Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Améliorer la traduction au niveau du document grâce au sur-échantillonnage négatif et au masquage ciblé

Gaëtan Caillaut[1] Mariam Nakhlé[1,2] Jingshu Liu[1] Raheel Qader[1]
(1) Lingua Custodia, Paris, France (2) Université Grenoble Alpes, CNRS, Grenoble INP, LIG, France
firstname.lastname@linguacustodia.com,
firstname.lastname@univ-grenoble-alpes.fr

RÉSUMÉ

Ces travaux visent à améliorer les capacités des systèmes de traduction automatique à tenir compte du contexte dans lequel se trouve la phrase source, et donc, ultimement, à améliorer les performances globales des systèmes de traduction automatique. L'approche que nous proposons repose uniquement sur les données et la manière dont elles sont fournies au modèle durant l'entraînement et est complètement agnostique de l'architecture du modèle. Nous montrons que les performances des modèles de traduction, sur la paire en-fr, peuvent être améliorées simplement en fournissant des données plus pertinentes vis-à-vis de la tâche cible, et ce sans modifier ni complexifier les architectures existantes, en particulier l'architecture Transformer couramment utilisée par les systèmes de TAL modernes. Pour ce faire, nous présentons deux stratégies d'augmentation de données (sur-échantillonnage négatif et masquage ciblé) conçues pour inciter le modèle à s'appuyer sur le contexte. Nous montrons, au travers de métriques appropriées, que ces méthodes permettent d'améliorer les performances des systèmes de traduction sans pour autant modifier ni l'architecture du modèle, ni le processus d'entraînement.

ABSTRACT

Improve Context-Aware Machine Translation with Negative Sampling and Focused Masking

This work aims at enhancing the context awareness of machine translation models without requiring modification on their architecture. Instead, we took a data-driven approach and explore different data augmentation strategies. We show that performance, on the en-fr pair, can be improved solely by improving the « relevance » of the train data according to the target task, instead of refining and/or complicating the transformer architecture, commonly used by modern machine translation systems. Hence, we propose two simple data augmentation strategies (Negative Sampling and Focused Masking) crafted in order to encourage the model to look at the context. We show through the use of appropriate test suite, as well as traditional BLEU, that these data augmentation strategies improve context-level machine translation performances without requiring change in the model architecture nor the training pipeline.

MOTS-CLÉS : traduction au niveau du document, traduction automatique, transformer.

KEYWORDS: context-level machine translation, transformer.

1 Introduction

Les moteurs de traduction traditionnels traduisent les phrases d'un même document indépendamment les unes des autres. Il est de notoriété publique que les prédictions effectuées par systèmes modernes

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 295–306.

de traduction automatique neuronale (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014; Luong *et al.*, 2015; Vaswani *et al.*, 2017) sont généralement très satisfaisantes. Cependant, ces systèmes bénéficieraient largement à davantage prendre en considération le contexte du texte en cours de traduction.

Nous cherchons à faire évoluer les systèmes de traduction centrés sur des phrases individuelles vers des systèmes en mesure de considérer un contexte plus large. En effet, la plupart des travaux autour de la traduction automatique se concentrent généralement sur le problème de la traduction de phrases individuelles (Bahdanau *et al.*, 2014). Toutefois, il semble communément admis que les informations contextuelles, souvent présentes dans les phrases précédentes, peuvent s'avérer décisives pour comprendre pleinement le sens d'une phrase, en particulier lorsqu'elle contient des références implicites comme des ellipses ou des deixis (Voita *et al.*, 2019). Plus généralement, il est courant qu'une phrase référence une précédente, notamment via l'usage de pronoms qui devront être traduits en genre et en nombre. Considérons un document contenant les deux phrases suivantes en anglais : « the girls are getting ready for their exam. they are a bit anxious. ». Si l'on traduit ces deux phrases vers le français de manière classique, c'est à dire indépendamment les unes des autres, on obtiendrait la traduction suivante : « les filles se préparent pour leur examen. Ils sont un peu anxieux. ». En revanche, la prise en compte de la première phrase lors de la traduction de la deuxième, permettrait de lever toutes ambiguïtés et d'obtenir une traduction correcte : « les filles se préparent pour leur examen. Elles sont un peu anxieuses. ». Bien que simple, cet exemple montre qu'il est véritablement crucial de tenir compte du contexte si l'on veut être en mesure de traduire un texte convenablement.

Par ailleurs, tenir compte du contexte ne vise pas uniquement à améliorer la qualité des traductions générées par le modèle, mais peut aussi permettre de réduire certains biais, tels que les biais de genre (Stanczak & Augenstein, 2021), puisque le modèle pourra s'appuyer sur le contexte afin de désambiguïser correctement les éléments critiques.

Plusieurs approches permettant d'incorporer ces informations contextuelles au sein de la représentation de la phrase source ont déjà été proposées. Toutefois, plusieurs travaux ont montré que, curieusement, cela n'améliore pas les systèmes de traduction automatique. Pire encore, ces informations contextuelles semblent être considérées comme du bruit (Kim *et al.*, 2019), ce qui, certes, semble améliorer la robustesse des modèles lorsqu'il s'agit de traduire des phrases individuelles, mais n'améliore pas nécessairement la qualité de la traduction lorsqu'il s'agit de traduire un paragraphe complet. Ces approches consistent généralement en une extension de l'architecture Transformer (Vaswani *et al.*, 2017) avec un encodeur dédié aux phrases contextuelles (Zheng *et al.*, 2020; Maruf *et al.*, 2019).

Les travaux présentés ici s'inspirent de ceux de Lupo *et al.* (2022), mais nous avons choisi d'explorer une piste différente. Les auteurs introduisent une évolution de l'architecture Transformer, censée être plus à même de tenir compte des informations contextuelles. Nous pensons que l'architecture originale est tout à fait capable de prendre le contexte en considération, et que le problème provient principalement de la manière dont les données sont fournies au modèle durant l'entraînement. Ainsi, nous proposons dans ces travaux de repenser le jeu de données et l'objectif ciblé durant la phase d'entraînement afin de pousser le modèle à utiliser de manière appropriée la phrase à traduire et (surtout) son contexte. À cet effet, nous proposons les deux stratégies **sur-échantillonnage négatif** (*negative sampling*) et **masquage ciblé** (*focused masking*). La première consiste à fournir au modèle des traductions erronées et à l'entraîner à ne pas les produire. La seconde consiste, durant l'entraînement, à masquer certains mots fortement dépendants du contexte afin de forcer le modèle à extraire cette information du contexte. Nous évaluons nos stratégies à l'aide du traditionnel score BLEU, mais aussi à l'aide de ContraPro (Lopes *et al.*, 2020) et GenPro (Post & Junczys-Dowmunt,

2023), deux méthodes centrées sur l'évaluation de la traduction des pronoms, et donc plus adaptées à notre objectif. Nous montrons que nos stratégies permettent d'améliorer la qualité des traductions lorsque le contexte est présent, tout en ne la dégradant pas lorsque celui-ci est absent.

2 Travaux connexes

La prise en compte du contexte est crucial afin de construire un système de traduction automatique fiable et robuste. Pourtant, il est encore difficile d'entraîner efficacement un modèle sur ce problème particulier. L'une des raison étant la faible disponibilité des ressources contextuelles : la plupart des jeux de données pour la traduction no sont disponibles que sous la forme de phrases parallèles, dépourvues de leurs contextes d'origine (Schwenk *et al.*, 2019a,b; Koehn, 2005). De plus, les systèmes actuels sont limités par la taille des phrases qu'ils sont capable de traiter, bien que ce problème est en passe d'être en partie résolu avec le développement des LLM de dernières génération, permettant de traiter plusieurs milliers de mots (Jiang *et al.*, 2024; AI@Meta, 2024). Toutefois, les statistiques de la compétition WMT23 (Kocmi *et al.*, 2023) montrent que l'adoption des LLM par la communauté de la traduction automatique est encore très faible.

De nombreux travaux dans ce domaine se concentrent sur le problème de la traduction des pronoms (Guillou, 2012; Le Nagard & Koehn, 2010), ces derniers étant très dépendants du contexte par nature. De plus, certaines phrases peuvent être intraduisibles lorsqu'un pronom se s'accorde pas en genre et/ou en nombre dans la langue source, mais s'accorde dans la langue cible. C'est typiquement le cas avec le pronom anglais « they », pouvant se traduire par « ils » ou « elles » en français. Ainsi, Hardmeier & Guillou (2018) montrent que les modèles reposant sur l'architecture Transformer ont tendance à traduire correctement les pronoms non-anaphoriques et ceux coréférents avec un élément de la même phrase. En revanche, ces modèles ne parviennent pas à traduire convenablement les pronoms faisant références à des éléments présents dans les phrases précédentes.

Lupo *et al.* (2022) estiment qu'il est difficile, en traduction automatique, de prendre en compte le contexte, car les informations pertinentes sont très éparses : généralement, seuls un nombre limité de tokens ont un impact sur la phrase courante. Par conséquent, les modèles ont tendance à ignorer le contexte, puisqu'il ne contient finalement que peu de signaux utiles à la traduction. Pourtant, ces signaux sont d'une importance capitale afin de résoudre certaines ambiguïtés linguistiques, comme les ellipses ou les deixis (Voita *et al.*, 2019). Ils proposent de réduire ce problème à l'aide d'un entraînement en trois étapes. La première étape consiste à entraîner un modèle de traduction standard, sans données contextuelles. La seconde étape, appelée `d&r` (*divide and rule*), consiste à réitérer l'entraînement, mais en scindant les phrases en K morceaux. Les $K - 1$ premiers morceaux sont alors utilisés en tant que contexte et le modèle est entraîné à traduire le dernier morceau. De cette manière, le modèle est incité à porter plus d'attention sur le contexte, puisque la dernière portion de phrase a de fortes chances d'être difficile à traduire sans avoir connaissance du début de la phrase. Une dernière étape d'entraînement est requise afin de recadrer le modèle, puisqu'on attend de celui-ci qu'il traduise des phrases complètes, et non pas des fragments de phrases.

Les récentes avancées en Traitement Automatique de la Langue (TAL) ont montré que les grands modèles de langage (LLM) avec architecture décodeur sont en mesure de répondre efficacement à ce problème, puisqu'ils sont capables de traiter de très longues séquences (plusieurs milliers de tokens). Ces grands modèles se sont également montrés surprenant performants sur des tâches de traduction, sans pour autant avoir été entraînés spécifiquement pour (Chowdhery *et al.*, 2022; Scao *et al.*, 2022;

Touvron *et al.*, 2023). Toutefois, la flexibilité offerte par ces modèles requiert d'énormes capacités de calculs, qui ne sont pas nécessairement justifiées si le but visé ne concerne que la traduction. En outre, Raffel *et al.* (2020) ont montré que l'architecture décodeur semble moins adaptée à la traduction automatique que l'architecture encodeur-décodeur. C'est pourquoi nous nous concentrerons sur des modèles encodeur-décodeur de tailles raisonnables dans la suite de ce document.

Les améliorations apportées par la bonne prise en compte du contexte par les systèmes de traduction automatique se révèlent assez difficile à observer à l'aide des métriques traditionnelles, typiquement BLEU (Papineni *et al.*, 2002), car les quelques tokens dépendant du contexte sont noyés par la masse de tokens n'en dépendant pas. C'est pourquoi Lopes *et al.* (2020) introduisent `ContraPro`, un jeu de données dédié à l'évaluation de modèles contextuels. Ce jeu de données est constitué de phrases provenant de OpenSubtitles2018 (Lison *et al.*, 2018) dans lesquelles le genre de certains pronoms ont été remplacés par le genre opposé. Par exemple, certains « il » sont changés en « elle ». Les phrases présentes dans le jeu de données sont sélectionnées afin que le contexte puisse permettre de sélectionner le bon pronom. Les modèles sont alors évalués selon les probabilités qu'ils attribuent aux phrases originales et à leurs versions altérées : la phrase originale doit être considérée comme plus probable que l'autre.

Post & Junczys-Dowmunt (2023) pensent que les modèles doivent être évalués en fonction des phrases qu'ils génèrent réellement, pas seulement selon les probabilités qu'ils donnent à des phrases prédéfinies. En effet, si l'on considère une traduction A et sa version altérée B, `ContraPro` considère le modèle comme étant juste si et seulement si la probabilité $P(A)$ est supérieure à $P(B)$. Or, $P(A)$ peut tout à fait être tellement faible qu'en pratique cette phrase ne puisse pas être générée par le modèle. Dans ce cas, doit-on considérer le modèle comme juste ? Post & Junczys-Dowmunt (2023) cherchent à limiter ce biais en proposant `GenPro`, une méthode d'évaluation basée sur `ContraPro` valorisant un modèle quand il génère effectivement le pronom attendu.

3 Méthode

Les approches existantes cherchent généralement à introduire l'information contextuelle ajoutant un module dédié à l'architecture transformer, couramment utilisée de nos jours pour répondre aux problématiques de TAL. Or, Li *et al.* (2020) ont montré que ces modifications n'aident pas réellement le modèle à mieux prendre en compte le contexte. Comme montré par Kim *et al.* (2019), les modèles tendent à considérer le contexte comme du bruit, et apprennent à ne pas en tenir compte. Ce comportement n'est pas déraisonnable, puisque l'immense majorité des informations nécessaires se trouvent effectivement dans la phrase à traduire, et non dans le contexte passé. Par conséquent, le modèle est naturellement incité, lors du processus d'entraînement, à se concentrer sur la phrase en cours de traduction uniquement, et à ignorer le reste. Cela permet au modèle de converger rapidement vers une solution acceptable. Toutefois, nous supposons qu'une fois ce niveau atteint, il est difficile, pour le modèle, de *faire marche arrière* afin de sortir de cet état où le contexte est ignoré.

Nous pensons que l'architecture Transformer originale est naturellement capable de sélectionner et d'utiliser les bonnes informations contextuelles. Le mécanisme d'attention est suffisamment puissant pour extraire les signaux pertinents pour chaque token. Selon nous, si le modèle traite les phrases contextuelles comme du bruit, c'est parce qu'il a été entraîné de façon sous-optimale, et que le modèle doit être davantage incité à regarder du côté du contexte, au lieu d'apprendre à l'ignorer.

Lupo *et al.* (2022) proposent une méthode dont afin de forcer le modèle à extraire de l'information du côté du contexte. Toutefois, leur approche en trois étapes nous semble contraignante et perfectible. En effet, le principe de leur méthode est de fournir au modèle des phrases dont la structure est *cassée* de manière à ce qu'elles ne puissent pas être traduites sans leur contexte. Le modèle est par la suite ré-entraîné sur des phrases non altérées, tout en figeant les paramètres du modèle chargés d'extraire l'information contextuelle. Le modèle résultant de ce processus est donc censé extraire de l'information contextuelle, mais il est nécessairement sous-optimal puisqu'il n'a été entraîné que sur des contextes synthétiques et incomplets.

Nous nous inspirons de ces travaux et proposons une approche en deux étapes consistant à entraîner le modèle sur un large jeu de données de phrases parallèles, puis à adapter ce modèle sur un jeu de données de taille réduite que l'on augmente avec des données synthétiques. Nous considérons le même jeu de données utilisé par Lupo *et al.* (2022) , c'est-à-dire WMT14 [1] pour entraîner le modèle de base et IWSLT17 [2] pour affiner le modèle sur des données contextuelles. Dans ces travaux, nous n'étudions que le cas de la traduction de l'anglais au français (EN-FR). Le contexte et la phrase à traduire sont encodés de la manière suivante : `<s> <ctx> CONTEXT TOKENS </ctx> SOURCE TOKENS </s>`.

Nous pensons que le comportement du modèle vis-à-vis du contexte doit être guidé par les données vues durant la phase d'entraînement, c'est pourquoi nous explorons différentes manières de pousser le modèle à extraire de l'information du contexte, lorsque cela s'avère nécessaire.

3.1 Sur-échantillonnage négatif

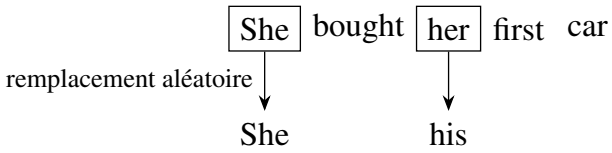

FIGURE 1 – Augmentation du jeu de données avec des échantillons négatifs. La phrase source est scannée pour détecter des mots genrés, qui sont remplacé selon une certaine probabilité.

L'objectif du sur-échantillonnage négatif (*negative sampling*) est de fournir au modèle davantage d'exemples négatifs afin de l'entraîner à ne pas les reproduire. Pour chaque phrase cible, nous générons une version altérée dans laquelle les mots *critiques* (pronoms et déterminants) sont remplacés par leurs homologues du genre opposé, comme illustré en Figure 1. Nous avons sélectionné les mots à remplacer en effectuant une simple recherche basée sur les caractères au lieu de reposer sur des méthodes plus raffinées, par exemple en s'appuyant sur l'analyse morphosyntaxique des mots. Ensuite, seul un sous-ensemble aléatoire des mots ainsi sélectionnés sont remplacés. La liste des déterminants et pronoms que nous considérons est la suivante : il/elle, ils/elles, le/la, un/une, mon/ma, ton/ta, son/sa, ce/cette, tous/toutes, quel/quelle et lequel/laquelle.

Le modèle est entraîné de façon à minimiser la fonction de perte associée au sur-échantillonnage (*negative sampling loss*) $\mathcal{L}_{ns}$, correspondant à la moyenne des probabilités des tokens remplacés, telles que prédites par le modèle (les *logits*). Puisque les tokens remplacés sont incorrects, leurs

1. http://www.statmt.org/wmt14/translation-task.html
2. https://sites.google.com/site/iwsltevaluation2017/data-provided

probabilités doivent être faibles, nous cherchons donc à minimiser $\mathcal{L}_{ns}$. La fonction objectif finale est donnée par $\mathcal{L} = \mathcal{L}_{translation} + \mathcal{L}_{ns}$ où $\mathcal{L}_{translation}$ est l'entropie-croisée, couramment utilisée pour entraîner les modèles de traduction.

3.2 Masquage ciblé

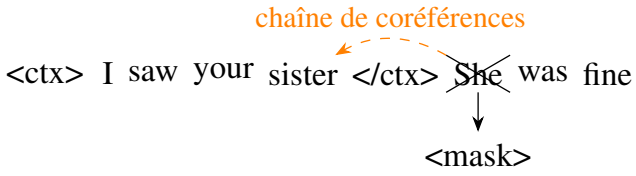

FIGURE 2 – Masquage des mots coréférents. Les éléments de la chaine de coréférence [she, sister] présents dans la source sont masqués selon une certaine probabilité.

L'idée du masquage ciblé est relativement similaire à celle du sur-échantillonnage négatif, mais au lieu d'altérer la cible, nous modifions la phrase de façon à masquer certains mots, censés être très dépendants du contexte. Puisque l'objectif est d'inciter le modèle à extraire de l'information du contexte, les tokens du contexte ne sont jamais masqués, seuls les tokens de la phrase à traduire le sont. Nous avons exploré deux stratégies, l'une consiste à masquer les mots coréférents comme montré en Figure 2, l'autre à masquer les mots apparaissant à la fois dans le contexte et la source. Nous avons utilisé `coreferee` [3] et `spaCy` [4] pour extraire les chaînes de coréférence. Par exemple, dans la phrase « J'ai vu ta sœur. Elle allait bien. », les mots « sœur » et « elle » sont coréférents. Nous supposons que masquer le pronom « elle » forcera le modèle à analyser le contexte pour en déduire le bon pronom. De même que pour la stratégie précédente, seulement une portion aléatoire d'éléments des chaînes de coréférence sont masqués, et seulement s'ils apparaissent dans la source afin de ne pas décourager le modèle à puiser de l'information dans le contexte.

3.3 Divide and rule

L'objectif de la stratégie divide-and-rule proposée par Lupo *et al.* (2022) est de forcer le modèle à observer le contexte en divisant la phrase initiale en morceaux de tailles similaires, et de réutiliser les premiers morceaux en guise de contexte. De cette manière, le modèle est entraîné sur des phrases complètement déstructurées, de sorte qu'il soit nécessaire de prendre en compte le contexte pour traduire la phrase. Dans les travaux originaux, les auteurs entraînent un encodeur dédié au contexte en plus de celui dédié à la phrase source à traduire, puis fusionnent les représentations de ces deux encodeurs avant de générer la traduction par le décodeur.

Dans le cadre de nos travaux, nous pensons que l'encodeur dédié au contexte est superflu, mais la stratégie divide-and-rule nous semble toutefois pertinente, notamment car elle permet de créer de l'information contextuelle à partir de phrases isolées. C'est pourquoi nous avons généré des données synthétiques en divisant chaque phrase source en trois morceaux x, y et z de tailles équivalentes. Ainsi, chaque phrase nous permet de générer les trois paires contexte/source suivantes :

3. https://github.com/richardpaulhudson/coreferee
4. https://spacy.io/

— <s> <ctx> </ctx> x y z </s>
— <s> <ctx> x </ctx> y z </s>
— <s> <ctx> x y </ctx> z </s>

Les phrases cibles sont générées de la même manière, ce qui s'avère être l'un des défauts majeurs de cette stratégie, puisque les mots présents dans la source et la cible n'apparaissent pas nécessairement dans le même ordre.

4 Expérimentations

Plusieurs modèles ont été entraînés afin de mesurer l'impact des différentes stratégies présentées dans ces travaux : sur-échantillonnage négatif, masquage ciblé et *divide and rule*. Les modèles ont été entraînés dans un premier temps sur la direction EN-FR du corpus WMT14 pendant une *epoch*, puis affinés sur la direction EN-FR du corpus IWSLT17 pour 15 *epochs*. Nos modèles reprennent l'architecture encodeur-décodeur standard, avec 6 couches d'encodeurs et 6 couches de décodeurs, 8 têtes d'attentions et la dimension des embeddings est fixée à 512. Les modèles sont entraînés avec l'optimiseur AdamW, un taux d'apprentissage de $5e^{-5}$ et la taille du batch est fixée à 32 phrases.

Les scores BLEU sont calculés sur l'ensemble de test du corpus IWSLT17 à l'aide de SacreBLEU (Post, 2018). Nous calculons également les scores ContraPro (Lopes *et al.*, 2020) et GenPro (Post & Junczys-Dowmunt, 2023), puisqu'ils sont censés mieux capturer la capacité d'un modèle à tenir compte des informations contextuelles. Ces deux dernières métriques mesurent essentiellement la capacité du modèle à traduire correctement les pronoms, qui sont des éléments très dépendants du contexte.

Chaque modèle a été entraîné avec et sans contexte afin d'observer si l'information portée par le contexte permet réellement d'améliorer la qualité des traductions. Dans cette expérience, le contexte correspond à la phrase précédant la phrase source. Les scores de tous les modèles évalués sont donnés en Tableaux 1 et 2. Les modèles dont le nom portent la mention « coref » et « samewords » ont été entraînés avec, respectivement, les stratégies consistant à masquer les mots coréférents et les mots partagés entre le contexte et la source. Pour chaque stratégie, deux expériences ont été faites avec des probabilités de masquage de 0,5 et 0,8. Le modèle « d&r » a été entraîné à l'aide de l'approche *divide and rule*.

Notre première observation est que les modèles contextuels (auxquels on fournit la phrase précédente) sont globalement plus performants que les autres modèles, comme l'indiquent les scores en Tableau 1, supérieurs à ceux en Tableau 2. Les différences entre chaque scores sont statistiquement significatives ($p < 0.05$), il semble donc qu'il soit bénéfique d'augmenter l'entrée du modèle, la phrase à traduire, avec son contexte. Cela est d'autant plus perceptible sur les scores ContraPro et GenPro, là où les fluctuations sur les scores BLEU sont plus limitées. Cette dernière observation corrobore les conclusions d'autres travaux, notamment ceux ayant motivé la création de ContraPro et GenPro, selon lesquelles BLEU ne serait pas adapté pour l'évaluation de modèles contextuels (Nakhlé, 2023; Jin *et al.*, 2023). De plus, ces expériences montrent qu'il est effectivement nécessaire d'inclure des informations contextuelles afin de traduire correctement les pronoms. Il en va certainement de même pour d'autres éléments fortement dépendants du contexte, mais ContraPro et GenPro ne nous permettent de nous prononcer que sur les pronoms.

Nous remarquons également que les différentes stratégies de masquage, de même que la méthode

	affiné avec SN			*affiné sans SN*		
Modèle	**BLEU**	**ContraPro**	**GenPro**	**BLEU**	**ContraPro**	**GenPro**
Pré-entraîné avec SN						
no masking	39,0	0,84	46,4	39,3	0,83	52,1
d&r	39,3	0,83	43,5	39,9	0,82	51,7
coref50	38,9	0,84	46,3	39,4	0,83	52,2
coref80	38,9	0,84	46,5	39,3	0,83	52,3
samewords50	39,0	0,83	47,2	39,7	0,82	52,3
samewords80	39,0	0,83	47,4	39,7	0,82	52,4
Pré-entraîné sans SN						
no masking	39,0	0,83	43,3	39,8	0,79	46,8
d&r	39,3	0,82	39,7	40,1	0,78	44,5
coref50	39,0	0,83	43,3	39,8	0,79	46,8
coref80	39,0	0,83	43,3	39,7	0,79	46,8
samewords50	39,4	0,82	41,6	39,9	0,78	43,8
samewords80	39,9	0,78	43,8	40,0	0,78	43,8
Lupo et al. *(2022)*						
K1				41,93	0,84	
K1 d&r				41,78	0,79	

TABLE 1 – Performances sur la traduction de phrases accompagnées de leurs contextes. SN signifie « sur-échantillonnage négatif ». Tous les scores sont à maximiser. Les différences entre les modèles affinés avec SN (gauche) et sans (droite) sont significatives ($p < 0.05$). Les différences entre les modèles pré-entraînés avec (haut) et sans SN (bas) ne sont significatives ($p < 0.05$) que pour les scores *Pro.

divide and rule, n'ont finalement que très peu d'impact sur les performances finales. Pire, le sur-échantillonnage négatif semble dégrader légèrement les scores BLEU des modèles contextuels. Cependant, cette stratégie semble tout de même améliorer les scores ContraPro et GenPro. C'est en effet la seule méthode permettant d'améliorer systématiquement le score ContraPro, indiquant que cette stratégie permet bel et bien au modèle d'extraire de meilleures informations contextuelles. Toutefois, cette stratégie semble aussi dégrader les scores GenPro lorsqu'elle est appliquée pour affiner le modèle. Cela signifie que le sur-échantillonnage négatif permet au modèle de mieux identifier les erreurs sur les pronoms, sans pour autant l'inciter à générer les bons. Cela reste tout de même à considérer avec un peu plus de hauteur, puisque GenPro repose sur une méthode heuristique pour aligner les pronoms entre les phrases sources et cibles. Il est possible que de nombreux faux négatifs soient comptabilisés.

Enfin, nous comparons les résultats obtenus par nos modèles avec ceux publiés par Lupo *et al.* (2022). Nos expérimentations sont comparables à celles effectuées sur le modèle intitulé *K1 Low Res*, c'est pourquoi nous ne reprenons que ces résultats en Tableau 1. Les scores BLEU des modèles *K1* sont plus élevés que les nôtres, nous supposons que cela provient de paramétrages plus optimaux, soit au niveau des hyperparamètres des modèles, soit au niveau de l'entraînement (taille du batch, durée de l'entraînement, ...). Nous nous intéressons davantage aux gains apportées par l'approche d&r sur les scores ContraPro, et nous observons que nos modèles parviennent à concurrencer le modèle *K1 d&r*, même en l'absence d'encodeur dédié à la prise en charge du contexte. Cela semble valider notre

	affiné avec SN			*affiné sans SN*		
Modèle	**BLEU**	**ContraPro**	**GenPro**	**BLEU**	**ContraPro**	**GenPro**
Pré-entraîné avec SN						
no masking	37,8	0,78	33,6	38,6	0,78	40,5
d&r	39,3	0,78	31,6	39,7	0,77	40,8
coref50	38,2	0,78	33,6	38,5	0,78	40,4
coref80	38,0	0,78	33,7	38,7	0,78	40,5
samewords50	38,4	0,78	34,6	39,1	0,77	40,6
samewords80	38,5	0,78	34,7	39,0	0,77	40,6
Pré-entraîné sans SN						
no masking	35,2	0,77	30,9	36,0	0,77	39,7
d&r	39,5	0,78	31,1	40,1	0,76	41,4
coref50	35,2	0,77	31,0	35,9	0,77	39,7
coref80	35,2	0,77	30,9	35,9	0,77	39,6
samewords50	35,9	0,78	30,5	36,8	0,76	39,2
samewords80	36,7	0,76	39,2	36,8	0,76	39,2

TABLE 2 – Performances sur la traduction de phrases individuelles. SN signifie « sur-échantillonnage négatif ». Tous les scores sont à maximiser. Les différences entre les modèles affinés avec SN (gauche) et sans (droite) ne sont significatives ($p < 0.05$) que pour GenPro. Les différences entre les modèles pré-entraînés avec (haut) et sans SN (bas) ne sont significatives ($p < 0.05$) que pour les scores BLEU et ContraPro.

hypothèse initiale (il n'est pas nécessaire de complexifier l'architecture du modèle pour améliorer la prise en charge du contexte).

En résumé, nos expériences montrent que la meilleure stratégie semble être d'augmenter le jeu de données de pré-entraînement avec des exemples négatifs (sur-échantillonnage négatif), puisque les modèles entraînés de cette manière tendent à trouver le meilleur équilibre entre la qualité de la traduction (BLEU) et prise en compte du contexte (ContraPro et GenPro). Les autres stratégies évaluées ne semblent pas apporter de réels gains.

5 Conclusion

Nous proposons, des méthodes permettant d'introduire des éléments d'information contextuelles au sein du processus de traduction. Cette étude se concentre sur la prise en compte des informations présentes dans la phrase précédente, mais nous envisageons d'étendre le contexte à davantage de phrases. Nous proposons et évaluons différentes stratégies d'entraînement et d'augmentation de données destinées à améliorer la prise en compte du contexte par modèles de traduction automatique. En effet, comme le montrent nos expériences, inclure le contexte, même s'il ne s'agit que de la phrase précédente, permet d'augmenter les performances du système de traduction. L'une de nos stratégies consiste à augmenter le jeu de données initiales afin d'ajouter des exemples négatifs. Les résultats montrent que cette stratégie permet effectivement d'améliorer la prise en charge du contexte par le modèle, ce qui se matérialise par des scores ContraPro et GenPro plus élevés, deux métriques conçues

pour évaluer les modèles contextuels. Toutefois, cette stratégie semble dégrader la qualité générale de la traduction, telle que reportée par le score BLEU. Nous montrons qu'il est possible d'éviter ce phénomène en appliquant le sur-échantillonnage négatif lors du pré-entraînement uniquement. Notre approche se limite à introduire un signal négatif sur le genre et le nombre de certains pronoms et déterminants. L'une des pistes à explorer serait d'envisager d'autres méthodes de corruptions, par exemple remplacer un mot par son antonyme, ou encore modifier le temps de certains verbes.

La seconde stratégie que nous proposons consister à sélectionner des termes dépendants du contexte et d'en masquer une proportion aléatoire. Nous avons exploré deux pistes : masquer les mots apparaissant à la fois dans le contexte et la source et masquer certains éléments des chaînes de coréférences. Nos expériences montrent que cette approche n'apporte aucune plus-value. Cependant, cela ne suffit pas pour rejeter complètement cette stratégie, puisqu'il est possible qu'elle porte ses fruits avec des choix plus avisés de mots à masquer.

Références

AI@META (2024). Llama 3 model card.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

CHOWDHERY A., NARANG S., DEVLIN J., BOSMA M., MISHRA G., ROBERTS A., BARHAM P., CHUNG H. W., SUTTON C., GEHRMANN S. *et al.* (2022). Palm : Scaling language modeling with pathways. *arXiv preprint arXiv :2204.02311*.

GUILLOU L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1–10.

HARDMEIER C. & GUILLOU L. (2018). Pronoun translation in english-french machine translation : An analysis of error types. *arXiv preprint arXiv :1808.10196*.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., HANNA E. B., BRESSAND F. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv :2401.04088*.

JIN L., HE J., MAY J. & MA X. (2023). Challenges in context-aware neural machine translation. *arXiv preprint arXiv :2305.13751*.

KIM Y., TRAN D. T. & NEY H. (2019). When and why is document-level context useful in neural machine translation ? *arXiv preprint arXiv :1910.00294*.

KOCMI T., AVRAMIDIS E., BAWDEN R., BOJAR O., DVORKOVICH A., FEDERMANN C., FISHEL M., FREITAG M., GOWDA T., GRUNDKIEWICZ R. *et al.* (2023). Findings of the 2023 conference on machine translation (wmt23) : Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, p. 1–42.

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x : papers*, p. 79–86.

LE NAGARD R. & KOEHN P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, p. 252–261 : Association for Computational Linguistics.

LI B., LIU H., WANG Z., JIANG Y., XIAO T., ZHU J., LIU T. & LI C. (2020). Does multi-encoder help ? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, p. 3512–3518, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.322.

LISON P., TIEDEMANN J. & KOUYLEKOV M. (2018). Opensubtitles2018 : Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* : European Language Resources Association (ELRA).

LOPES A. V., FARAJIAN M. A., BAWDEN R., ZHANG M. & MARTINS A. F. (2020). Document-level neural mt : A systematic comparison. In *22nd Annual Conference of the European Association for Machine Translation*, p. 225–234.

LUONG M.-T., PHAM H. & MANNING C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv :1508.04025*.

LUPO L., DINARELLI M. & BESACIER L. (2022). Divide and rule : Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4557–4572.

MARUF S., MARTINS A. F. & HAFFARI G. (2019). Selective attention for context-aware neural machine translation. *arXiv preprint arXiv :1903.08788*.

NAKHLÉ M. (2023). 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)*, p. 143–159.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

POST M. & JUNCZYS-DOWMUNT M. (2023). Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv :2304.12959*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

SCHWENK H., CHAUDHARY V., SUN S., GONG H. & GUZMÁN F. (2019a). Wikimatrix : Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv :1907.05791*.

SCHWENK H., WENZEK G., EDUNOV S., GRAVE E. & JOULIN A. (2019b). Ccmatrix : Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv :1911.04944*.

STANCZAK K. & AUGENSTEIN I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv :2112.14168*.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, **27**.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VOITA E., SENNRICH R. & TITOV I. (2019). When a good translation is wrong in context : Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. *arXiv preprint arXiv :1905.05979*.

ZHENG Z., YUE X., HUANG S., CHEN J. & BIRCH A. (2020). Towards making the most of context in neural machine translation. *arXiv preprint arXiv :2002.07982*.