Exploring NMT Explainability for Translators Using NMT Visualising Tools

Gabriela Gonzalez-Saez, Mariam Nakhlé, James Robert Turner, Fabien Lopez, Nicolas Ballier, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Raheel Qader, Caroline Rossi, et al.

Exploring NMT Explainability for Translators Using NMT Visualising Tools

Gabriela Gonzalez-Saez[1], **Mariam Nakhlé**[1 5], **James Robert Turner**[4], **Fabien Lopez**[1], **Nicolas Ballier**[3], **Marco Dinarelli**[1], **Emmanuelle Esperança-Rodier**[1], **Sui He**[4], **Raheel Qader**[5], **Caroline Rossi**[2], **Didier Schwab**[1], **Jun Yang**[4]
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG 38000 Grenoble, France; [2]Université Grenoble Alpes; [3]Université Paris Cité, LLF & CLILLAC-ARP, 75013 Paris, France; [4]Swansea University; [5]Lingua Custodia, France
gabriela-nicole.gonzalez-saez@univ-grenoble-alpes.fr

Abstract

This paper describes work in progress on Visualisation tools to foster collaborations between translators and computational scientists. We aim to describe how visualisation features can be used to explain translation and NMT outputs. We tested several visualisation functionalities with three NMT models based on Chinese-English, Spanish-English and French-English language pairs. We created three demos containing different visualisation tools and analysed them within the framework of performance-explainability, focusing on the translator's perspective.

1 Introduction

The development of machine translation (MT) is influenced by a wide range of actors and agents, ranging from the investors to general public. A stakeholder approach to MT enables us to examine the effects of MT on each of the different interest groups, with particular reference to levels of involvement with MT (e.g., translators, students and trainees, end users, MT investors and developers, translation agencies, and academic researchers) (Guerberof-Arenas and Moorkens, 2023).

Upon refining the landscape of MT to include the directly associated stakeholders, several distinct categories emerge. The primary category consists of MT developers, typically computer scientists, whose focus lies in enhancing the accuracy and fluency of translations. In contrast, a second group of stakeholders, comprised of linguists and translators with expertise in translation studies, may argue that translation quality cannot be regarded as a static or absolute concept; instead, it is influenced by both subjective and objective factors that may change over time, as evidenced by semantic/communicative and functional translation approaches. Their concern centres on how MT fits into their practical translation workflow, and MT's ability to handle cultural-specific items and nuances – a realm in which translators take great pride. Moreover, industry surveys, such as the Freelance Translator Survey 2023 by Inbox Translation, and CIOL Insights 2022,[1] have also demonstrated that translators are primarily concerned about the effects of MT in their professional status, i.e., decreased translation/post-editing rates, clients' unrealistic expectations and other people's perception of their professionalism. Moreover, end-users constitute another critical group, prioritising usability, speed, and the cost-effectiveness of translations (Vieira et al., 2023).

Although these three groups of stakeholders are closely related to the development of MT, they do not always understand each other's work or demands, underscoring the need for continuous dialogue between each group. This work is of part of the MAKE-NMTViz project, which, by bringing together key stakeholders in MT development, aims to connect MT researchers with professional translators, taking into consideration their needs and preferences, whilst improving translators' MT literacy. This project is a starting point for facilitating communication ensuring that MT is developed and utilised effectively.

Central to our investigation is the role of vi-

[1]https://inboxtranslation.com/resources/research/freelance-translator-survey-2023/ and https://www.ciol.org.uk/ciol-insights-languages-professions-2

sualisation systems in Explainable Artificial Intelligence (XAI) for Neural Machine Translation (NMT). We aim to assess the utility of various NMT visualisation tools for professional translators, examining how these tools contribute to their understanding of NMT models' decisions. Through a comprehensive review of existing explainability visualisation systems, we implement selected ones in the form of three demos available on HuggingFace Spaces, in order to determine their effectiveness in helping translators comprehend whether MT models produce accurate translations for appropriate reasons. By facilitating communication and understanding among key stakeholders, our project aims to promote the effective development and use of NMT systems.

The contributions of this paper are the following: (i) a revision and typology of state-of-the-art visualisation functionalities for the explainability for NMT; (ii) a set of ready-to-use explainability and visualisation tools available in the Hugging Face Spaces for the translator's use;[2] and, (iii) a translator-focused evaluation of explainability visualisations for NMT. The paper is structured as follows: Section 2 provides an overview of previous research on NMT explainability methods and existing visualisation systems. In Section 3, a typology of functionalities is examined. Section 4 details the methodology for assessing explainability functionalities for translators. Section 5 presents an analysis of the visualisation systems from the translator's perspective. Discussion and conclusions are presented in Sections 6 and 7.

2 Previous Research on Visualisation Tools for NMT

In this section, we present the state-of-the-art of visualisation methods and tools employed for Explainable NMT (XNMT). Visualisation is a key component of XNMT methods identified by Stahlberg (2020) in his survey paper. It is used in Model-intrinsic interpretability methods, post-hoc interpretability methods (interpreting predictions with input analysis), and the analysis of Confidence Estimation in Translation. In a more recent survey by Madsen et al. (2022), several post-hoc methods for NLP interpretability were reviewed, which has been further specialised by Leiter et al. (2023), who specifically reviewed methods for NMT metrics. In this work, we focus on the review of existing systems that implement such methods, aiming to make XNMT accessible for translators.

[2]https://huggingface.co/gabrielanicole

2.1 XNMT Methods

We present Explainability Methods specifically for NMT implemented using a Transformer architecture (Vaswani et al., 2017). The translation process starts with an **input** sequence of words in the source language that undergoes different steps as defined in the **process**, and concludes with the generation of a **output** sequence of words in the target language (this process is detailed in Figure 1). We categorise XNMT methods into two types: *Inspection methods* and *Attribution methods*. Each type considers different aspects of the translation process to provide explanations to the final user.

2.1.1 Inspection methods

Inspection methods present a single-point decision made by the NMT system. The challenge lies in selecting valuable information directly from a model decision or parameter.

We inspect the NMT model in several parts of the process. On the input side, we consider the presentation of the tokenised input sequence that is being fed to the NMT system. On the output side, the presentation of the NMT probability of every generated token, and the visualisation of the decoding algorithm, such as the beam search sequence generation. These inspection methods are used as part of debugging techniques and provide transparency to the NLP pipeline (Alharbi et al., 2021). Inspection Methods can also be extended using manipulation procedures. In this case, the raw values of the NMT system are post-processed to be more easily interpretable. An example is the use of weights computed by the attention mechanisms to describe how the NMT system relates the source and output sentences (Wiegreffe and Pinter, 2019). The attention values have also been used to compute a Confidence estimation metric, as presented by Rikters and Fishel (2017).

2.1.2 Attribution methods

Attribution methods aim to elucidate the relationship between different parts of the translation process and the impact that one part has on another. These methods are often referred to as feature importance algorithms, as they model one part (e.g., the input tokens) as a set of features responsible for the generated output (Zhou et al., 2022).

There are different levels of attributions

(Kokhlikyan et al., 2020). *Primary* attribution focuses on the relationship between the input features and the corresponding generated outputs of the model (e.g. (Sundararajan et al., 2017), (Ding et al., 2019)). It uses the gradients (i.e., internal data) of the NMT with respect to the input, helping to visualise the impact of the input tokens on the output tokens. *Layer* Attribution variant extends attribution to all neurons in a hidden layer, and *Neuron* Attribution methods attribute specific internal, hidden neurons to the inputs or output of the model. For instance, Bau et al. (2018) presented a method to detect the neuron responsible for a particular linguistic property, and manipulating that neuron would alter the linguistic property in the output. Detecting the relationship between a specific part of the NMT model and a linguistic behaviour is also known as a *probing method*. For example, in Linguistic correlation Analysis (Dalvi et al., 2019a), a supervised method learns the most relevant neurons for an extrinsic task as Part-of-Speech classification. This approach helps uncover the linguistic properties encoded within the NMT model's internal representations.

Inspection and Attribution methods can both be categorised as Model-Intrinsic or Post-Hoc methods, depending on whether they utilise the model's internal data or only the inputs and outputs of the model. While Inspection Methods aim to explain a single decision made by the model, Attribution Methods are more complex as they analyse the interaction between different parts of the model that may not directly interact. Together, these methods provide both decision and model understanding of the NMT outputs.

2.2 XNMT Systems and Tools

While survey papers on explainability in AI encompass many systems from Computer Vision to Text Generation, existing reports on XAI (Phillips et al., 2021) or on Visual Analytics (e.g. (Cui, 2019)) do not focus on the task and processes of translation *per se*. Though acknowledging two main types of visualisation techniques for texts, Bodria et al.'s (2021) all-encompassing survey paper fails to capture all the investigation techniques based on visualisation that have been developed for NMT. Even if some NMT toolkits like THUMT (Tan et al., 2020) or JoeyNMT (Kreutzer et al., 2019a) propose cross-lingual attention as a standard functionality, visualisation is hardly exploited to the best of its potential for XNMT. Instead, the focus is not exclusively on visualisations but rather on probing strategies (de Seyssel et al., 2022).

In the following, we review existing visualisation systems and subsequently propose a typology of implemented functionalities within these systems. This recap encompasses methods, toolboxes, or libraries used for visualising NMT.

2.2.1 Main Visualisation tools

Various tools are available that implement functionalities to explain the outputs and internals of NMT. These tools are available in the form of libraries and systems. Our analysis primarily focuses on visualisations designed for the `Transformer` architecture, but we also consider related tools that focus on sequence-to-sequence tasks (e.g. seq2seq-viz (Strobelt et al., 2018)). The described tools offer a comprehensive overview of various functionalities that would enhance our understanding of translation as a task.

We review tools based on one of the following NMT toolkits: Fairseq (Ott et al., 2019), OpenNMT (Klein et al., 2017), JoeyNMT (Kreutzer et al., 2019b), and HuggingFace Transformers (HF-Tranformers) (Wolf et al., 2020). The toolkit is the base of the XNMT tool, as the visualisation features are developed using the internals and outputs provided by the toolkit. We list XNMT method implementations detailed in Section 2.1. Table 1 summarizes eight analysed libraries and systems, detailing their creation year, NMT toolkit and if it is designed for the Transformer architecture (TR).

System	Year	NMT toolkit	TR
Seq2Seq-Vis	2018	OpenNMT	no
BertVis	2018	HF-Transformers	yes
Neurox	2019	HF-Transformers	yes
LIT	2020	HF-Transformers	yes
Captum	2020	HF-Transformers	yes
NMTViz	2021	py-torch	yes
Ecco	2021	HF-Transformers	yes
InSeq	2023	HF-Transformers	yes

Table 1: Libraries and systems overview. TR: Transformer.

2.2.2 Libraries

BertVis (Vig, 2019) is an inspection tool, which focuses on the NMT process by visualising the internals of the models, more specifically it provides detailed information of each multilayer and

multi-head attention of a neural model, supporting encoder-decoder architectures. It is specific for NLP models and works directly in Jupyter Notebook. *NeuroX* (Dalvi et al., 2019b) implements probing methods at a neuron and layer level. Additionally, it facilitates the manipulation of neuron values to explore architecture alternatives at tokenisation and neuronal levels. NeuroX also supports quality evaluation and analysis through Ablation studies, allowing users to analyse the impact of modifications on the generation of translated text, as demonstrated in previous research (Bau et al., 2018). *Captum* (Kokhlikyan et al., 2020) is a Python library designed for PyTorch models, offering access to model internals and computation of primary, neuron, and layer attribution methods. *Ecco* (Alammar, 2021) is an interactive inspection and attribution tool that operates within Jupyter Notebook. It supports a selection of models such as GPT-2, BERT, and RoBERTa. Similar to *Ecco*, *Inseq* (Sarti et al., 2023) is also based on *Captum* and provides comparable functionalities. However, *Inseq* extends its support to a wider range of models and is adaptable to various systems beyond Jupyter notebooks.

2.2.3 Systems

The following tools, presented in the form of systems, are standalone platforms tailored to facilitate explainability in NMT tasks.

Seq2Seq-Vis (Strobelt et al., 2018) is a system focused on aiding neural model developers in error detection through a set of inspection functionalities. It includes three main functionalities: (i) Inspection, presenting embedding space visualisation based on similar tokens from the training data for the encoder and decoder, attention visualisation between them, and probabilities for the generation of each token and the beam search; (ii) *What-if* Translations, allowing modification of selected tokens using the most probable ones, beam search, or attention between source and target tokens; and, (iii) Human error search for debugging, utilising the previous functionalities and relying on the NMT model's understanding to identify bugs in the analysed architecture. It works on OpenNMT encoder-decoder architectures before the transformer era.

The *Language Interpretability Tool (LIT)* (Tenney et al., 2020) implements various tasks for explaining datasets, embeddings, and token representations. It utilises different primary attribution methods for analysing model behaviour. Additionally, LIT visualises attention matrices, compares different data points and models, and provides performance evaluation. This versatile tool is compatible with various NMT toolkits, such as HF-transformer. *NMTVis* (Munz et al., 2021) Is the only tool that targets the translator user, offering functionality for exploring various translation alternatives using the generated target text and the beam search, along with the visualiastion of attention between source and target sentences. The user can manually modify a translation, which updates the remainder of the target sentence, and navigate across different generation options to refine translations using the beam search.

While libraries are easier to incorporate into a new tool, systems are closed platforms that pose challenges when integrating with different models or third-party systems.

3 A Typology of Implemented Functionalities

In this section, we present a typology of implemented functionalities in state-of-the-art systems. To exemplify each functionality, we map them to the Transformer architecture (Figure 1, original figure taken from Vaswani et al. (2017)). Our survey of XNMT as a task and process follows an input, process, output analysis of the functionalities:

(i) Tokenisation (input/output) visualisation illustrates how input sequences are divided into tokens, which are then represented as embeddings in the encoder. Similarly, decoder output is presented in terms of tokens. Various XNMT tools display this functionality, like *BertViz* by showing attention links between tokens rather than words.

(ii) Embeddings (input/process) representation is depicted as a 2D or 3D projection through dimension reduction techniques, such as UMAP or t-SNE. As it is a space of points, multiple samples are used to relate different tokens. For instance, *LIT* illustrates the embedding space using several input sentences.

(iii) Attention weights (process) visualisation relates the input and output with its context at the encoder and decoder. It is represented as a bipartite graph (as in *BerViz*), or as a Heatmap Matrix (*InSeq*). In the encoder, self-attention relates the input sequence to itself. In the decoder, two attention types are used: self-attention relates the out-

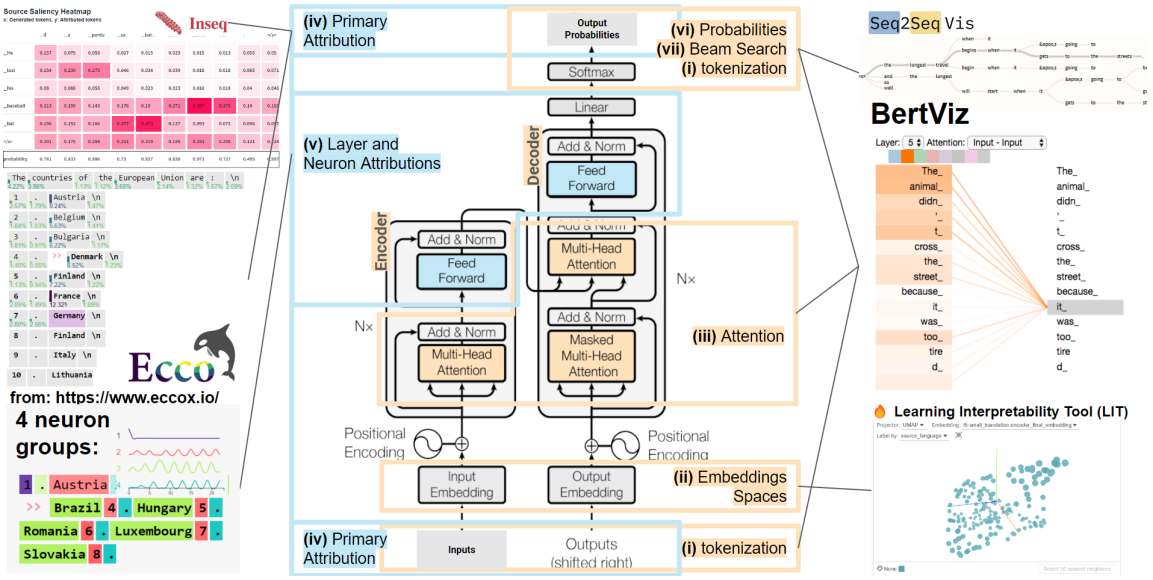

Figure 1: XNMT functionalities mapped in the transformer architecture and a corresponding example.

put to the generated tokens, and cross-attention relates the output tokens to the encoder output. This involves multiple layers and heads, each computing different attention weights, resulting in complex visualisations of multi-head and multi-layer weights for each attention type. [3]

(iv) Primary Attribution (Input/output) visualisation, which can be computed using different methods aims to illustrate the importance relationship between input and output tokens, attributing responsibility to specific input parts for generating each output part. While it reveals the relationship between input and output without explaining the process, it may utilise process information to compute a more accurate attribution metric. *Inseq* system presents this visualisation as a heatmap between inputs (rows) and outputs (columns), with darker colors indicating higher importance weight.

(v) Neuron and Layer Attribution (input/process/output) visualisation tries to pinpoint the responsibility of an output to a specific part of the architecture (a neuron or layer). For instance, *Ecco* illustrates this relationship by considering groups of neurons with a common linguistic task, and *NeuroX* use textual heatmaps to associate inputs with neuron values.

(vi) Probabilities (output) visualisation shows the prediction of the next token across the target dictionary. At each generation step, tokens more likely to appear next have higher probabilities. The visualisation displays top probable tokens, with darker colours indicating higher probabilities.

(vii) Decoding strategy (output) is the final step of the translation process. Here, the search for the sequence translation is exemplified, such as by visualising the beam search. This reveals each generated token step-by-step and how the optimal solution changes with respect to the search strategy and the beam size (i.e., the number of generated solutions). This visualisation is typically presented as a tree graph (e.g., *NMTViz*), offering the advantage of providing translation alternatives.

(viii) Training Data visualisation tries to present the datasets used to train the NMT systems. For example, *seq2seq-vis* uses them in the embedding representation to show similar input and output tokens, and *LIT* presents different data clusters and computes description metrics on them.

The described functionalities comprehensively map the Transformer architecture, although ongoing improvements are needed to develop better methods. In this work, we focus on evaluating them from the translator's perspective.

4 Material and Methods

4.1 Data: Challenge sets

We adopted and, in part, adapted a challenge set (Isabelle et al., 2017), from which a selection of

[3] Attention weights reveal internal model values, yet the link to model outputs is not clear (Jain and Wallace, 2019). We present this visualization to translators without assuming it offers explanations, enabling them to assess its utility.

example segments likely to be mistranslated by NMT were identified. Since the aforementioned challenge set has previously been used to examine translations into French, we decided to apply this to English-Spanish and English-Chinese tests. To assess the explainability of the visualisation tools, ten test segments were selected using a modified version of the challenge set created by Isabelle et al. (2017). Five segments come directly from Isabelle et al. (2017) challenge set and were selected for their varying morpho-syntactic properties, and five additional segments were created to complement this list, in Table 2 (details in appendix A).

(Isabelle et al., 2017)
1. The repeated calls from his mother [should] have alerted us.
2. The woman who [saw] a mouse in the corridor is charming.
3. I requested that families not [be] separated.
4. She was perfect tonight, [was she not]?
5. [Whom] is she going out [with] these days?
New test segments
1. The door [slammed shut].
2. He lost his [baseball bat].
3. The government's new programme [was rolled out] last month.
4. [Berry] is a gifted student.
5. We will [leave no stone unturned] to hold [those responsible] to account

Table 2: Challenge sets

4.2 Models

We resorted to the Helsinki NLP opus models available on Hugging Face (Tiedemann and Thottingal, 2020). The three models (English-French, English-Spanish and English-Chinese) have an encoder-decoder Transformer architecture and use the Sentencepiece algorithm (Kudo and Richardson, 2018). They were chosen for pedagogical and interoperability purposes.

4.3 Visualisation

To present the functionalities to our translation experts, we developed an online web interface available on Hugging Face Spaces (details in Appendix B). Each functionality is built based on a specific state-of-the-art library, as follows:

Top-K and Beam Search Sequence: output probabilities and decoding sequence generation based on *NMTVis*.

Attention: modified version of *BertViz* for the visualisation of attention weights.

Attribution: *Inseq* heatmaps of input X gradient method (Simonyan et al., 2013).

We explore how the explanability visualisations provide information about the challenge sets.

4.4 Explainability Evaluation

As a final global appraisal, we adapt the performance-explainability framework proposed by (Fauvel et al., 2020) to describe the translator analysis in specific for visualisation tools in XNMT. Following (Phillips et al., 2021), we include the evaluation of Meaningfulness, Accuracy, Knowledge Limits Explanations, as follows:

Meaningfulness: Is the explanation intelligible and understandable to the translator? Possible values: 1=no, 2=somewhat, 3=yes

Faithfulness: Can we trust the explanations? Possible values: the explanations are 1=incorrect, 2=imperfect, 3=perfect

Accuracy: Does the explanation accurately reflect the NMT processing? Possible values: 1=no, 2=somewhat, 3=yes

Knowledge limits: Does the explanation show the uncertainties of the NMT prediction? Possible values: 1=no, 2=somewhat, 3=yes

User: What is the target user category of the explanations? Possible values: 1=NMT expert, 2=translation expert, 3=broad audience

Usage: What is the intended use? Possible values: 1=debugging, 2=training, 3=professional use

Information: Which kind of information does the explanation provide? Possible values: 1=inspection, 2=inspection with post-processing, 3=attribution

We conducted a focus group with six translators working in English-Chinese (2), English-Spanish (1) and English-French (3). Each functionality is tested using the same ten source sentences by all users, and finally, the evaluation is the result of a group discussion with the support of NMT experts. We distinguish between translators and NMT experts because each group possesses a different set of skills and knowledge.

5 Results : XNMT evaluation

5.1 Top-K and Beam Search Sequence

Generally, the results produced by the Top-K and beam search sequence tool are both interesting and useful for viewing alternatives, particularly where synonymous words have been considered. For example, for English-Spanish, the example 'I requested that families not be separated' the final target translation uses the verb *solicitar* (literally, *to request*) however *pedir* (*to ask for*/*to request*) was also considered by the machine. Yet the information outlined within the Top-K feature demonstrates that the former received a higher probability, although the tokenisation of the word into three separate tokens *so-licit-é* (literally, *(I) requested*) makes it difficult to obtain any concrete statistical data confirming the probability of the word as a single lexical unit. Similarly for the English-French translation of this sentence, the Top-K visualisation shows a full list of synonymous translations for the English word *requested*, these being: *demander* (kept by the beam search algorithm), *exiger*, *prier*, *réclamer*, *vouloir*, *solliciter*.

In a similar light, when analysing the sentence 'Berry is a gifted student', which becomes *Berry es una estudiante talentoso* (literally, 'Berry is a student talented'), and 'Berry' is tokenised as *Ber-ry*), it is difficult to assess the level of attention given to the more literal alternative such as *baya* (referring to the fruit as opposed to the name). In contrast, this type of tokenisation is useful for translating proper nouns, names in particular, into Chinese. Transliteration is the main way of addressing names between English and Chinese. The way in which 'Berry' is tokenised as *Ber-ry* shows how the model transliterates the name from a phonological perspective.

Moreover, within the same segment, the English-Chinese translation of the word *gifted* can be rendered in various ways, depending on the collocation embedded in the translation. The Top-K probable tokens can thus inform translators of the different possibilities available, therefore aiding translators to make more contextually-aware decisions. In addition, the complementary insights shown across the two visualisations are helpful for translators to navigate themselves among the different possible translations ranging from the level of tokens (i.e., within the Top-K) to the level of semantic trunks or even sentences (i.e., within the beam search sequence tree).

5.2 Attention

The visualisation tool of the multi-layer and multi-headed attention mechanisms can be instrumental in facilitating collaborations between computer scientists and translation practitioners in order to identify at what stage things go right or wrong during the processing stage, thus facilitating the potential to improve the overall performance of the MT model. However, its usefulness for translators is somewhat less optimistic. Within English-Spanish translation, the sentence 'I requested that families not be separated' yields interesting results whereby the subjunctive mood is correctly triggered due to a change of subject, yet the cross attention also demonstrates a high level of attention between the verb *se separaran* (literally,*they would be separated*) in the target translation and the subject of the verb in the English source text. Using the visualisation tool, it is possible to see that the particle *se* (a marker of the medial passive) places a greater amount of importance on the verb 'separated' – potentially suggesting the machine's recognition of the English passive as a grammatical structure. Meanwhile, the verb *seperaran* is tokenised as *separar-an*, with *-an* (the element indicating subject-verb conjugation) placing greater importance on the subject of the verb (families), which again may suggest the machine's ability to recognise verb-subject agreement.

When analysing the English–French translation of the sentence 'He lost his baseball bat.', it is interesting to notice that the encoder's self-attention shows a higher attention weight between the tokens *bat* and *baseball*. This might suggest that the presence of the latter word helps to disambiguate the polysemous word *bat*, and obtain the correct translation 'Il a perdu sa batte de baseball.'.

For English–Chinese translation, this tool is particularly helpful to identify where things start to go wrong, especially when analysing the 'Cross Attention'. Using this tool, users can walk through the layers and locate the layer in which the information started to go wrong. For example, in the sentence 'We will leave no stone unturned to hold those responsible to account', the translation output is "我们将不遗余力地追究责任者的责任" (literally, we will spare no efforts to hold 责任者*zerenzhe* [responsible person] accountable). Here, *zerenzhe* is not commonly used in this context; however, it is a literal translation for 'those responsible', as indicated within the different lay-

ers of the decoder.

Overall, the different language groups involved within this study consider the attention tool the most complex. In many cases, understanding layers as the different stages within the translation is fairly easy to grasp and thus deploy within teaching-based scenarios. For future translators, attention weights are a way to revisit an onomasiological/semasiological approach. Students could be asked to identify the most relevant links where constituents are properly delimited with the attention weights. Conversely, they could use their constituent detection competence to characterise the division of labour for the different layers. Nevertheless, an important limitation should be pointed out: attention-weight visualisation on long sentences is more difficult. From a translation perspective, it would be useful to gain insights into how the model processes long and complex sentences. And, provided with context-sensitive NMT models, it would be interesting to analyse greater-than-sentence-level textual features. This will help to assess the model's reliability on contextual analysis, for example, overall coherence of the translation, consistency of proper nouns and issues with co-referentiality. However, the current visualisation output of long sentences is difficult to interpret and thus the data becomes less meaningful.

5.3 Attribution

The attribution heatmaps have the potential to provide useful insights for the translator, particularly in the case of the source saliency heatmap, which makes it possible to see how words within the source text influence the final target translation. Similarly, the target saliency heatmap focuses on how the previously translated words influence the determination of the following words. Both tools can potentially allow translators to evaluate the efficiency of an MT model from the perspective of contextual cohesion, as well as the model's performance in producing natural collocations. However, the current version of the visualisation contains less focused information than a translator would need. A more interactive user interface might be helpful to enhance the usability of this tool. For example, when demonstrating the English-Spanish sentence 'The repeated calls from his mother should have alerted us', the source saliency heatmap yields no noteworthy results; in the English-Chinese direction, the heatmap shows correct syntactic attention, but due to the fact that it failed to provide a semantically correct translation ('calls' mistranslated as 呼吁 *huyu* (appeal)), it can result in translators' confusion: is the visualisation trustworthy whereas the actual problem might be a lack of training data? However, the target saliency heatmap, shows an increased amount of saliency being given to the verb *deberían* (they should) to confirm its translation of *habernos alertado* (literally, having alerted us), which in Spanish is typically formed with the use of a modal verb such as *deber* (to have to) as we see here. There was one scenario in which the target saliency heatmap proved largely redundant when considering the sentence 'The woman who [saw] a mouse in the corridor is charming' as no colours appeared within the heatmap itself.

5.4 Global appraisal of Functionalities

Among the translators who tested the tools, three of them (one for each language pair) provided a fine-grained evaluation of every visualisation tool in terms of the criteria evoked in section 4.4. However, the following criteria were not rated by the translators but by the NMT experts: Faithfulness, Information, Accuracy and User. The reason for this is that in order to rate how accurately and faithfully a visialisation tool reflects the NMT processes, a detailed understanding of its inner workings is necessary, therefore this information was provided by the experts in the field. Similarly, for the Information, an in-depth understanding of the tool and data processing is needed. As for the User, we consider the definition of a "user by design", predefined by the creators of the tools. These criteria can be considered as being objective, while the remaining can be considered subjective. The latter were assessed by the translators. All the results can be found in figure 2. It presents separately the objective criteria (upper-left) and the subjective criteria, which are presented by language pair (and hence by evaluator).

We remark that the tools that were found most useful (for debugging, training and professional use) were the beam-search tree and the Top-K probabilities visualisation. These two also rank highest in meaningfulness, which indicates that this tool speaks to the translators the most. They also seem most capable of showing the limits in the model. This is probably due to their capacity to show alternative translations. Every

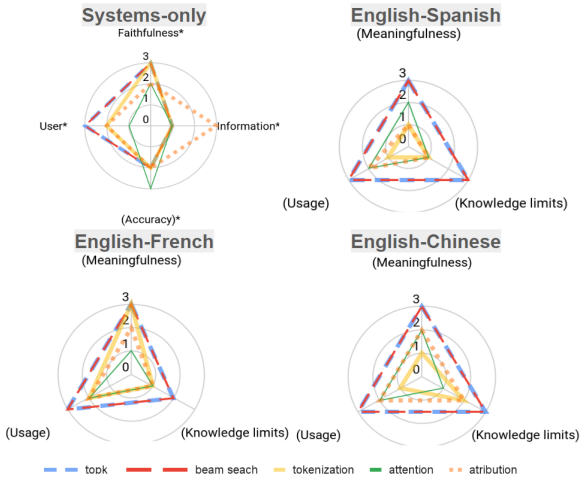

Figure 2: Global XNMT evaluation, upper-left graph shows the assessment done by NMT experts, followed by a graph per language pair (per evaluator) assessed by translators.

evaluator rated these two tools (Top-K and beam search) identically, which indicates a strong similarity between them. The attention visualisation tool proved to be only somewhat meaningful even though it is the most accurate of the visualisations. The evaluators agree that this visualisation doesn't permit them to detect the knowledge limits of the MT model, in fact this tool rated lowest in both Knowledge limits and Meaningfulness criteria. However, they do state that it can be useful not only for debugging purposes but as well during training of translation students. The visualisation of tokenisation is the tool on which the evaluators were less unanimous in terms of its meaningfulness. It was rated as not meaningful (1 out of 3) by two evaluators and as meaningful (3 out of 3) by one evaluator. This is probably caused by the fact that some evaluators found it misleading to see the probabilities of generating tokens rather than the the probability of the whole word. The Attribution method was rated mildly meaningful and not very capable of showing the model's limits. In combination with its low faithfulness and low accuracy due to the inner workings of this tool, it doesn't seem to be very useful for translators.

6 Discussion

6.1 XNMT for translation

From the translator perspective, one could hypothesise that there are parallels between traditional translation approaches (i.e., human translation) and the visualisation functionalities that we have tested. For example, Nida (1964) concept of three-stage translation systems emphasises source text analysis, kernel extraction, the transfer, and the restructuring of meaning in the target language. The inspection methods of NMT (in particular, cross attention) resemble a deep understanding of the source text that is required to grasp its semantic and syntactic nuances; namely, the layers of attention forming part of the way in which NMT analyses and encodes the source sentence into representations capturing its meaning. Extracting the kernel, or comprehending the fundamental meaning of a sentence, could be linked to attribution methods (e.g., saliency), through which translators gain insights into the semantic and syntactic elements of a sentence that the model pays attention to. This process helps translators to understand how the model makes decisions and restructures the target translation to fit the norms of the target language. These links have the potential to help translators and trainees gain a superficial insight into the 'thinking process' of NMT models and expand their perceptions of the trustworthiness and reliability of such models. The inspection and attribution methods are similar to how a human translator might refine their understanding and approach to translation, which will lead to a continuous human-informed improvement cycle for NMT explainability.

6.2 Visualisation in CAT tools

We need to take into account the current computer-assisted translation (CAT) tools available to professional translators and discuss visualisation tools already at their disposal. Professional translators use Translation management systems (TMS), typically CAT tools, for their daily work. The main features of a TMS include project management, translation memory, terminology management, collaboration and review, reporting and analytics, automation and other systems integrations, etc. A TMS enables translators to leverage resources including translation memories, terminology databases and MT engines. This allows them to reuse previously translated segments and/or use raw MT output as a starting point for human translation, whilst maintaining consistency in terminology and phrasing. Since MT is usually an integrated feature of a TMS, translators either use MT suggestions as reference or directly post-edit the raw MT outputs. The working processes of a MT are not a primary concern for translators during

their regular workflows. Nevertheless, their ongoing automation anxieties need to be addressed and MT literacy is part of the overarching strategy.

It is important to take into account the levels of visualisation that translators and trainees can take in when promoting NMT explainability. The deep visualisations that we tested are very distant from their daily work. The visualisations in a TMS are functional-oriented which typically include the indicators of matches found in translation memories and terminology databases in the form of colour-coding, underlining, or other visual cues, the list of suggestions generated from concordance search, flagged potential quality issues, and progress bars, etc. These visualisations are set to present the complex information in a more intuitive and user-friendly manner, helping translators work more efficiently. In contrast, deep visualisation requires basic knowledge in NMT and clear guidelines to ensure correct interpretation. This is also the next step of our project: a workshop to disseminate the visualisation toolkit and to test the translators' and trainees' reception, and explore its wider usage.

6.3 Additional Functionalities

With the advent of Large Language Models (LLMs), it is tempting to use LLMs for Automatic Post-Editing of translations, a pipeline already implemented for Automatic Speech Recognition (ASR) systems such as WhisperingLlama (Radhakrishnan et al., 2023), which uses Llama (Touvron et al., 2023) to regularise and optimise Whisper's outputs (Radford et al., 2023) for ASR transcriptions.

We are already witnessing an integration of generative-AI with TMS; for example, the web-based system Wordscope. Along with the essential TMS features, the system integrates ChatGPT with ready-made prompts that allow translators to look up terms, search a topic or a concept, explore alternative expressions, back translate into the source language for quality check, proofread, post-editing, and more. A potential research question here is: can LLMs facilitate XNMT? Considering the increasing integration of generative-AI into the workflows of tech-savvy translators, is it possible to use LLMs to enhance the interpretation of NMT visualisations? For example, using generative-AI to help analyse the linguistic challenges that might be overlooked by human, and compare the results with the visualisations to foster more comprehensive evaluation.

In addition to the proposed research questions, the focus group highlighted several desires and requirements for XNMT to be fully deployed within the translator's workflow. One of the more divisive of which included potentially changing the presentation of subtokens within the final visualisation output (i.e., presenting the whole word as a single lexical item e.g., *berr-ry* as 'Berry'). Whilst it is generally understood that tokenisation forms an indispensable element of how the machine understands and process language (a feature enjoyed by the tech-savvy and developers of XNMT), the lack of a single overall probability for the entire lexical unit makes it challenging for translators to obtain meaningful statistical data that could be used to inform the translator's decision making processes.

7 Conclusion

In this paper, we have summarised the main visualisation tools adapted for XNMT, detailing the functionalities implemented in a prototype and discussing the potential benefits for translators. Our innovation lies in highlighting the translator's viewpoint and utilising XNMT to provide accountability for translators. This entails gaining a better grasp of the training data, monitoring the learning phase, or finding ways to understand the entire NMT process. As future work, we will continue exploring additional visualisation tools and evaluating their use, specifically focusing on one of the following translation moments: (i) initiating use of a new technology to understand NMT system workings during training, (ii) beginning a project to comparing, trusting, and selecting the best NMT, (iii) analyzing the translation process to identify reasons for poor output, and (iv) evaluating translation results e.g. to test alternatives.

Acknowledgement

This paper emanated from research 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)). This work was also supported by the CREMA project (Coreference REsolution into MAchine translation) funded by the French National Research Agency (ANR), contract number ANR-21-CE23-0021-01.

References

Alammar, J. 2021. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations*, pages 249–257.

Alharbi, Mohammad, Matthew Roach, Tom Cheesman, and Robert S Laramee. 2021. Vnlp: Visible natural language processing. *Information Visualization*, 20(4):245–262.

Bau, Anthony, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*.

Bodria, Francesco, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and survey of explanation methods for black box models. *ArXiv*, abs/2102.13076.

Cui, Wenqiang. 2019. Visual analytics: A comprehensive overview. *IEEE access*, 7:81555–81573.

Dalvi, Fahim, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019a. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Dalvi, Fahim, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019b. Neurox: A toolkit for analyzing individual neurons in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9851–9852.

de Seyssel, Maureen, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. *arXiv preprint arXiv:2203.16193*.

Ding, Shuoyang, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. *arXiv preprint arXiv:1906.10282*.

Fauvel, Kevin, Véronique Masson, and Elisa Fromont. 2020. A performance-explainability framework to benchmark machine learning methods: application to multivariate time series classifiers. *arXiv preprint arXiv:2005.14501*.

Guerberof-Arenas, Ana and Joss Moorkens. 2023. Ethics and machine translation: The end user perspective. In *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 113–133. Springer.

Isabelle, Pierre, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In Palmer, Martha, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jain, Sarthak and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Bansal, Mohit and Heng Ji, editors, *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Kokhlikyan, Narine, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Kreutzer, Julia, Jasmijn Bastings, and Stefan Riezler. 2019a. Joey nmt: A minimalist nmt toolkit for novices. *arXiv preprint arXiv:1907.12484*.

Kreutzer, Julia, Jasmijn Bastings, and Stefan Riezler. 2019b. Joey NMT: A minimalist NMT toolkit for novices. In Padó, Sebastian and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China, November. Association for Computational Linguistics.

Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Leiter, Christoph, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2023. Towards explainable evaluation metrics for machine translation. *arXiv preprint arXiv:2306.13041*.

Madsen, Andreas, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Munz, Tanja, Dirk Väth, Paul Kuznecov, Ngoc Thang Vu, and Daniel Weiskopf. 2021. Visualization-based improvement of neural machine translation. *Computers Graphics*.

Nida, Eugene Albert. 1964. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Ott, Myle, 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 NAACL-HLT 2019: Demonstrations*.

Phillips, P Jonathon, P Jonathon Phillips, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. 2021. Four principles of explainable artificial intelligence.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Radhakrishnan, Srijith, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023. Whispering llama: A cross-modal generative error correction framework for speech recognition. *arXiv preprint arXiv:2310.06434*.

Rikters, Matīss and Mark Fishel. 2017. Confidence through attention. In Kurohashi, Sadao and Pascale Fung, editors, *Proceedings of Machine Translation Summit XVI: Research Track*, pages 299–311, Nagoya Japan, September 18 – September 22.

Sarti, Gabriele, Nils Feldhus, Ludwig Sickert, Oskar Van Der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. *arXiv preprint arXiv:2302.13942*.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Stahlberg, Felix. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Strobelt, Hendrik, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description 3: Grammatical categories and the lexicon*, pages 57–149.

Tan, Zhixing, Jiacheng Zhang, Xuancheng Huang, Gang Chen, Shuo Wang, Maosong Sun, Huanbo Luan, and Yang Liu. 2020. THUMT: An open-source toolkit for neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 116–122, Virtual, October. Association for Machine Translation in the Americas.

Tenney, Ian, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122*.

Tiedemann, Jörg and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vieira, Lucas Nunes, Carol O'Sullivan, Xiaochun Zhang, and Minako O'Hagan. 2023. Machine translation in society: insights from uk users. *Language Resources and Evaluation*, 57(2):893–914.

Vig, Jesse. 2019. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR workshop: Debugging machine learning models*, volume 3.

Wiegreffe, Sarah and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Liu, Qun and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Zhou, Yilun, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2022. Do feature attribution methods correctly attribute features? In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 36, pages 9623–9633.

A Appendix - Challenge Sets

Tables 3 and 4 present the full set of challenging translations used to evaluate each explainability visualisation tool. Following Isabelle's 2017 procedure, we define each translation example with its related challenge.

These segments have been included within this study for their various grammatical and linguistic features. Firstly, (Talmy, 1985) distinguishes between two main language groups: those that favour conflation of path with motion (e.g., the Romance languages), and those that favour conflation of manner with motion (e.g., Germanic, and Slavic languages). Chinese appears to fall somewhere between the two, albeit with a slight preference towards the latter. Verbs of manner and path, which (Isabelle et al., 2017) call 'crossing movement verbs', present a difficulty due to the lexical and syntactic challenges arising when translating between languages that conflate information differently. The example sentence 'The door [slammed shut]' was used to examine how the model responds to such verbs. We also presented challenges involving prepositional verbs, which can result in inaccuracies concerning active vs passive voice, as well as syntax, or overly literal translations (i.e., within the sentence 'The government's new programme [was rolled out] last month').

A second key difference between languages concerns the productivity of compounding as a means of word formation. Both English and Chinese are especially productive in this regard; however, this is not necessarily the case within languages such as French and Spanish. Compound nouns present difficulties by way of differences in phrasal word order (i.e., modifier + noun vs noun + modifier) in addition to potential issues with lexical ambiguity or polysemy. We used the sentence 'He lost his [baseball bat]' to test the model's ability to identify polysemous words such as 'bat' (i.e., object vs animal). And finally, in addition to testing polysemy with compound nouns, we also tested polysemy within proper nouns or names, with a particular focus on examining the visualisation of data where names are likely to be transliterated. In this instance, the sentence '[Berry] is a gifted student' was used.

Sentence	Challenge
The repeated calls from his mother [should] have alerted us.	Is subject-verb agrement correct? (Possible interference from distractors between the subject's head and the verb).
The woman who [saw] a mouse in the corridor is charming.	Are the agreement marks of the flagged participles the correct ones? (Past participle placed after auxiliary AVOIR agrees with verb object iff object precedes auxiliary. Otherwise participle is in masculine singular form).
I requested that families not [be] separated.	Is the flagged verb in the correct mood? (Certain triggering verbs, adjectives or subordinate conjunctions, induce the subjunctive mood in the subordinate clause that they govern).
She was perfect tonight, [was she not]?	Is the English "tag question" element correctly rendered in the translation?
[Whom] is she going out [with] these days?	Is the dangling preposition of the English sentence correctly placed in the French translation?

Table 3: (Isabelle et al., 2017) challenges used to evaluate the explainability visualisation tools

B Appendix - Visualisation tools

In this section, we describe the implemented functionalities[4].

General Interface We have created three demos available as spaces on the HuggingFace platform, all built using Gradio and Javascript. In Figure 3, we present the general interface, where translators can either choose a challenge or input a source text in English. The text is subsequently translated based on the selected model (en-zh for Chinese, en-es for Spanish, and en-fr for French).

Probabilities: Top-k Figure 4 shows the top-k most probable tokens to be generated, where in this case, k=10. The probability is represented on a scale of grey colours. At each generation step,

[4]Publicly available at anonymous

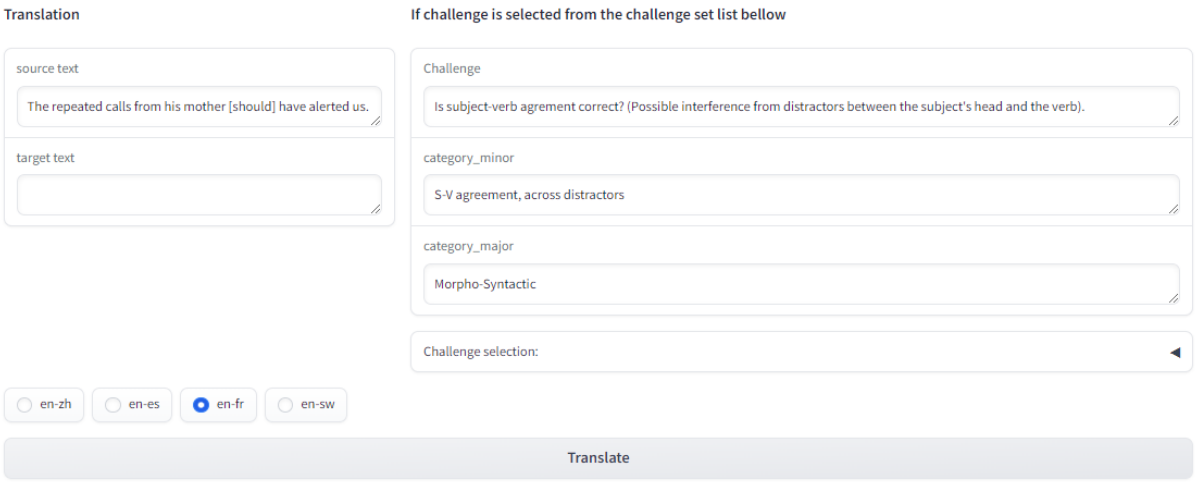

Figure 3: General Interface

Sentence	Challenge
The door [slammed shut].	Verb of manner and path - How has the manner and path been conflated, and does this follow the typical patterns of the target language?
He lost his [baseball bat].	Has baseball bat been translated as a compound noun, or two separate lexical items?
The government's new programme [was rolled out] last month.	Similiar to verb of manner and path with added syntatical difficulties and passive vs active voice.
[Berry] is a gifted student.	Has Berry's name been translated literally? Transliterated?
We will [leave no stone unturned] to hold [those responsible] to account.	How has the idiomatic expression been translated? Has the syntax been adjusted accordingly?

Table 4: New challenge set used to evaluate the explainability visualisation tools

the top-k probable tokens are presented. According to the tokenisation used, one or several tokens could correspond to a single word. For instance, the word *alerter* was generated in two steps: first *alerte*, and then *r*.

Decoding Strategy: Beam Search Sequence Generation The beam search visualisation is a simplified representation of the "beam search" decoding strategy, aiming to find the best "global" translation, i.e., the best sequence of translated tokens. Figure 5 displays the beam search decoding sequence generation using a beam size of 4. This visualisation presents the sequences (4) of output tokens in a tree structure, allowing users to notice the differences between alternatives. The top branch represents the sequence with the highest probability, while less likely sequences are displayed below.

Attention The attention visualisation shows the multi-layer and multi-head attention mechanism used in the transformer architecture. Each layer comprises several heads, each learning different weights between compared elements (tokens of the source or translated sentence). In the visualisation, each head is represented by a colour, with darker colours indicating higher attention weights. This information is represented through connection lines and coloured boxes. Three attention options are presented in the visualisation: (i) *encoder self-attention*, which relates the tokens of the source text to each other; (ii) *decoder self-attention*, which relates the translated tokens to the previously generated tokens; and (iii) *cross-attention*, which relates the translated tokens and

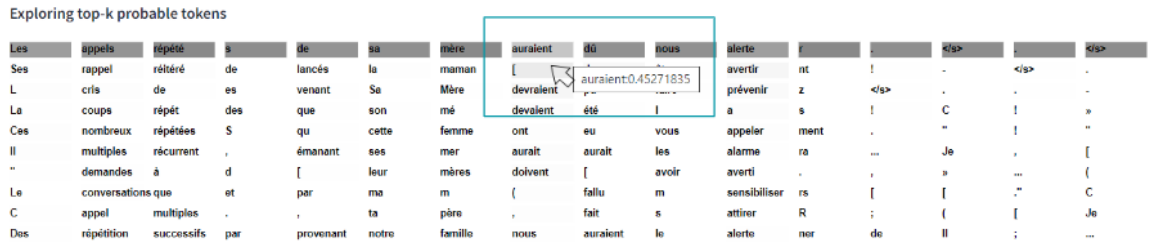

Figure 4: TopK probable tokens

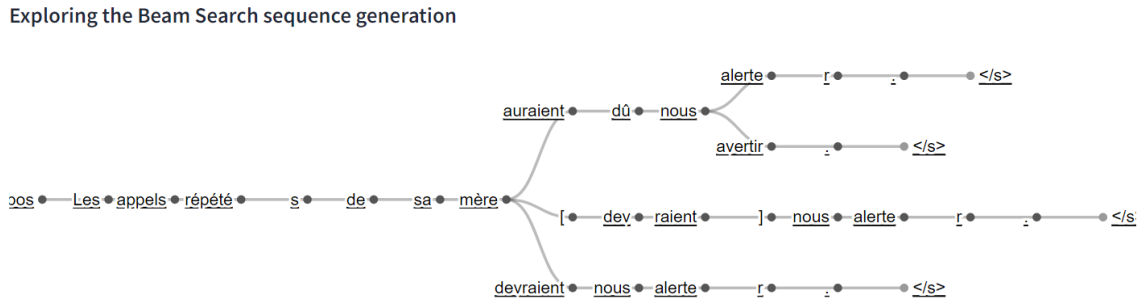

Figure 5: Beam Search Sequence Generation

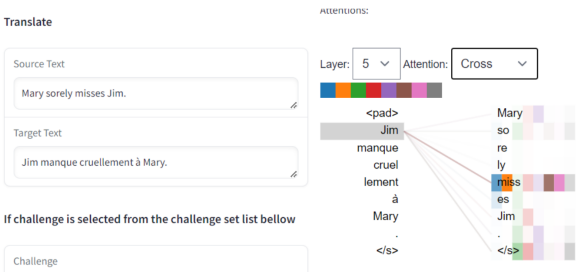

Figure 6: Attention Visualisation

Figure 7: Primary Attribution

the source tokens. For example in Figure 6, in the 5th layer of cross-attention, *Jim* is strongly related to *miss* for heads represented by the colours blue, orange, brown and pink.

Attribution The attribution visualisation presents the importance of each token of the source text (rows) in generating the tokens of the translated text (columns). This attribution is computed using the input X gradient method (Simonyan et al., 2013). In the heatmap, the importance of compared tokens is indicated by the darkness of the colour. For example, in Figure 7, the most important token for generating *doué* is *gift*.

TopK and Beam Search Sequence Generation functionalities are based on state-of-the-art tools. However, they are implemented by us. For attention visualisation, we adapted the *BertViz* library to make it compatible with Gradio, while the *Inseq* library made possible the attribution visualisation.