Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention

Marco Dinarelli, Dimitra Niaouri, Fabien Lopez, Gabriela Gonzalez-Saez, Mariam Nacklé, Emmanuelle Esperança-Rodier, Caroline Rossi, Didier Schwab, Nicolas Ballier

Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention

Marco Dinarelli[1] — **Dimitra Niaouri**[1] — **Fabien Lopez**[1] — **Gabriela Gonzalez-Saez**[1] — **Mariam Nakhlé**[1,2] — **Emmanuelle Esperança-Rodier**[1] — **Caroline Rossi**[3] — **Didier Schwab**[1] — **Nicolas Ballier**[4]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, Grenoble, France.* [2] *Lingua Custodia.* [3] *Univ. Grenoble Alpes, ILCEA4.* [4] *Université Paris Cité, LLF & CLILLAC-ARP, Paris, France*

ABSTRACT. *Model explainability has recently become an active research field. Many works are published supporting or criticizing attention weights as model explanation. In this work we adhere to the former and analyze attention as explanation for Context-Aware Neural Machine Translation (CA-NMT). Since its evaluation often concerns the evaluation of models in resolving discourse phenomena ambiguity, we perform analyses and evaluations over coreference links in a parallel corpus. We propose a human evaluation over heatmaps, strengthened by a quantitative evaluation based on attention weights over coreference links and with different metrics purposely designed for this work. Such metrics provide a more explicit evaluation of the CA-NMT models than evaluations using contrastive test suites.*

RÉSUMÉ. *L'explicabilité des modèles est devenue un champ de recherche très actif. Beaucoup de travaux ont vu le jour, à la fois soutenant et critiquant l'utilisation de l'attention comme explication du comportement des modèles. Dans cet article, nous adhérons au premier type de travaux et analysons l'attention pour interpréter le comportement des modèles de traduction neuronale en contexte (CA-NMT). Puisque cette évaluation concerne souvent la résolution de l'ambiguïté des phénomènes discursifs, nous effectuons des analyses et évaluations sur les liens de coréférence annotés dans un corpus parallèle. Nous proposons une évaluation humaine sur des* heatmaps*, renforcée par une évaluation quantitative basée sur les poids d'attention des liens de coréférence, avec trois métriques conçues explicitement pour ce travail. Celles-ci constituent une évaluation plus directe des modèles pour la CA-NMT que celles fondées sur les* test suite *contrastives.*

KEYWORDS: *Machine Translation, Explainability, Coreference resolution, CA-NMT evaluation*

MOTS-CLÉS : *Traduction automatique neuronale, Explicabilité, Résolution de coréférences, Évaluation de la traduction automatique neuronale en contexte*

1. Introduction

Since the adaptation of the attention mechanism (Bahdanau *et al.*, 2014) to translation, its integration in neural models (Bahdanau *et al.*, 2014; Luong *et al.*, 2016), and its heavy use in several domains of computer science thanks to the invention of the *Transformer* model (Vaswani *et al.*, 2017), this mechanism has been used extensively to show and explain the behavior of neural models in performing predictions. In the original paper introducing the attention mechanism (Bahdanau *et al.*, 2014), authors draw attention weights to show how a neural end-to-end model for machine translation learns the alignment between source and target sentences. Since then, attention weights have been instrumental in providing visual explanation of models behavior. E.g. in Lee *et al.* (2017), authors show through attention weights the *soft-head* learned by the model to represent mentions in neural coreference resolution. In Darcet *et al.* (2023), attention is used to show the behavior of neural models for image classification.

While attention constitutes an intuitive mean to explain models' behavior visually, a whole research domain named *explainability* arose to understand how neural models store and use information based on probing models (Pasad *et al.*, 2021; de Seyssel *et al.*, 2022). This approach has been used especially for analyzing large neural models learned by self-supervision like *BERT* (Devlin *et al.*, 2019). In parallel, the attention mechanism has also increasingly been used for models' explainability (see Paul (2023) for an overview), but some doubts have been raised concerning whether attention is indeed explanation (Bibal *et al.*, 2022). In the context of Neural Machine Translation (NMT) for instance, Ding *et al.* (2019) started from the observation that attention weights may be inconsistent with the actual predicted target tokens when performing beam search, and proposed a solution based on *token saliency* computation. Despite doubts and counter examples of attention working as explanation, intuitively and empirically, that is visually, attention still constitutes a useful mean for understanding models behavior at inference phase.

In this paper, we analyze the behavior of context-aware neural machine translation (CA-NMT) systems on discourse phenomena, namely coreferences, using the attention weights over the current and the previous sentences, that is the context. While there have been already works in this respect (Tiedemann and Scherrer, 2017a; Jaziriyan and Ghaderi, 2023), most of the time the ability of CA-NMT systems to exploit a context is only measured indirectly and quantitatively through automatic metrics on the system output, such as BLEU (Papineni *et al.*, 2002) or COMET (Rei *et al.*, 2020) or on purposely designed contrastive test suites (Bawden *et al.*, 2018; Müller *et al.*, 2018; Voita *et al.*, 2019a; Lopes *et al.*, 2020) and other *challenge sets* (Isabelle *et al.*, 2017). While the latter are an interesting method for evaluating CA-NMT, they only provide an indirect evaluation as models are only asked to score sentences in their context, without having to actually generate them. We propose a human evaluation over heatmaps, strengthened by a quantitative evaluation based on attention weights over coreference links and with different metrics designed on purpose for this work. We believe such metrics constitute more explicit and direct evaluations of CA-NMT models' ability to use context than evaluations with contrastive test suites.

The rest of the paper is structured as follows. Section 2 summarizes previous research on NMT systems and contextualizes explainability for NMT and our contribution in this

respect. Section 3 presents our experimental methodology, the data and experimental design. Section 4 presents quantitative and qualitative results. In Section 5 we briefly discuss a particular aspect of models' behavior. Section 6 concludes the paper.

2. State of the Art and Related Work

Explainability in the context of NMT involves unravelling the decisions made by the model at different levels during the translation process in order to show the user how the system performs translation based on objective measures (Ali *et al.*, 2023). It comprises the provision of reasons for the model's output, with an ideal scenario ensuring that these explanations are meaningful, accurate, and bounded within the system's knowledge (Phillips *et al.*, 2021). In this paper, we explore explanations by leveraging the internal values of the NMT system, specifically focusing on the attention weights on coreference phenomena.

Attention weights have been used to compute feature attribution methods. These methods are used to explain the alignment of the source text to the translated text in NMT models. These methods measure the token-level importance obtained via input attribution methods with the generated output. Alvarez-Melis and Jaakkola (2017) used the attention scores to measure the relevance between two input-output tokens by perturbing the input sequence. He *et al.* (2019) used the same approach, changing the definition of relevance between the input and output attention scores integrating gradient based methods (Sundararajan *et al.*, 2017). Ding *et al.* (2019) proposed to compute saliencies to obtain word alignment interpretation of NMT prediction based on gradients and attention weights. Their results show that the gradient-based methods present lower alignment error rates than methods using attention weights. In the same line, Ghader and Monz (2017) showed that attention and conventional alignment methods exhibit certain similarities, although there are variations depending on the specific attention mechanism and the type of word being translated. Notably, their study revealed that attention patterns were influenced by the grammatical function of the target word.

In our exploration, we do not use gradient-based attribution methods to understand the importance of contextual text in translation. We exclusively rely on attention weights, emphasizing their suitability for exploring the real assignment values of source words and the relative importance of contextual words during translation. While some works question the explanatory power of attention, others acknowledge that it is one of the diverse explanation tools (Wiegreffe and Pinter, 2019). In this line, Vig and Belinkov (2019) posit attention mechanisms as valuable explanatory tools, particularly for tasks related to syntax in NLP.

Examining specific NMT and CA-NMT models, Yin *et al.* (2021) observed a reliance on source context over target context for pronoun and polysemous word disambiguation, highlighting the significance of attention scores on contextual words. Voita *et al.* (2018) delved into incorporating discourse phenomena, finding that attention weights played a pivotal role in capturing contextual information related to pronoun translation. Clark *et al.* (2019) investigated attention heads in BERT, revealing preferences for different types of information across layers, reinforcing attention as a plausible explanation for syntactic dependency tagging and coreference resolution. Raganato and Tiedemann (2018) further

emphasized the diverse semantic patterns identified in attention weights across different layers, reinforcing the nuanced interpretability provided by attention mechanisms. In this work, we propose to continue this research path analyzing and explaining how NMT models use context in translation, based on the attention weights.

The literature on models explainability is large (Bibal *et al.*, 2022), with a lot of works both sustaining and criticizing attention as explanation and interpretation mean. Works on tasks involving syntax and semantic seem to be more on the first category (Vashishth *et al.*, 2019), especially works on NMT (Vashishth *et al.*, 2019; Moradi *et al.*, 2021). In particular Vashishth *et al.* (2019) identifies the reason of wrong conclusions drawn from two important works refusing the explainability power of attention (Jain and Wallace, 2019; Serrano and Smith, 2019) in the use of classification models, opposed to sequence prediction models, like in NMT, where attention seems to play a crucial role. Wiegreffe and Pinter (2019) use two previously defined categories of explainability of attention weights named *plausibility* and *faithfullness*. The latter concerns the degree to which attention explains model's predictions. The first concerns how plausible models behavior is as externalized by attention weights. As such, it may concern any aspect of a model. In this paper we adhere to the view of the first class of works falling into the *plausibility* category, and basing our work on the intuitive, visual interpretability of attention as explanation. Specifically we aim at analyzing the attention behavior in attending to the context in NMT when the model faces discourse phenomena like coreferences.

All the analyses and evaluations proposed in this work are based on the simple observation that CA-NMT models have access to the context only through attention mechanisms. Together with the empirical evidence we observed in our model's output, these motivate our work. Indeed, the attention patterns we observed over coreferences in the data support the fact that, at least when correctly learned, attention mechanisms do show interpretable behaviors with respect to coreference phenomena. In order to guarantee correct learning of attention mechanisms, as much as possible, we fine-tuned our models on a larger amount of document-level data with respect to what was used by other models in the literature evaluated on the same benchmark.

Like in Bibal *et al.* (2022) and Vashishth *et al.* (2019), we perform a human evaluation of attention over heatmaps displaying the current and one of the context sentences. However on the one side, we do not encounter the same issues raised in this type of evaluation since in our case the phenomenon we observe (coreference) is annotated in the data, which enables a very precise evaluation of the targeted phenomenon. We note that other discourse phenomena may occur in the same sentences, however thanks to coreference annotation, and to a post-processing we performed, explained in Section 3.3, analyses on coreference phenomena are made easier. On the other side, in the context of CA-NMT the phenomenon we observe rarely impacts model's predictions and thus it is not often involved in the model's loss signal at training phase. As a consequence its correct behavior on coreferences cannot be necessarily judged through the model's prediction. The model can indeed correctly put attention on coreferent mentions regardless of whether these are correctly translated; and the model can correctly translate mentions, especially pronominal anaphora, without putting significant attention weight on their correct antecedents: this can happen for example when translating using

the most occurring word is correct. It can happen also for proper nouns. The intuition behind these behaviors is that models for sequence generation, like NMT models, should learn to some extent the language structure in order to solve effectively the problem they are designed for. The latter observations motivate our quantitative evaluations based on attention weights.

2.1. *Context-Aware Neural Machine Translation Models*

CA-NMT models can be classified in two main categories (Lupo *et al.*, 2022a), concatenation approaches and multi-encoder approaches.

2.1.1. *Concatenation approaches*

The concatenation approach simply consists in concatenating the context to the current sentence before feeding it to a standard encoder-decoder architecture (Tiedemann and Scherrer, 2017b; Agrawal *et al.*, 2018; Junczys-Dowmunt, 2019; Ma *et al.*, 2020; Zhang *et al.*, 2020). The context can be on the source side, the target side, or both. Generation can then follow two strategies: the *many-to-many* strategy consists in translating all the source sentences and discarding contextual sentences; the *many-to-one* strategy consists in translating the current sentence only. Although concatenation approaches have the advantage of using the same architecture as standard sentence-level NMT models, their context is limited to few sentences because the complexity of the attention mechanisms scales quadratically with sentence length, although some recent works try to provide solutions to this constraint (Wang *et al.*, 2020; Tay *et al.*, 2020).

2.1.2. *Multi-encoder approaches*

Multi-encoder models augment a standard sentence-level NMT system, with parameters θ_S, with additional modules that encode and integrate the context of the current sentence for modeling the context either on source side, target side, or both. These modules account for *contextual parameters* θ_C. The full context-aware architecture has parameters $\Theta = [\theta_S; \theta_C]$. Note that a model based on the concatenation approach can thus be characterized in terms of parameters with only θ_S. Most of the multi-encoder models can be described as instances of two architectural families (Kim *et al.*, 2019), differing in the way the representations of the context and the current sentence are integrated.

Outside integration. In this approach, the encoded representations are merged outside the decoder (Maruf *et al.*, 2018; Voita *et al.*, 2018; Zhang *et al.*, 2018; Miculicich *et al.*, 2018; Maruf *et al.*, 2019; Zheng *et al.*, 2020). This can happen in different ways, such as by simple concatenation of the encodings, or with a gated sum.

Inside integration. Here the decoder attends to the context representations directly, using its internal representation of the decoded history as query of the attention mechanism (Tu *et al.*, 2018; Kuang *et al.*, 2018; Bawden *et al.*, 2018; Voita *et al.*, 2019b; Tan *et al.*, 2019).

In many of these works parameters of current-sentence and context encoders are shared (Voita *et al.*, 2018; Li *et al.*, 2020). In this way, the number of contextual parameters to learn, $|\theta_C|$ and the computational costs are reduced.

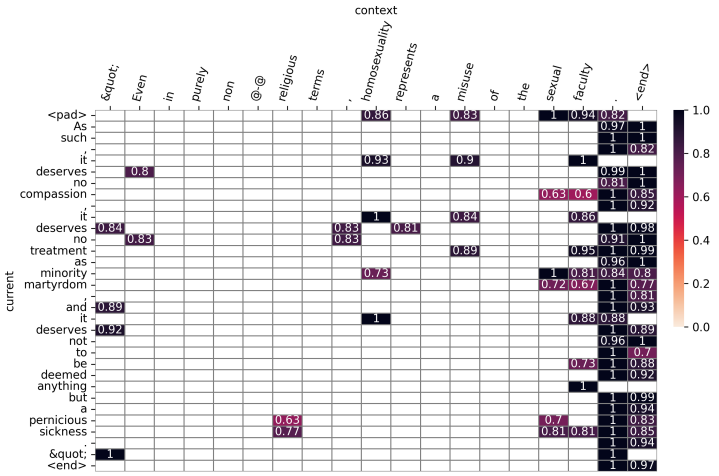

Figure 1. *Example of attention weights between current and context sentence from the multi-encoder model. This example can be compared with the one in Figure 2.*

2.1.3. *Two-step training*

CA-NMT models are commonly trained following a two-step strategy (Tu *et al.*, 2018; Zhang *et al.*, 2018; Miculicich *et al.*, 2018; Maruf and Haffari, 2018; Li *et al.*, 2020). The first step consists in training θ_S independently on a sentence-level parallel corpus. Then, in multi-encoder approaches, contextual parameters θ_C are trained on a document-level parallel corpus, while fine-tuning or freezing θ_S. In concatenation approaches θ_S are further tuned using document-level data.

2.1.4. *Attention Mechanism*

While the attention mechanisms used by multi-encoder and concatenation NMT models for attending to the context may have a functional difference, they can be generically defined in the same way in terms of sequences of queries, keys and values Q,K,V where each element $q_i \in \mathbb{R}^{d_1}, i = [1,...N]$, and $k_j, v_j \in \mathbb{R}^{d_2}, j = [1,...M]$. Attention weights are then computed as

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{j=1}^{M} exp(e_{ij})} \quad [1]$$

where e_{ij} computes an *association score* $a(q_i,k_j)$ between the query q_i and the key k_j. Attention weights α_{ij} are then used to obtain a weighted sum of values $c_i = \sum_j \alpha_{ij} v_j$, which results in a *contextualization* of queries with respect to the values.

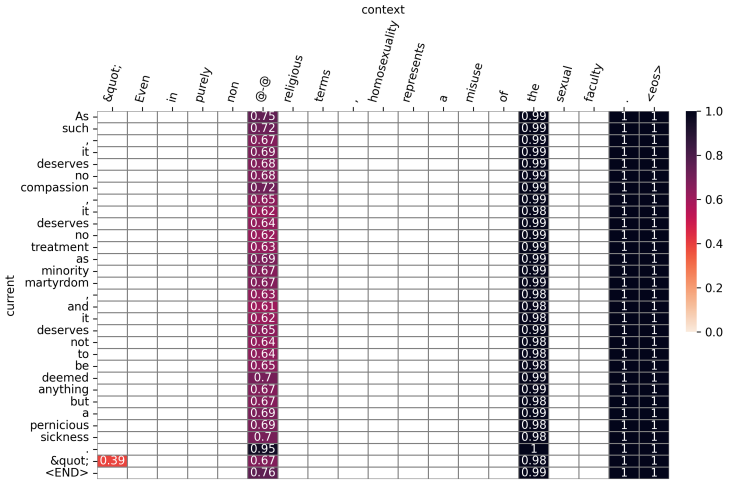

Figure 2. *Example of attention weights between current and context sentence from the concatenation model. This example can be compared with the one in Figure 1.*

3. Methodology

In this section we detail the whole experimental procedure and evaluation, starting by introducing the models we employed for CA-NMT.

3.1. *Employed CA-NMT Models*

In this work we analyze two CA-NMT models, one from each of the two broad approaches introduced in Section 2.1. Namely we use a variant of the multi-encoder Hierarchical Attention Network (HAN) approach proposed in Lupo *et al.* (2022a) where we exploit only the source-side context; as second model we use a concatenation approach based on the *Transformer* model proposed in Lupo *et al.* (2022b) which uses both source and target context. The two models keep a standard Transformer architecture, that is they have 6 encoder and decoder blocks with 512 dimensional hidden layers, 2,048 dimensional FFNN hidden layers, 8 attention heads. The number of token embeddings is determined by the use of BPE. We used a dictionary size of 32,000, sharing input and output vocabulary, as used often in the literature with the same data. The other hyper-parameters, including those for model training, are the same as in Vaswani *et al.* (2017).

Both concatenation and multi-encoder models can potentially process any number of context sentences, from the past or future. However, most of the approaches proposed in the literature focus on a few previous sentences, where most of the relevant context is concentrated, but also to reduce the computational cost related to the attention mechanism's complexity.

In the *self-attention* mechanism of *Transformers* (Vaswani *et al.*, 2017), used in the concatenation NMT model for attending to the context, queries, keys and values are the same vectors. In the HAN module (Miculicich *et al.*, 2018), used in the multi-encoder model, queries are hidden states of the encoder for the current sentence, keys and values are previously encoded hidden states of the encoder for the context sentences. The functioning of the attention for attending to the context is thus the same in the two models, the difference is that the multi-encoder model attend to each context sentence individually, the second level HAN mechanism allows the model to distinguish between different context sentences. The concatenation model attend to all the context sentences at the same time.

Equation 1 for computing attention weights implies that attention weights α_{ij} sum up to 1 over keys. As a consequence, since the concatenation model attends to the context with the self-attention module over the concatenation of context sentences to the current sentence, attention weights in the concatenation model are smaller than weights in the multi-encoder model, which make them not comparable. To overcome this issue we applied a post-processing on attention weights, detailed in Section 3.3.

3.2. *Dataset*

For the analyses and the evaluation focused on the ability of CA-NMT models in using context, we exploit the *ParCorFull2* corpus (Lapshinova-Koltunski *et al.*, 2022). This corpus is provided in four different languages: English, French, German and Portuguese. Data in all languages are document-level and are annotated with coreferences. Coreferences are mentions to the same entities of the world. For example (Lapshinova-Koltunski *et al.*, 2022):

... not to mention social networking platforms, allow [people]$_1$ to self-identify, to claim [their]$_1$ own descriptions of [themselves]$_1$, so [they]$_1$ can go align with global groups of [their]$_1$ own choosing.

All mentions in *[]* with the same index refer to the same entity of the world, they are coreferences. The *ParCorFull2* corpus contains not only pronominal anaphora, which are the most common examples of coreferences, but also coreferences involving noun phrases, elliptical constructions, clauses or set of clauses. This comes from the choice of the authors to annotate events as antecedents.

We perform the analyses and evaluations on the English-German language pair only. Our analyses are performed on the source-side first-level *HAN* attention in the multi-encoder model (we refer the reader to Miculicich *et al.* (2018) for details), and on the *self-attention* mechanisms of the encoder in the concatenation model, which is the attention attending to the source context. We do not analyse *cross-attention* mechanisms in any model. While this mechanism may be forced to attend to the context for coreference disambiguation by the loss function training signal, since it learns the alignment between source and target sentences, its functioning from an explainability perspective is more complex, and there can be interference because encoder's hidden states are already contextualized through attending to the source context.

Language	*Sentences*	*Tokens*	*Mentions*	*Coref. chains*
English	2,280	42,798	4,206	425
German	2,280	40,261	3,377	306

Table 1. *Statistics on the English-German data from the ParCorFull2 corpus used for our analyses.*

Language	*Sentences*	*Tokens*	*Mentions*	*Coref. chains*
English	74	557	135	73
German	74	605	132	73

Table 2. *Statistics on our selected sentences for human evaluation.*

Some statistics of the data used for our analyses are depicted in Table 1. For our human evaluation we selected a subset of such data made of 73 examples of coreference links. Statistics are shown in Table 2. The column *Mentions* shows the number of annotated coreferent mentions, while in the column *Coref. chains* is reported the total number of coreference chains. For more details on the full *ParCorFull2* corpus we refer the reader to the original paper (Lapshinova-Koltunski *et al.*, 2022).

In order to come up with robust and effective CA-NMT models, we perform the two-step training mentioned in Section 2.1, where models are first pre-trained on large sentence-level corpora, and then refined on document-level data, which are in general less available. The multi-encoder model we use in this work is exactly the one proposed and trained for Lupo *et al.* (2022a). The concatenation model is the one described in Lupo *et al.* (2022b). Both multi-encoder and concatenation models were learned with 3 previous sentences as context. The multi-encoder model is pre-trained with the *divide-and-rule* strategy which makes it very effective on the contrastive test suites (Lupo *et al.*, 2022a).

3.3. *Experimental Design*

In order to perform the analyses and evaluations planned in this work, we performed the following processing steps on the ParCorFull2 English-German data and with the two CA-NMT models targeted in our analyses.

First of all we translated the ParCorFull2 data with the two CA-NMT models. The models were modified to generate also attention weights from the current sentence to the context sentences, for the source-side context only in the multi-encoder model, for both source and target side context in the concatenation model. For analyses presented in this work, we used attention weights obtained as the average of all attention heads. In the multi-encoder model we used the attention heads of the first level of the HAN module. In the concatenation model we used attention heads from the last layer of the encoder, for the source-side context, or decoder, for the target-side context.

The second step was to align the system's input and output sequences to sentences in the corpus. While alignment of input sequences should not be necessary, we found that sentences in the corpus were poorly tokenized. We thus provided raw sequences extracted from the corpus to the system and we re-performed a tokenization from scratch in order to guarantee a better match with the training data of the CA-NMT models. Alignment was performed with Levenshtein distance augmented with the *token-swap* operation. More details are given below.

Using alignments, we retrieved tokens in the system's input and output sequences belonging to coreferent mentions, with the corresponding attention weights. At this point, we were able to compute attention scores over coreference links between mentions in the system's current sentence and mentions in the system's context sentences. These scores were used to perform two analyses: i) a qualitative analysis performed manually over the subset of sentences introduced in Section 3.2; ii) a quantitative automatic analysis based on three metrics we designed on purpose for this work. This second analysis has been performed also on the target side of the concatenation model.

In order to compare the behavior of our two CA-NMT models through our analyses, and also to make attention weights more readable, we performed some post-processing on the attention matrices. From an explainability perspective, we would intuitively expect that a model which correctly exploits the context, when translating tokens involved in discourse phenomena, should put very high attention weights from these tokens to tokens instantiating their antecedent, and very low weights on the other tokens. In practice this behavior is rarely observed, but we keep it as a conceptual upper bound for the model's explainability evaluation. One of the worst behaviors from the same point of view would be when the model assigns the same weight to all tokens of a context sentence. In practice, model behaviors stay in between these two extreme cases.

We post-process attention weights as follows: i) we filter out attention weights smaller or equal to the value w_u for a given context sentence, $w_u = \frac{1.0}{N}$, where N is the context sentence length. This post-processing allows us to have cleaner attention matrices for manual inspection and leaves only weights potentially meaningful for analyzing the model's behavior; ii) we re-normalize attention weights with respect to the maximum weight in a given context sentence. This post-processing converts into 1.0 the maximum weight, allowing us to immediately spot the tokens where the model put the maximum attention. However it can generate more than one *1.0* weight in the same sentence. Additionally, it allows us to compare the multi-encoder to the concatenation model. Since the latter processes concatenated sentences and attention weights sum up to 1, its attention weights have in general smaller values. Renormalizing attention weights over context sentences separately allows us to bring back values to the same scale as the multi-encoder model.

Qualitative analysis. For our qualitative analysis, we identified the coreferent mentions in the current sentence and their corresponding antecedents in the context. Then we analyzed whether the attention from the former is indeed the highest toward the tokens in the context representing their antecedent. We focused on the potential mismatches and observed which tokens had the highest weights in this case. We also observed potentially ambiguous cases and commented on how the attention weights were distributed across the other tokens. This analysis was performed in the perspective of explainability of

machine translation, meaning that the objective was to understand if the disambiguation of the coreference was useful for the final translation, and in the perspective of CA-NMT evaluation with respect to disambiguating coreferences.

Quantitative analysis. While existing evaluation of CA-NMT based on test suites provides interesting insights on the ability of NMT models to use context, such an evaluation is only implicit, as models are only asked to score purposely chosen sentences in context, they are not used to generate translations for explicitly evaluating models. In order to provide a more direct and explicit evaluation of the ability of models in using context, we designed three evaluation metrics based on attention weights from tokens in the current sentence to tokens in the context sentences. The underlying hypothesis is that CA-NMT model's only way to access context is through the attention mechanism. Thus, the higher the attention, from tokens needing context to be disambiguated to the context, the more the model is correct in using the context, which is a much more direct way to assess the ability of models to use context.

All metrics exploit discourse phenomena annotated on the *ParCorFull2* corpus by aligning the corpus data to the system input and output sequences. Once the alignment is performed, tokens in the system's input and output sequences belonging to coreferent mentions can be spotted, and scores for these coreference links can be computed with attention weights from the mentions in the current sentence to their corresponding antecedents in the context sentences.

Data alignment is performed simply with an edit distance considering also the token swapping operation in addition to the traditional edit operations (insertion, deletion and substitution). We note that for input sequences edit distance is perfectly fine, as corpus and system sequences on the source side are basically the same, only a slight difference can be found due to system's tokenization. Indeed, computing the match rate of tokens belonging to mentions in the corpus and system's input sequences, we found that almost 96% of tokens match exactly. Tokens not matching differ indeed just because of the tokenization. Using the edit distance on the target side can be more problematic, as NMT models may generate target sequences matching perfectly the meaning of the gold target sentence, but using different tokens, e.g. synonyms. The mention tokens match rate was indeed around 55% on the target side. But analyzing a sample of corresponding target sequences, we found out that most of the time the meaning is preserved, that is the edit distance still align correctly, most of the time, mention tokens, even if the surface form is different, which is why the match rate is lower on target side.

We define the three evaluation metrics based on attention weight as follows:

1) *Max-weight* metric: is the percentage of coreference links for which the model gave the maximum attention weight compared to the attention weights to all tokens in the same context sentence. The intuition is that when a model has learned to exploit the context perfectly, it should give all the attention weight, that is 1.0, to the coreference link and ignore, that is attention weight 0.0, all the other tokens;

2) *Non-zero weight* metric: is the percentage of coreference links for which the model gave an attention weight greater than zero. We note that because of the post-processing

NMT model	*ContraPro Accuracy*
Baseline	45.00
(Zhang *et al.*, 2018)	42.60
(Tu *et al.*, 2018)*	45.20
(Müller *et al.*, 2018) concat21	48.00
(Müller *et al.*, 2018) concat22*	70.80
(Maruf *et al.*, 2019)*	39.15
(Voita *et al.*, 2018)	42.55
(Stojanovski and Fraser, 2019)	52.55
(Müller *et al.*, 2018)* best	58.13
Multi-encoder	61.09
Concat*	74.39

Table 3. *Quantitative results in terms of accuracy on the ContraPro test suite, obtained with the CA-NMT models. We show a comparison to a baseline context-agnostic model, and the best models from the literature. Models marked with * use both source and target context.*

performed on attention weights (see Section 3.3), the fact that a coreference link receives a non-zero weight is significant. This metric is much less restrictive than the *Max-weight* metric, the intuition for this is that the ideal situation where the model gives the total attention weight to the coreference link and zero to all the other tokens is too hard to reach. In practice, and basically because of the way attention mechanism is learned during the training phase, models spread attention to all tokens in a sentence;

3) *Average weight* metric: is the average attention weight the model gives to coreference links. This metric is computed by simply summing up the attention weights on all coreference links and dividing the sum by the number of coreference links.

We note that coreferent mentions may be composed of multiple tokens, and the attention mechanism of the model assigns a weight from each token in the current sentence to each token in the context sentence. In order to have only one attention weight for each coreference link, we chose to select the maximum weight. While this may give higher evaluation scores, given the difficulties in learning the attention mechanisms in NMT, mentioned in Section 3.3, we believe this choice does not change the overall picture.

4. Results

Beyond quantitative and qualitative evaluation of the CA-NMT models based on attention weight analyses and metrics, in order to show the effectiveness of the same models in terms of more traditional evaluation metrics compared to the literature, we also provide the accuracy on the English-German test suite ContraPro (Müller *et al.*, 2018).

En→De ContraPro (Müller *et al.*, 2018) is a large-scale test set from OpenSubtitles2018 (Lison *et al.*, 2018) that measures translation accuracy of the English anaphoric pronoun *it* into the corresponding German translations *er*, *sie* or *es*. Examples are balanced across the three pronoun classes (4,000 examples each). Each example requires identification of the pronominal antecedent, either in the source or target side, that can be found in the current sentence or any of the previous ones.

NMT model / Metric	*BLEU*	*COMET*	*ChrF*	*TER*
Multi-encoder	32.17	0.83	59.04	56.53
Concat*	32.08	0.81	58.62	57.38

Table 4. *Quantitative results in terms of BLEU[a], COMET[b], ChrF and TER scores, obtained with the multi-encoder and concatenation models on the ParCorFull2 corpus.* * *means the model use both source and target context.*

a) Using sacrebleu (Post, 2018), signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1.

b) Using model wmt22-comet-da: https://huggingface.co/Unbabel/wmt22-comet-da.

Quantitative results in terms of accuracy on the ContraPro test suite are provided in Table 3. We would like to underline some aspects concerning evaluation in Table 3: i) these results are provided with the only purpose of showing that we are using strong CA-NMT models for our analyses, and thus attention mechanisms on which we are basing our analyses have been properly learned; ii) accuracy on the ContraPro test suite is more predictive of the ability of the model to exploit context information than traditional metrics such like BLEU (Papineni *et al.*, 2002), but as we previously mentioned it only provides an implicit evaluation; iii) while systems from the literature showed in Table 3 were also evaluated in terms of BLEU, they were not evaluated on the same test set, or not with the same evaluation script, making BLEU results not comparable.

From results in Table 3 we can see that our concatenation model provides the best result in terms of accuracy on the ContraPro test suite. Our multi-encoder model reaches also a strong result, the only model from the literature providing a better accuracy on ContraPro being the concat22 in Müller *et al.* (2018) which also integrates the target side context. We attribute the strong performances of our two models on the ContraPro test suite to the larger amount of document-level data used for fine-tuning the models. Indeed we use a concatenation of News-Commentary-v12, Europarl-v7 and TED talks subtitles released by IWSLT17 (Cettolo *et al.*, 2012), accounting for $\sim$2.29M sentences. While the other models from the literature fine-tune CA-NMT models only on IWSLT17.

Additional results are displayed in Table 4. These results are computed on the 2,280 sentences from the English-German part of ParCorFull2. We can see that results in terms of BLEU, COMET, ChrF (Popović, 2015) and TER (Snover *et al.*, 2006) metrics are very similar for the two models, making their comparison through our analyses on these data more reliable.

Qualitative analysis

In this analysis, we examine how attention weights on context sentences, focusing in particular on coreference links, allow us to explain the result provided by the translation model. From an explainability point of view, we must distinguish two cases in our analysis: 1) cases where disambiguation of the antecedent is needed for a correct translation of a coreferent mention and 2) cases where there is no ambiguity, so the disambiguation of the antecedent is not needed. For example, the German pronoun “sie” for the third plural person doesn’t distinguish between genders. Therefore in order to translate “they” into German, the model doesn’t need to identify the correct antecedent in the context.

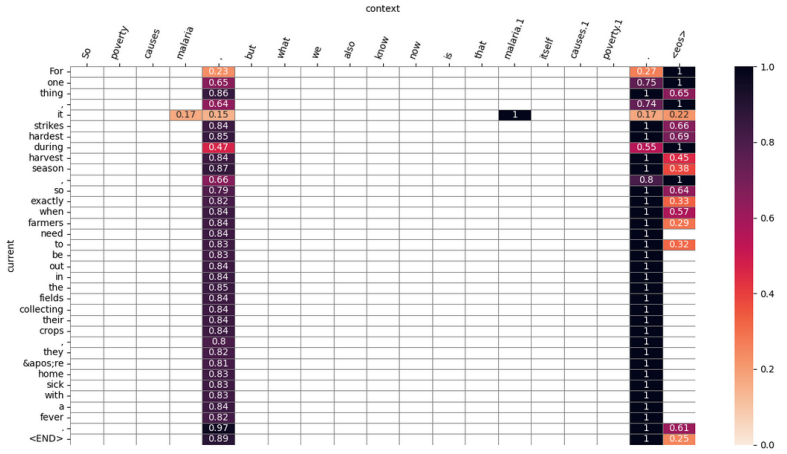

Figure 3. *An example of heatmap from the concatenation model, showing the register tokens problem. The model can still spot the coreference link between "it" and "malaria".*

For facilitating the understanding of heatmap images and discussions, we note that heatmaps must be read line by line, as tokens of the current sentence to be translated from the model are on the left-most column, while we specify in the discussion or in the caption of the image the distance of the context sentence from the current one in the document (1, 2 or 3). Additionally we recall that attention weights have been post-processed as described in Section 3.3.

From analysis of attention heatmaps displaying attention weights, not surprisingly attention is spread over more tokens than intuitively expected, that is attention is not concentrated on tokens belonging to coreferent mentions only. Both models suffer from giving high attention weights to function tokens (e.g. punctuation, articles, or the *end-of-sentence* symbol). This behavior has already been observed previously (Bibal *et al.*, 2022), and our interpretation is similar to the *register issue* described in Darcet *et al.* (2023). We give more details in Section 5. The multi-encoder model spreads attention more than the concatenation model, and increasingly more as the context sentence is at increasing distance, unless the context sentence contains antecedents for mentions of the current sentence. We can observe this behavior for example comparing Figures 1 and 2 which are attention heatmaps respectively from the multi-encoder and the concatenation model. They show attention for the same sentences, in particular from current sentence to a context sentence at distance 2. As we can see, while they both suffer from the *register issue*, and the multi-encoder model gives useless attention to some tokens, concerning the 3 mentions "it" coreferent with "homosexuality", the multi-encoder model is very precise in using the attention as the highest weight is always on the correct antecedent. The concatenation model instead does not spot any coreferent "it". Both models correctly translate each occurrence of "it", which is surprising for the concatenation model since we did not find any attention weight on a correct clue, either on the source or target side.

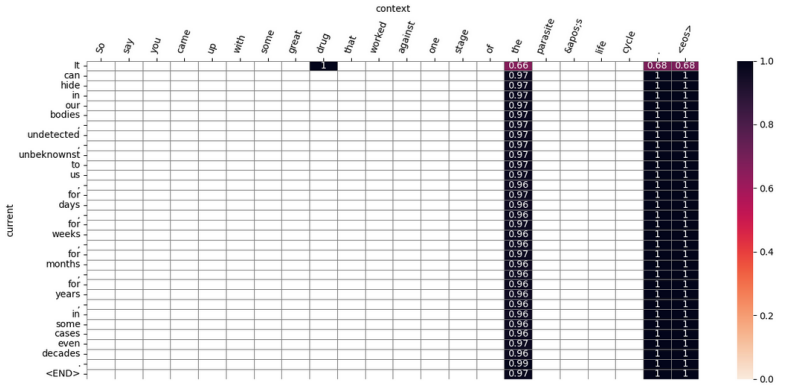

Figure 4. *An example of heatmap from the concatenation model, showing the register tokens problem. The model spots a wrong coreference link between "it" and "drug".*

In the case of the sentence shown in Figure 3 from the concatenation model, we can observe that for the token "it", the attention weights are the highest for the token "malaria" which is the correct antecedent of this pronoun. In German, "malaria" translates as "Malaria" (feminine noun) and the proposed translation of "it" is "sie" (feminine pronoun), so the translation is correct. The most *obvious* translation for the English "it" without any context would in fact be "es", we can therefore deduct that the disambiguation of the coreferent mention with the context helped for generating a correct translation.

As shown in the example in Figure 4 from the concatenation model, for the token "It" the highest attention weight is on the context token "drug", which is not the correct antecedent for this mention. The correct antecedent is the token "parasite" but it is not attended to by the model. Verifying the translation, we saw that the model translated "It" as "Es". This model has also access to the target-side context, therefore we can consider in our analysis the antecedent in the target language. The true antecedent is "Parasiten" (masculine noun) and the attended but incorrect token is "Medikament" (neuter noun). The generated pronoun "Es" is of neuter form, which doesn't agree with the correct antecedent but it agrees with the attended token "Medikament". We note that, in the perspective of purely evaluating the use of the context by a NMT model, this case should not be penalized like a full mistake, since the model translated the pronoun coherently with the translation of the token attended in the context.

Concatenation and multi-encoder models do not use attention mechanism in the same way. The concatenation model computes attention from the current sentence to all context sentences at the same time, making attention weights dependent one from each other. The multi-encoder model computes attention weights from the current sentence to each context sentence one at a time. As consequence, the multi-encoder model may make attention mistakes when context and current sentences contain coreferent mentions of different entities used in the same context. Examples in Figures 5 and 6 show this kind of issue. The current sentence contains the ambiguous mention "they", which can be disambiguated with both "women" in the context sentence at distance 1 (correct), and "men" in the context

sentence at distance 2. From a context usage point of view, this ambiguity will be difficult to resolve for the multi-encoder model, which processes context sentences separately. This is in general not an issue for the concatenation model which processes all context sentences at the same time. Examples 7 and 8 from the concatenation model show that although the token "men" still receives attention, the attention weight is much higher on "women", which is the correct antecedent. Both models generate a correct translation for this sentence, which is thus another case where an explicit evaluation of context usage would be more explainable than traditional metrics.

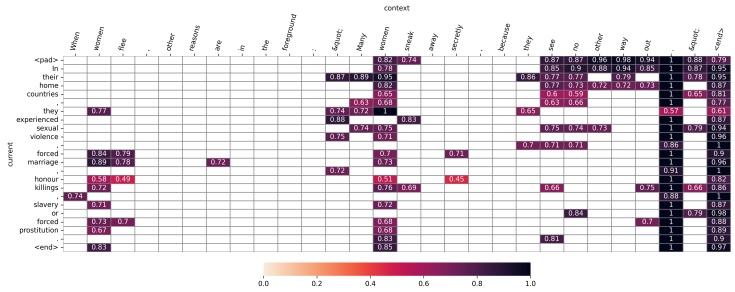

Figure 5. *Example of attention weights with the multi-encoder model between the current sentence and the context sentence at distance 1.*

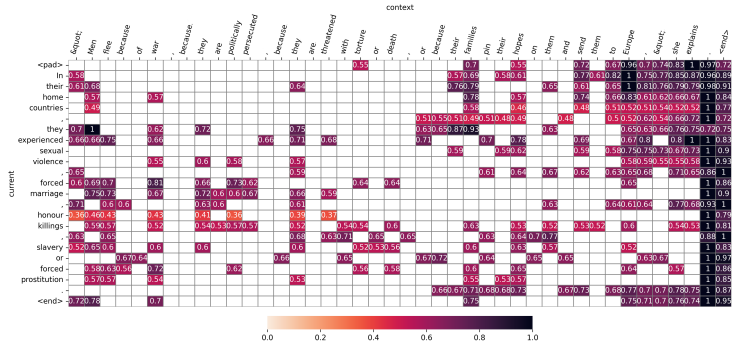

Figure 6. *Example of attention weights with the multi-encoder model between the current sentence and the context sentence at distance 2.*

The examples we analyzed are representative of what we observed in the subset of selected sentences for our human evaluation. Overall results of this evaluation are summarized in Table 5. We note that for the manual evaluation we considered all coreferent mentions we found in the selected sentences, not only the 135 coreferent mentions annotated in the ParCorFull 2.0 corpus. This is reflected in the total number of mentions (**# of mentions**) which is roughly 220.

We split results of Table 5 in 3 groups: the first reports results for all coreference cases (**All cases**); the second for coreference cases where the antecedent in the context is crucial

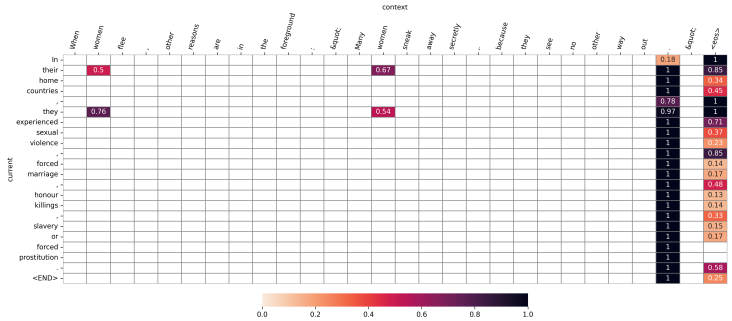

Figure 7. *Example of attention weights with the concatenation model between the current sentence and the context sentence at distance 1.*

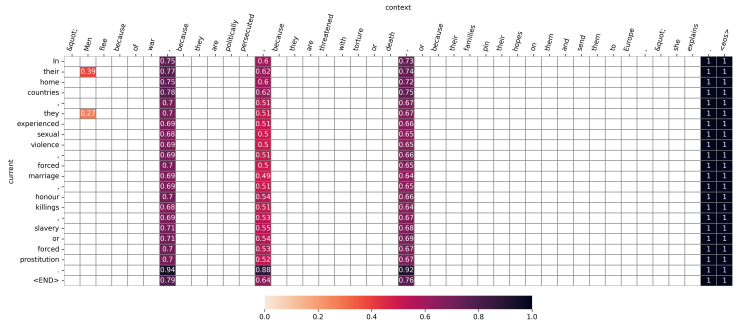

Figure 8. *Example of attention weights with the concatenation model between the current sentence and the context sentence at distance 2.*

for disambiguating the mention (**Ctx needed**), and thus presumably also for the translation; the third group reports results for cases where the context is needed for disambiguating the mention and the coreference case is considered as *hard*. We consider a coreference case as hard if some or all words in the mention are different from words in its antecedent, excluding function words (and thus pronouns). The additional group **Positive attention** in the table, refers to all cases where the model puts significant attention weights on tokens in the context sentence, regardless if they are in correct antecedents or not. We note that even in the latter case positive attention was most of the time *justified*, e.g. the correct antecedent can be ambiguous. This group helps us to understand the precision of the model.

Over table lines, together with results corresponding to the same three evaluation metrics introduced in Section 3.3 for the quantitative evaluation, that will be discussed in the next section, we report also if the model actually found just a naive coreference link (**Naive links**). This could mean that either the correct coreference link annotated in the corpus is naive, e.g. both mention and its antecedent are the same mention ("it" → "it"); or mention and its antecedent contain some common tokens, and the model put attention weights only on common

tokens, for example "the pink one" → "a pink ballon" and the model put a significant attention weight only on "pink". In the last line of the Table 5 (**Dispersion**), we give the average number of tokens in the context sentence, excluding function tokens used as *registers* (see Section 4, paragraph **Qualitative analysis**), attended by the model from each token in the current sentence with a significant attention weight. This value summarizes in a number what can be visually observed in the heatmaps: in some cases, the model spreads attention weights over a relatively high number of tokens, while in other cases it does not pay much attention, except for function words, while there is still a correct coreference link that should be spot. [1]

We summarize results in Table 5 as follows:

– 82.4% of times the multi-encoder model puts a significant attention weight on the correct antecedent, versus 41.5% of times for the concatenation model (**All cases** group). In such cases the concatenation model puts the maximum attention value more often (91.3% of times) than the multi-enc model (67.9% of times). However, on average the attention to the correct antecedent is larger for the multi-encoder model (0.886) than for the concatenation model (0.467);

– 11.2% of times the context is needed for disambiguating a coreference, but the multi-enc model puts insignificant attention weight (below the uniform distribution) on the correct antecedent (25 of 224 mentions). When the context is needed for disambiguation, the multi-enc model shows a small improvement in the Non-Zero-weight metric (84% versus 82%), showing that the model puts more significant attention when the contextual information could be necessary to generate a correct translation. In the case of the concatenation model, 41.1% of mentions need disambiguation and are not significantly attended (92 of 224 mentions);

– 53% of times the coreference is considered as hard (ctx needed & hard coref column). In this cases, both models present a drop in their performance, the multi-enc model attended with a maximum value in only 50% of cases, and the concatenation model attended to a mention in 37% of hard cases;

– 84% of the significantly attended antecedents are a coreference in the multi-encoder model and 83.1% in the concatenation model. The precision of the concatenation model is the highest one if we consider only the Max-weight value achieving 78.8% of correct coreferences resolved with the maximum value.

The human evaluation, together with observations we made in Section 2, motivate our quantitative analyses with the three metrics based on attention weights. The aim is to find a metric which better explains the behavior of models we observed over heatmaps.

Quantitative analysis

Results obtained with the three metrics based on attention weights over coreference links are shown in Table 6 for the whole English-German data of ParCorFull2, while in Table 7 we show results with the same metrics on the sentences selected for the human evaluation. For the concatenation model we show evaluation scores for both source-side (src) and target-side (tgt) context. Results in the two tables follow the same trend, and they have also similar trend

1. In all sentences of the ParCorFull 2.0 corpus, there is at least one annotated coreference case.

	All cases		Ctx needed		Ctx needed & hard coref		Positive attention	
	Model							
	multi-enc	concat	multi-enc	concat	multi-enc	concat	multi-enc	concat
# of mentions	224	224	160	160	116	119	215	118
Naive links (%)	13.1%	14.3%	3.6%	4.3%	2.6%	5.8%	13.5%	27.1%
Max-weight	58.1%	41.5%	55.2%	39.7%	50%	34.5%	60%	78.8%
Non-zero weight	82.4%	43.8%	84%	42.2%	80.1%	37%	84.1%	83.1%
Average weight	0.887	0.467	0.894	0.476	0.887	0.50	0.886	0.467
Dispersion	6.43	3.42	6.63	1.11	7.49	1.06	7.53	1.02

Table 5. *Human (manual) evaluation statistics on the 73 selected examples for the multi -encoder (multi-enc) and concatenation (concat) models.*

NMT model / Metric	Max-weight	Non-zero weight	Average weight
Multi-encoder (src)	45.91%	88.83%	0.8183
Concat (src)	10.45%	50.98%	0.2994
Concat (tgt)	13.25%	33.22%	0.2136

Table 6. *Quantitative results with three different evaluation metrics (see the text), over discourse phenomena in the* ParCorFull2 *corpus, based on attention weights of CA-NMT models.*

NMT model / Metric	Max-weight	Non-zero weight	Average weight
Multi-encoder (src)	49.31%	92.36%	0.8574
Concat (src)	9.49%	51.82%	0.3039
Concat (tgt)	10.71%	29.46%	0.2117

Table 7. *Quantitative results with three different evaluation metrics (see the text), over discourse phenomena in the selected subset of 73 examples, based on attention weights of CA-NMT models.*

as the same metrics computed in the manual evaluation, shown in Table 5. These agreements among different tables make the scores more reliable, but also prove to some extent the correcteness of our automatic evaluation methodology based on alignments. As we can see, these metrics provide an evaluation much more in favour of the multi-encoder model, in contrast to traditional and official evaluation metrics as shown in Table 4, including the evaluation based on the ContraPro contrastive test suite in Table 3. This is not surprising for the *Average-weight* metric, since on the analyzed subset of sentences we observed higher weights on coreference links for the multi-encoder model. The other two metrics confirm quantitatively on the whole data set what we observed on the subset, in particular the *Max-weight* metric which is the most restrictive (the model must put the maximum attention weight of the analyzed context sentence on the coreference link). While computing these metrics demands the availability of an expensive resource like the ParCorFull2 corpus, they provide a more explicit and intuitive evaluation of the behavior of models in using the context.

5. Discussion

Drawing an analogy with explainability for vision recognition, it seems that some function words are assigned attention weights that do not seem to convey specific information *per se* but seem to play a role in how the information flux is organized. In Darcet *et al.* (2023) authors suggest that, for vision transformers, some pixels are used to store attention weight information. For a picture task identification, these pixels are meaningless but seem to be used to store information like a buffer, what they call "registers". It may be the case that the special tokens (e.g. ',', '.', and '<eos>' in Figure 8) are potentially similarly used as registers, as heatmaps in our work and in Darcet *et al.* (2023) present similar patterns, and our models are also based on *Transformer*.

6. Conclusions

In this paper we proposed a human evaluation of heatmaps generated with attention weights from current to context sentences for CA-NMT models. We analyzed two different models, belonging to the two main approaches for CA-NMT: multi-encoder and concatenation. Despite some reasonable divergence from what can be intuitively expected from the attention behavior in the targeted context, at least on discourse phenomena like coreferences, in the limit of data used for the analyses, attention weights exhibit a sufficient interpretability from a *plausibility* perspective that let us adhere to the party of *attention is explanation* in the debate raised by Bibal *et al.* (2022). The human evaluation is completed by a quantitative evaluation based on attention weights over coreference links, and with three evaluation metrics. The results obtained with this evaluation confirm those observed with the human evaluation, and let us believe that the proposed metrics may constitute a more explicit and direct evaluation of the ability of CA-NMT models to use context when facing coreference phenomena.

As a limitation of our work, we note that focusing on coreference analysis in the case of CA-NMT is a particularly favourable case and some other linguistic phenomena may not be that easily captured with attention weights and, conversely, it may well be that some higher attention weights may be assigned without such an obvious linguistic correlation, and therefore any explanatory power. Additionally we performed our manual analyses on the source side only, while the concatenation model uses also the target context, which may alter the way the model needs to attend to the source context. In the same line of thoughts, also the cross-attention mechanism, which we did not consider at all in this work, may alter to some extent the behavior of the other mechanisms. We leave deeper and more comprehensive analyses on these points for future work.

We note that annotating two system outputs is time-consuming, but in future work we may use our annotations to perform a logistic regression with the attention weight as the predictors and the accuracy of the identification of the referent as the predicted variable.

Acknowledgements

Work supported by: the MAKE-NMTVIZ project, funded under the 2022 Grenoble-Swansea Centre for AI Call for Proposals/ GoSCAI Grenoble-Swansea Joint Centre in Human Centred AI and Data Systems (MIAI@Grenoble Alpes (ANR-19-P3IA-0003)); the CREMA project (Coreference REsolution into MAchine translation) funded by the French National Research Agency (ANR), contract number ANR-21-CE23-0021-01.

7. References

Agrawal R. R., Turchi M., Negri M., "Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides", *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, p. 11-20, 2018.

Ali S., Abuhmed T., El-Sappagh S., Muhammad K., Alonso-Moral J. M., Confalonieri R., Guidotti R., Del Ser J., Díaz-Rodríguez N., Herrera F., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence", *Information Fusion*, vol. 99, p. 101805, 2023.

Alvarez-Melis D., Jaakkola T. S., "A causal framework for explaining the predictions of black-box sequence-to-sequence models", *arXiv preprint arXiv:1707.01943*, 2017.

Bahdanau D., Cho K., Bengio Y., "Neural Machine Translation by Jointly Learning to Align and Translate", *arXiv e-prints*, vol. 1409, p. arXiv:1409.0473, September, 2014.

Bawden R., Sennrich R., Birch A., Haddow B., "Evaluating Discourse Phenomena in Neural Machine Translation", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, p. 1304-1313, June, 2018.

Bibal A., Cardon R., Alfter D., Wilkens R., Wang X., François T., Watrin P., "Is attention explanation? an introduction to the debate", *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 3889-3900, 2022.

Cettolo M., Girardi C., Federico M., "WIT3: Web Inventory of Transcribed and Translated Talks", *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, European Association for Machine Translation, Trento, Italy, p. 261-268, May 28–30, 2012.

Clark K., Khandelwal U., Levy O., Manning C. D., "What does bert look at? an analysis of bert's attention", *arXiv preprint arXiv:1906.04341*, 2019.

Darcet T., Oquab M., Mairal J., Bojanowski P., "Vision Transformers Need Registers", *arXiv preprint arXiv:2309.16588*, 2023.

de Seyssel M., Lavechin M., Adi Y., Dupoux E., Wisniewski G., "Probing phoneme, language and speaker information in unsupervised speech representations", *Proc. Interspeech 2022*, p. 1402-1406, 2022.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171-4186, June, 2019.

Ding S., Xu H., Koehn P., "Saliency-driven Word Alignment Interpretation for Neural Machine Translation", *in* O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, K. Verspoor (eds), *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Association for Computational Linguistics, Florence, Italy, p. 1-12, August, 2019.

Ghader H., Monz C., "What does attention in neural machine translation pay attention to?", *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. , p. 30-39, 2017.

He S., Tu Z., Wang X., Wang L., Lyu M., Shi S., "Towards Understanding Neural Machine Translation with Word Importance", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 953-962, 2019.

Isabelle P., Cherry C., Foster G., "A Challenge Set Approach to Evaluating Machine Translation", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2486-2496, 2017.

Jain S., Wallace B. C., "Attention is not Explanation", *in* J. Burstein, C. Doran, T. Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 3543-3556, June, 2019.

Jaziriyan M. M., Ghaderi F., "Automatic Post-editing of Hierarchical Attention Networks for Improved Context-aware Neural Machine Translation", *Journal of AI and Data Mining*, vol. 11, nº 1, p. 95-102, 2023.

Junczys-Dowmunt M., "Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation", *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Association for Computational Linguistics, Florence, Italy, p. 225-233, August, 2019.

Kim Y., Tran D. T., Ney H., "When and Why is Document-level Context Useful in Neural Machine Translation?", *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, Association for Computational Linguistics, Hong Kong, China, p. 24-34, November, 2019.

Kuang S., Xiong D., Luo W., Zhou G., "Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches", *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 596-606, August, 2018.

Lapshinova-Koltunski E., Ferreira P. A., Lartaud E., Hardmeier C., "ParCorFull2.0: a Parallel Corpus Annotated with Full Coreference", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 805-813, June, 2022.

Lee K., He L., Lewis M., Zettlemoyer L., "End-to-end Neural Coreference Resolution", *in* M. Palmer, R. Hwa, S. Riedel (eds), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, p. 188-197, September, 2017.

Li B., Liu H., Wang Z., Jiang Y., Xiao T., Zhu J., Liu T., Li C., "Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation", *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, Association for Computational Linguistics, Online, p. 3512-3518, July, 2020.

Lison P., Tiedemann J., Kouylekov M., "OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, May, 2018.

Lopes A., Farajian M. A., Bawden R., Zhang M., Martins A. T., "Document-level Neural MT: A Systematic Comparison", *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisbon, Portugal, p. 225–234, 2020.

Luong T., Le Q. V., Sutskever I., Vinyals O., Kaiser L., "Multi-task Sequence to Sequence Learning", *International Conference on Learning Representations*, 2016.

Lupo L., Dinarelli M., Besacier L., "Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models", *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, p. 4557-4572, May, 2022a.

Lupo L., Dinarelli M., Besacier L., "Focused Concatenation for Context-Aware Neural Machine Translation", *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), p. 830-842, December, 2022b.

Ma S., Zhang D., Zhou M., "A Simple and Effective Unified Encoder for Document-Level Machine Translation", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 3505-3511, July, 2020.

Maruf S., Haffari G., "Document Context Neural Machine Translation with Memory Networks", *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, p. 1275-1284, July, 2018.

Maruf S., Martins A. F. T., Haffari G., "Contextual Neural Model for Translating Bilingual Multi-Speaker Conversations", *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Brussels, Belgium, p. 101-112, October, 2018.

Maruf S., Martins A. F. T., Haffari G., "Selective Attention for Context-aware Neural Machine Translation", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 3092-3102, June, 2019.

Miculicich L., Ram D., Pappas N., Henderson J., "Document-Level Neural Machine Translation with Hierarchical Attention Networks", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, p. 2947-2954, October-November, 2018.

Moradi P., Kambhatla N., Sarkar A., "Measuring and Improving Faithfulness of Attention in Neural Machine Translation", *in* P. Merlo, J. Tiedemann, R. Tsarfaty (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, Online, p. 2791-2802, April, 2021.

Müller M., Rios A., Voita E., Sennrich R., "A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation", *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Brussels, Belgium, p. 61-72, October, 2018.

Papineni K., Roukos S., Ward T., Zhu W.-J., "Bleu: a Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 311-318, July, 2002.

Pasad A., Chou J.-C., Livescu K., "Layer-Wise Analysis of a Self-Supervised Speech Representation Model", *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 914-921, 2021.

Paul B., "Advancements and Perspectives in Machine Translation: A Comprehensive Review", *1st-International Conference on Recent Innovations in Computing, Science & Technology*, 2023.

Phillips P. J., Hahn A. C., Fontana P. C., Broniatowski D. A., Przybocki M. A., "Four principles of explainable artificial intelligence", 2021.

Popović M., "chrF: character n-gram F-score for automatic MT evaluation", *in* O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (eds), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Lisbon, Portugal, p. 392-395, September, 2015.

Post M., "A Call for Clarity in Reporting BLEU Scores", *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Belgium, Brussels, p. 186-191, October, 2018.

Raganato A., Tiedemann J., "An analysis of encoder representations in transformer-based machine translation", *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, The Association for Computational Linguistics, p. 287-297, 2018.

Rei R., Stewart C., Farinha A. C., Lavie A., "COMET: A Neural Framework for MT Evaluation", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2685-2702, 2020.

Serrano S., Smith N. A., "Is Attention Interpretable?", *in* A. Korhonen, D. Traum, L. Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 2931-2951, July, 2019.

Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J., "A Study of Translation Edit Rate with Targeted Human Annotation", *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, p. 223-231, August 8-12, 2006.

Stojanovski D., Fraser A., "Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning", *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, European Association for Machine Translation, Dublin, Ireland, p. 140-150, August, 2019.

Sundararajan M., Taly A., Yan Q., "Axiomatic attribution for deep networks", *International conference on machine learning*, PMLR, p. 3319-3328, 2017.

Tan X., Zhang L., Xiong D., Zhou G., "Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 1576-1585, November, 2019.

Tay Y., Dehghani M., Bahri D., Metzler D., "Efficient Transformers: A Survey", *CoRR*, 2020.

Tiedemann J., Scherrer Y., "Neural Machine Translation with Extended Context", *in* B. Webber, A. Popescu-Belis, J. Tiedemann (eds), *Proceedings of the Third Workshop on Discourse in*

Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, p. 82-92, September, 2017a.

Tiedemann J., Scherrer Y., "Neural Machine Translation with Extended Context", *Proceedings of the Third Workshop on Discourse in Machine Translation*, Association for Computational Linguistics, Copenhagen, Denmark, p. 82-92, September, 2017b.

Tu Z., Liu Y., Shi S., Zhang T., "Learning to Remember Translation History with a Continuous Cache", *Transactions of the Association for Computational Linguistics*, vol. 6, p. 407-420, 2018.

Vashishth S., Upadhyay S., Tomar G. S., Faruqui M., "Attention Interpretability Across NLP Tasks", *arXiv preprint*, arXiv, 2019.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., "Attention is all you need", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Curran Associates Inc., Long Beach, California, USA, p. 6000-6010, December, 2017.

Vig J., Belinkov Y., "Analyzing the structure of attention in a transformer language model", *arXiv preprint arXiv:1906.04284*, 2019.

Voita E., Sennrich R., Titov I., "Context-Aware Monolingual Repair for Neural Machine Translation", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 877-886, November, 2019a.

Voita E., Sennrich R., Titov I., "When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 1198-1212, July, 2019b.

Voita E., Serdyukov P., Sennrich R., Titov I., "Context-Aware Neural Machine Translation Learns Anaphora Resolution", *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, p. 1264-1274, July, 2018.

Wang S., Li B. Z., Khabsa M., Fang H., Ma H., "Linformer: Self-Attention with Linear Complexity", *arXiv:2006.04768 [cs, stat]*, June, 2020. 00013 arXiv: 2006.04768.

Wiegreffe S., Pinter Y., "Attention is not not explanation", *arXiv preprint arXiv:1908.04626*, 2019.

Yin K., Fernandes P., Pruthi D., Chaudhary A., Martins A. F., Neubig G., "Do context-aware translation models pay the right attention?", *arXiv preprint arXiv:2105.06977*, 2021.

Zhang J., Luan H., Sun M., Zhai F., Xu J., Zhang M., Liu Y., "Improving the Transformer Translation Model with Document-Level Context", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, p. 533-542, October-November, 2018.

Zhang P., Chen B., Ge N., Fan K., "Long-Short Term Masking Transformer: A Simple but Effective Baseline for Document-level Neural Machine Translation", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, p. 1081-1087, November, 2020.

Zheng Z., Yue X., Huang S., Chen J., Birch A., "Towards Making the Most of Context in Neural Machine Translation", *in* C. Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, International Joint Conferences on Artificial Intelligence Organization, p. 3983-3989, February, 2020.