, pages 3092–3102. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 61–72. Association for Computational Linguistics.

Prashanth Nayak, Rejwanul Haque, John D. Kelleher, and Andy Way. 2022. Investigating contextual influence in document-level translation. *Inf.*, 13(5):249.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

Matiss Rikters and Toshiaki Nakazawa. 2021. Revisiting context choices for context-aware machine translation. *CoRR*, abs/2109.02995.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2023. Quantifying the plausibility of context reliance in neural machine translation. *arXiv eprint 2310.01188*.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006*, pages 223–231. Association for Machine Translation in the Americas.

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, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 82–92. Association for Computational Linguistics.

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: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1126–1140. Association for Computational Linguistics.

Longyue Wang, 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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2826–2831. Association for Computational Linguistics.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 788–801. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 30–34. The Association for Computational Linguistics.