Research papers published by Lingua Custodia Lab

English version

GraphRAG: Leveraging Graph-Based Efficiency to Minimize Hallucinations in LLM-Driven RAG for Finance Data

Mariam Barry, Gaëtan Caillaut, Pierre Halftermeyer, Raheel Qader, Mehdi Mouayad, Dimitri Cariolaro, Fabrice Le Deit, Joseph Gesnouin

→ *https://aclanthology.org/2025.genaik-1.6.pdf*

The aim of this study is to propose novel approaches to integrate structured data from knowledge graphs into RAG systems. We aim to overcome the major shortcomings of LLMs used for RAG, which are hallucinations and excessive compute resources consumption. To this end, we propose several integration methods and show that providing contextual information from a knowledge graph reduces the probability of producing hallucinations. We also show that it is possible to drastically reduce the consumption (in terms of input tokens) of the LLM by submitting information encoded in the form of *edges*, or RDF triplets, without compromising data quality.

DOLFIN - Document-Level Financial Test-Set for Machine Translation

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, Hervé Blanchon

→ *In review for NAACL 2025.*

In this work we aim to fill the gap in document-level testing data. Despite the research interest in document-level Machine Translation, the test-sets dedicated to this task are still scarce and fall short on specialised domains, such as finance. Also, despite their document-level aspect, they still follow a sentence-level logic that doesn't allow for including certain linguistic phenomena such as information reorganisation. Therefore, we propose DOLFIN, a novel document-level test-set built from specialised financial documents and that makes a step towards true document-level MT by abandoning the paradigm of perfectly aligned sentences, presenting data in units of sections rather than sentences.

Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention. *Revue TAL : traitement automatique des langues*, 2024, 64 (3).

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

→ *https://pasteur.hal.science/ILCEA4/hal-04581509v1*

In this journal paper we analyse Machine Translation models' attention weights as an explanation method for Context-Aware Neural Machine Translation (CA-NMT). Since its evaluation often concerns the evaluation of resolving discourse phenomena ambiguity, we perform analyses and evaluations over coreference links in a

parallel corpus. We propose a human evaluation over attention 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.

The MAKE-NMTViz Project: Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators. EAMT : European Association for Machine Translation, Jun 2024, Sheffield, United Kingdom.
Gabriela Gonzalez-Saez, Fabien Lopez, Mariam Nakhle, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Caroline Rossi, Didier Schwab, Jun Yang, James Robert Turner, Nicolas Ballier
→ https://hal.science/hal-04638945/document
This paper describes the MAKE-NMTViz project, which is designed to help translators without a Machine Learning background understand Neural Machine Translation using explainability visualisation tools.

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, Didier Schwab, Jun Yang
This paper describes work on Explainability and Visualisation to foster collaborations between translators and computational scientists. We describe how visualisation methods can be used to explain Neural Machine Translation outputs to translation professionals without a Machine Learning background. We tested several visualisation functionalities and performed an evaluation with translators (professionals and trainees) within the framework of performance-explainability, focusing on the translator's perspective.

Scaling Laws of Decoder-Only Models on the Multilingual Machine Translation Task
Gaëtan Caillaut, Raheel Qader, Mariam Nakhlé, Jingshu Liu and Jean-Gabriel Barthélemy
→ *https://aclanthology.org/2024.wmt-1.124/ WMT 2024*
This work aims to establish a transition from Lingua Custodia's machine translation activities to more modern ones based on generative models (LLM). To this aim, we study the behaviour of decoder-only models on the task of multilingual and multi-domain machine translation. This architecture is indeed not very widespread in the field of machine translation, but is the reference one in the current state of generative AI.
As part of this study, we trained 6 models with sizes ranging from 70M to 7B parameters. These experiments have validated the relevance of decoder-only architectures for machine translation, but also leave doubts regarding the confidence that we place in scaling laws. Indeed, we show that although all models follow the same trend, which can be described by a power law, it is difficult (if possible) to identify a real universal law that applies to all language directions and all domains.
Finally, this work highlights a "flaw" in the way generative models are trained. The problem is related to a non-uniform use of certain special tokens during the model training phase.

Improve Context-Aware Machine Translation with Negative Sampling and Focused Masking.
Gaëtan Caillaut, Mariam Nakhlé, Jingshu Liu et Raheel Qader.
→ *https://aclanthology.org/2024.jeptalnrecital-taln.20/ TALN 2024*
This work is part of the efforts to develop context-aware translation models. It proposes a method to encourage the model to take context more into account by augmenting the training data. It consists of two data augmentation strategies. Negative sampling, i.e., adding incorrect examples (deliberately corrupted) with a loss function that penalises the probability assigned to the incorrect word, and focused masking which masks tokens in the source sentence with a strong link to the previous context to force the model to seek relevant information in the context rather than in the source sentence.

Large language model adaptation for financial sentiment analysis. In Proceedings of the 6th Workshop on Financial Technology and Natural Language Processing (FinNLP). 2023
Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaëtan Caillaut, and Jingshu Liu.
→ *https://aclanthology.org/2023.finnlp-2.1*
This work proposes a method for adapting large language models to the financial domain. The adaptation of Pythia-1.4B and OPT-1.3B models was studied through continued pre-training on financial domain training data and instruction training on specific financial tasks (such as sentiment prediction in financial texts). The experiments show that the resulting models achieve performance equal to or higher than much larger models trained on larger amounts of data.

Lingua Custodia's Participation at the WMT 2023 Terminology Shared Task. In Proceedings of the Eighth Conference on Machine Translation, pages 897–901, Singapore. Association for Computational Linguistics. 2023
Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout, and Raheel Qader.
→ **https://aclanthology.org/2023.wmt-1.81/**
This work presents Lingua Custodia's participation in the Terminology Shared Task at the WMT 2023 conference. The proposed method involves extracting a bilingual glossary from the training data in a completely unsupervised manner and annotating the training data using this glossary. Thus, the model is led to learn to use the desired target term in the generated translation. The method also includes a re-sampling step of annotated data to ensure that the model has been trained with various types of terms. The method proves effective in the language pairs studied, notably English-Czech and German-English.

The MAKE-NMTVIZ System Description for the WMT23 Literary Task. In Proceedings of the Eighth Conference on Machine Translation, pages 287–295, Singapore. Association for Computational Linguistics. 2023
Fabien Lopez, Gabriela González, Damien Hansen, Mariam Nakhle, Behnoosh Namdarzadeh, Nicolas Ballier, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Sadaf Mohseni, Caroline Rossi, Didier Schwab, Jun Yang, Jean-Baptiste Yunès, and Lichao Zhu.
→ ***https://aclanthology.org/2023.wmt-1.30/***
This work presents the participation of the MAKE-NMT-VIZ team in the Literary Translation Shared Task at the WMT 2023 conference, which focused on translating Chinese to English for the "web novel" genre. Three models were proposed: a sentence-level model fine-tuned from mBART50, a context-sensitive model that models the source context by concatenating the three preceding sentences, and a contrastive model with a classical Transformer architecture without context modeling. The work also includes an in-depth human analysis presented by native Chinese-speaking translators.

L'évaluation de la traduction automatique du caractère au document : un état de l'art. In Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL). 2023.
Mariam Nakhlé.
→ ***https://aclanthology.org/2023.jeptalnrecital-rjc.10/***
This work provides an overview of the state of the art of machine translation evaluation. It outlines human evaluation approaches and automatic evaluation methods, differentiating between families of approaches—string-based and learned metrics—with a particular focus on document-level evaluation. A section is devoted to the meta-evaluation of metrics.

Lingua Custodia's Participation at the WMT 2022 Word-Level Auto-completion Shared Task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1170–1175, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. 2022.
Melissa Ailem, Jingshu Liu, Jean-gabriel Barthelemy, and Raheel Qader.
→ ***https://aclanthology.org/2022.wmt-1.118/***
This work presents Lingua Custodia's participation in the Word-Level Auto-completion Shared Task at the WMT 2022 conference. This task aims to develop systems capable of completing a translation after a user has inserted a few characters. The language pairs studied are German-English and English-German. The proposed method treats the sequence entered by the user as a constraint, encouraging the model to predict a word that starts with such a sequence. Additionally, the method uses a joint optimization strategy to account for different types of translation context.

Lingua Custodia's Participation at the WMT 2021 Machine Translation Using Terminologies Shared Task. In Proceedings of the Sixth Conference on Machine Translation. 2021.
Melissa Ailem, Jingshu Liu, and Raheel Qader.
→ ***https://aclanthology.org/2021.wmt-1.78/***
This work presents Lingua Custodia's participation in the Terminology Control Shared Task at the annual WMT 2021 conference. The method we used in this shared task is based on our previously published paper at ACL 2021. Empirical results show that in the language pairs studied (notably English-French, English-Russian, and English-Chinese), the method satisfies the majority of terminological constraints while maintaining high translation quality.

Encouraging Neural Machine Translation to Satisfy Terminology Constraints. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1450–1455, Online. Association for Computational Linguistics. 2021.
Melissa Ailem, Jingshu Liu, and Raheel Qader.
→ ***https://aclanthology.org/2021.findings-acl.125/***
This work presents a new approach to forcing neural machine translation models to respect terminology constraints from a glossary. This method operates during training and involves adding training data containing terminological constraints surrounded by special tokens. The source term is masked to reinforce the copying behavior of the desired term into the generated translation. This method has the advantage of not adding any latency during inference.

French version

GraphRAG: Leveraging Graph-Based Efficiency to Minimize Hallucinations in LLM-Driven RAG for Finance Data. GenAIK, COLING 2025
Mariam Barry, Gaëtan Caillaut, Pierre Halftermeyer, Raheel Qader, Mehdi Mouayad, Dimitri Cariolaro, Fabrice Le Deit, Joseph Gesnouin
→ *https://aclanthology.org/2025.genaik-1.6.pdf*
Ces travaux de recherche ont pour objectif d'étudier et de proposer des approches pour intégrer les données structurées présentes dans un graphe de connaissance dans un système de RAG. L'objectif étant de paliers aux défauts majeurs des LLM utilisés pour le RAG, à savoir les problèmes d'hallucination et leurs consommations excessive de ressources. Pour ce faire nous proposons plusieurs méthodes d'intégration et nous montrons que fournir des informations contextuelles provenant d'un graphe de connaissances permet de réduire la

probabilité de produire des hallucinations. Nous montrons également qu'il est possible de réduire drastiquement la consommation (en termes de tokens d'entrée) du LLM en lui soumettant des informations encodées sous formes d'arc, ou de triplets RDF, sans pour autant compromettre la qualité des données.

DOLFIN - Document-Level Financial Test-Set for Machine Translation
Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, Hervé Blanchon
→ *NAACL 2025.*
Dans ce travail, nous visons à combler le manque de données de test au niveau des documents pour la Traduction Automatique. Malgré l'intérêt scientifique pour la Traduction Automatique (TA) au niveau des documents, les jeux de tests dédiés à cette tâche sont encore rares et ne couvrent pas les domaines spécialisés, tels que la finance. De plus, malgré le fait qu'ils sont au niveau du document, ils suivent toujours une logique de phrases qui ne permet pas d'inclure certains phénomènes linguistiques tels que la réorganisation de l'information. Par conséquent, nous proposons DOLFIN, un jeu de test innovant au niveau du document construit à partir de documents financiers spécialisés et qui fait un pas vers une véritable TA au niveau du document en abandonnant le paradigme des phrases parfaitement alignées et en présentant les données sous forme d'unités de sections plutôt que de phrases.

Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention. *Revue TAL : traitement automatique des langues*, 2024, 64 (3).
Marco Dinarelli, Dimitra Niaouri, Fabien Lopez, Gabriela Gonzalez-Saez, Mariam Nakhlé, Emmanuelle Esperança-Rodier, Caroline Rossi, Didier Schwab and Nicolas Ballier
→ *https://pasteur.hal.science/ILCEA4/hal-04581509v1*
Dans cet article de journal, nous analysons les poids d'attention des modèles de traduction automatique en tant que méthode d'explicabilité pour la Traduction Automatique Neuronale au niveau du document (CA-NMT). Comme son évaluation concerne souvent la résolution d'ambiguïté des phénomènes discursifs, nous effectuons des analyses et des évaluations sur les liens de coréférence dans un corpus parallèle. Nous proposons une évaluation humaine basée sur les heatmaps d'attention, renforcée par une évaluation quantitative basée sur les poids d'attention sur les liens de coréférence et avec différentes métriques conçues spécialement pour ce travail. Ces métriques fournissent une évaluation plus explicite des modèles CA-NMT que les évaluations utilisant les test-suites contrastives.

The MAKE-NMTViz Project: Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators. EAMT : European Association for Machine Translation, Jun 2024, Sheffield, United Kingdom.
Gabriela Gonzalez-Saez, Fabien Lopez, Mariam Nakhle, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Caroline Rossi, Didier Schwab, Jun Yang, James Robert Turner, Nicolas Ballier
→ *https://hal.science/hal-04638945/document*
Ce travail décrit le projet MAKE-NMTVis qui vise à aider les traducteurs sans connaissance en Apprentissage Automatique à mieux comprendre la Traduction Automatique Neuronale grâce aux outils de visualisation et d'explicabilité.

Exploring NMT Explainability for Translators Using NMT Visualising Tool. EAMT : European Association for Machine Translation, Jun 2024, Sheffield, United Kingdom.
Gabriela Gonzalez-Saez , Mariam Nakhlé, James Robert Turner, Fabien Lopez, Nicolas Ballier, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Raheel Qader, Caroline Rossi, Didier Schwab, Jun Yang
→ *https://hal.science/hal-04581586/*
Cet article décrit le travail sur l'Explicabilité et la Visualisation pour promouvoir les collaborations entre les traducteurs et les informaticiens. Nous décrivons comment les méthodes de visualisation peuvent être utilisées pour expliquer les sorties de Traduction Automatique Neuronale aux traducteurs professionnels sans connaissance en Apprentissage Automatique. Nous avons testé un nombre de fonctionnalités de visualisation

et nous avons mené une évaluation avec des traducteurs (professionnels et étudiants) dans un cadre de performance et explicabilité, avec un accent sur la perspective du traducteur.

Scaling Laws of Decoder-Only Models on the Multilingual Machine Translation Task
Gaëtan Caillaut, Raheel Qader, Mariam Nakhlé, Jingshu Liu et Jean-Gabriel Barthélemy
→ *https://aclanthology.org/2024.wmt-1.124/, WMT 2024*
Ces travaux visent à établir une transition depuis les activités de Lingua Custodia liées à la traduction automatique vers celles, plus modernes, reposant sur les modèles génératifs (LLM). Pour ce faire, nous étudions le comportement de modèles 100% décodeur sur la tâche de traduction automatique multi-langues et multi-domaines. Cette architecture est en effet peu répandue dans le domaine de la traduction automatique, mais est la référence dans le domaine de l'IA générative.
Dans le cadre de cette étude, nous avons entraîné 6 modèles dont les tailles varient entre 70M et 7B paramètres. Ces expériences ont permis de valider la pertinence des architectures 100% décodeurs pour la traduction automatique, mais laissent également dubitatif vis-à -vis de la confiance que l'on accorde aux lois d'échelle (scaling laws). En effet, nous montrons que bien les modèles suivent tous une même tendance, pouvant être décrite par une loi de puissance (power law), il est au moins difficile de dégager une réelle loi universelle s'appliquant à toutes paires de langue et à tous les domaines.
Enfin, ces travaux ont permis de déceler une "faille" dans la manière dont les modèles génératifs sont entraînés. Le problème est lié à une utilisation non-uniforme de certains tokens spéciaux lors de la phase d'entraînement du modèle.

Améliorer la traduction au niveau du document grâce au sur-échantillonnage négatif et au masquage ciblé
Gaëtan Caillaut, Mariam Nakhlé, Jingshu Liu et Raheel Qader.
→ *https://aclanthology.org/2024.jeptalnrecital-taln.20/ TALN 2024*
Ce travail s'inscrit dans la lignée des travaux cherchant à développer des modèles de traduction sensibles au contexte. Il propose une méthode pour encourager le modèle à prendre en compte davantage d'informations contextuelles en agissant au niveau de la donnée d'entraînement. Nous proposons deux stratégies d'augmentation de données. L'échantillonnage négatif, c-à-d ajout d'exemples incorrectes (délibérément corrompus) avec une fonction de perte qui pénalise la probabilité attribuée au mot incorrecte, et le masquage ciblé qui masque les tokens de la phrase source avec un lien fort au contexte précédent pour ainsi forcer le modèle à chercher l'information pertinente dans le contexte plutôt que dans la phrase source.

Large language model adaptation for financial sentiment analysis. In Proceedings of Workshop on Financial Technology and Natural Language Processing (FinNLP). 2023
Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaëtan Caillaut, and Jingshu Liu.
→ *https://aclanthology.org/2023.finnlp-2.1*
Ce travail propose une méthode d'adaptation de Grands Modèles de Langage au domaine de la finance. L'adaptation des modèles Pythia-1.4B et OPT-1.3B a été étudiée, à travers d'entraînement continué sur la donnée d'entraînement du domaine financier et d'entraînement par instructions sur des tâches spécifiques à la finance (telle que la prédiction de sentiment dans les textes financiers). Les expériences montrent que les modèles ainsi obtenus atteignent des performances égales ou plus élevées aux modèles beaucoup plus larges, entraînés sur de larges quantités de données.

Lingua Custodia's Participation at the WMT 2023 Terminology Shared Task. In Proceedings of the Eighth Conference on Machine Translation, pages 897–901, Singapore. Association for Computational Linguistics. 2023
Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout, and Raheel Qader.
→ **https://aclanthology.org/2023.wmt-1.81/**

Ce travail présente la participation de Lingua Custodia à la tâche commune de Contrôle de Terminologie à la conférence WMT 2023. La méthode proposée consiste en extraction d'un glossaire bilingue de la donnée d'entraînement d'une façon complètement non-supervisée et en annotation de la donnée d'entraînement en utilisant ce glossaire. Ainsi, le modèle est mené à apprendre à employer le terme cible souhaité dans la traduction générée. La méthode inclut également une étape de re-échantillonage de données annotées pour s'assurer que le modèle a été entraîné avec des types de termes variés. La méthode s'avère efficace dans les paires de langues étudiées, notamment le anglais-tchèque et l'allemand-anglais.

The MAKE-NMTVIZ System Description for the WMT23 Literary Task. In Proceedings of the Eighth Conference on Machine Translation, pages 287–295, Singapore. Association for Computational Linguistics. 2023
Fabien Lopez, Gabriela González, Damien Hansen, Mariam Nakhle, Behnoosh Namdarzadeh, Nicolas Ballier, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Sadaf Mohseni, Caroline Rossi, Didier Schwab, Jun Yang, Jean-Baptiste Yunès, and Lichao Zhu.
→ ***https://aclanthology.org/2023.wmt-1.30/***
Ce travail présente la participation de l'équipe MAKE-NMT-VIZ dans la tâche commune de Traduction Littéraire au sein de la conférence WMT 2023, cette année-là concentrée sur la traduction de chinois vers l'anglais du "roman web". Trois modèles ont été proposés, notamment un modèle opérant au niveau de la phrase, fine-tuné à partir de mBART50. Ensuite, un modèle sensible au contexte modélisant le contexte source par concaténation de trois phrases précédentes. Et finalement, un modèle contrastif, d'architecture Transformer classique sans modélisation de contexte a été proposé. Le travail comprend également une analyse humaine approfondie présentée par des traducteurs locuteurs natifs du chinois.

L'évaluation de la traduction automatique du caractère au document : un état de l'art. In Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), pages 143–159, Paris, France. ATALA. 2023.
Mariam Nakhlé.
→ ***https://aclanthology.org/2023.jeptalnrecital-rjc.10/***
Ce travail propose une vue d'ensemble sur l'état de l'art de l'évaluation de la traduction automatique. Il expose les approches d'évaluation humaine, les méthodes d'évaluation automatiques, tout en différenciant entre les familles d'approches - métriques superficielles et apprises - avec une attention particulière à l'évaluation au niveau du document. Une partie est consacrée à la méta-évaluation des métriques.

Lingua Custodia's Participation at the WMT 2022 Word-Level Auto-completion Shared Task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1170–1175, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. 2022.
Melissa Ailem, Jingshu Liu, Jean-gabriel Barthelemy, and Raheel Qader.
→ ***https://aclanthology.org/2022.wmt-1.118/***
Ce travail présente la participation de Lingua Custodia à la tâche commune de Complétion Automatique au Niveau des Mots de la conférence WMT 2022. Cette tâche vise le développement des systèmes capables de compléter une traduction après l'insertion de quelques caractères par un utilisateur. Les paires de langues étudiées sont : allemand-anglais et anglais-allemand. La méthode proposée traite la séquence introduite par l'utilisateur comme une contrainte, encourageant le modèle à prédire un mot qui commence par une telle séquence. De plus, la méthode utilise une stratégie d'optimisation jointe pour prendre en compte les différents types de contexte de traduction.

Lingua Custodia's Participation at the WMT 2021 Machine Translation Using Terminologies Shared Task. In Proceedings of the Sixth Conference on Machine Translation, pages 799–803, Online. Association for Computational Linguistics. 2021.

Melissa Ailem, Jingshu Liu, and Raheel Qader.

→ ***https://aclanthology.org/2021.wmt-1.78/***

Ce travail présente la participation de Lingua Custodia à la tâche commune de Contrôle de Terminologie de la conférence annuelle WMT 2021. La méthode proposée est basée sur le travail publié précédemment. Les résultats empiriques montrent que dans les paires de langues étudiées (notamment anglais-français, anglais-russe et anglais-chinois) la méthode satisfait la majorité de contraintes terminologiques, tout en gardant une qualité de traduction élevée.

Encouraging Neural Machine Translation to Satisfy Terminology Constraints. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1450–1455, Online. Association for Computational Linguistics. 2021.

Melissa Ailem, Jingshu Liu, and Raheel Qader.

→ ***https://aclanthology.org/2021.findings-acl.125/***

Ce travail présente une nouvelle approche pour forcer les modèles de traduction neuronale à respecter les contraintes terminologiques provenant d'un glossaire. Cette méthode agit au moment de l'entraînement et consiste en ajout de données d'entraînement contenant des contraintes terminologiques entourés de tokens spéciaux. Le terme source est masqué pour renforcer le comportement de copie du terme souhaité dans la traduction générée. Cette méthode a l'avantage de ne pas ajouter de temps de latence à l'inférence.