L'évaluation de la traduction automatique du caractère au document : un état de l'art

Mariam Nakhlé [1, 2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
(2) Lingua Custodia, France
mariam.nakhle@univ-grenoble-alpes.fr,
mariam.nakhle@linguacustodia.com

RÉSUMÉ

Ces dernières années l'évaluation de la traduction automatique, qu'elle soit humaine ou automatique, a rencontré des difficultés. Face aux importantes avancées en matière de traduction automatique neuronale, l'évaluation s'est montrée peu fiable. De nombreuses nouvelles approches ont été proposées pour améliorer les protocoles d'évaluation. L'objectif de ce travail est de proposer une vue d'ensemble sur l'état global de l'évaluation de la Traduction Automatique (TA). Nous commencerons par exposer les approches d'évaluation humaine, ensuite nous présenterons les méthodes d'évaluation automatiques tout en différenciant entre les familles d'approches (métriques superficielles et apprises) et nous prêterons une attention particulière à l'évaluation au niveau du document qui prend compte du contexte. Pour terminer, nous nous concentrerons sur la méta-évaluation des méthodes.

ABSTRACT

The evaluation of machine translation from character to document : state of the art

The human and automatic evaluation of machine translation have undergone great disruption in recent years. In the face of significant advances in neural machine translation, evaluation has shown to be unreliable. Multiple new approaches have been proposed to improve evaluation protocols. The aim of this work is to provide an overview of the global state of Machine Translation evaluation. We will start by outlining the approaches and methods used in human evaluation, next, we will present automatic evaluation methods while distinguishing between families of approaches (string-based and learned metrics) and we will pay particular attention to context-sensitive document-level evaluation. Finally, we will focus on the meta-evaluation of the methods.

MOTS-CLÉS : évaluation de la traduction automatique, traduction automatique, traduction automatique au niveau du document.

KEYWORDS: machine translation evaluation, machine translation, document-level machine translation.

1 Introduction

Nous présentons un état de l'art de l'évaluation de la traduction automatique (TA), en particulier dans le contexte de l'arrivée des approches neuronales dans la TA. La traduction automatique a connu de grands progrès ces dernières années, notamment suite à l'introduction du modèle Transformer

*. Institute of Engineering Univ. Grenoble Alpes

(Vaswani *et al.*, 2017), ce qui a permis une grande amélioration de la qualité des moteurs de traduction. Face à cette nouvelle génération de modèles de traduction, l'évaluation de la TA manque de robustesse (Bojar *et al.*, 2018).

Le score BLEU (Papineni *et al.*, 2002), la méthode d'évaluation automatique la plus connue, a été conçue pendant l'ère de la traduction statistique pour un usage avec de multiples phrases de référence. Avec l'arrivée des approches neuronales, ce score (de la même manière que les approches similaires) s'est montré peu fiable face aux nouveaux systèmes. Le score BLEU était controversé même avant l'arrivée de la TA neuronale et cette nouvelle génération de systèmes n'a fait qu'aggraver les critiques de la métrique. Les nouveaux systèmes ont un comportement - forces et faiblesses - très différent des systèmes statistiques du début du siècle. En effet, ils sont capables de formuler des phrases de plus en plus naturelles, employer des synonymes divers et des constructions syntaxiques variées, c'est pourquoi une simple comparaison de la traduction automatique avec la traduction de référence (comme le fait BLEU) n'est plus suffisante. Le développement des systèmes de traduction capables de traduire des segments plus longs que la phrase requiert une évaluation qui va au-delà de la phrase également. De nouvelles approches sont donc nécessaires.

La conférence annuelle sur la TA, appelée WMT, essaie de fournir une réponse à ce besoin en organisant chaque année une tâche commune (*shared-task*) sur les métriques automatiques (Freitag *et al.*, 2022, 2021b; Mathur *et al.*, 2020b). Ainsi, de nombreuses nouvelles métriques ont émergé. Face à cette pluralité de nouvelles approches, la communauté fait face un nouveau défi : quelle est la plus performante et comment mesurer cette performance ? Plusieurs travaux se sont proposés de comparer leur performance et fiabilité (Mathur *et al.*, 2020a; Kocmi *et al.*, 2021; Freitag *et al.*, 2022). Selon les résultats les plus récents, la méthode qui semble montrer les meilleures performances est le score COMET (Rei *et al.*, 2020), qui est en train de devenir le nouveau score standard.

Une autre famille d'approches propose de répondre au problème du contexte dans l'évaluation. Vu qu'il a été démontré que des erreurs de traduction passent inaperçues sans le contexte, ces méthodes ciblent l'évaluation sur les phénomènes linguistiques, préalablement identifiés comme problématiques, et dont la traduction correcte dépend du contexte. L'évaluation se fait en utilisant des ensembles de test spécialement conçus pour la tâche. Ces ensembles s'appellent des *test-suites* et ils servent à étudier la performance des modèles de traduction sur le phénomène linguistique ciblé. Généralement un taux de traductions correctes est calculé, qui sert de mesure d'évaluation complémentaire, surtout pour l'évaluation des systèmes de TA sensibles au contexte. Il s'agit par exemple de la traduction des pronoms anaphoriques ou des mots polysémiques (Vojtěchová *et al.*, 2019; Rios *et al.*, 2018; Müller *et al.*, 2018).

Devant cette pluralité d'approches d'évaluation, il peut être difficile de choisir la plus adaptée. Plusieurs méthodes ont été essayées pendant des années et de nouvelles approches émergent tous les ans. Comme une évaluation solide est la base du développement, nous proposons d'offrir une vue globale sur l'état de l'art du domaine, en présentant les différentes approches étudiées et une réflexion sur les zones d'améliorations. Nous organisons l'article en trois parties : dans la section 2 nous décrivons les méthodes d'évaluation humaine, dans la section 3 nous présentons les méthodes d'évaluation automatique en les classant selon le type d'approche. Nous faisons la distinction entre les méthodes qui évaluent les phrases isolées (sous-section 3.1) et les approches qui évaluent les phrases dans leur contexte (sous-section 3.2). Dans la sous-section 3.3, nous montrons comment évaluer les métriques automatiques et finalement dans la section 4 nous présentons une discussion sur le sujet.

2 Évaluation humaine

Déterminer la qualité de la traduction est une tâche compliquée qui n'a pas de réponse unique. Contrairement à certaines tâches du traitement automatique des langues où l'évaluation est plutôt simple à effectuer (comme par exemple la reconnaissance vocale), il est difficile de déterminer la qualité de la traduction automatique et cela même pour les humains. Plusieurs traductions d'une seule phrase peuvent être correctes et pour les évaluateurs (surtout les non-professionnels) il est compliqué de distinguer la meilleure. Plusieurs aspects jouent un rôle sur le choix : objectif, style, contexte, domaine, etc. du texte à traduire. Pourtant un protocole d'évaluation humaine clair est crucial pour l'obtention de résultats fiables et par conséquent pour le développement de la TA. Dans la recherche du meilleur protocole d'évaluation, plusieurs approches ont été essayées et étudiées :

- **Mesures d'effort de post-édition** : Ces approches mesurent l'effort nécessaire pour post-éditer une traduction automatique jusqu'à ce qu'elle soit correcte. Typiquement trois dimensions sont considérées : le temps passé, l'effort technique (nombre de frappes nécessaires) et l'effort cognitif (Koponen, 2016). De nos jours, cette piste n'est plus beaucoup étudiée.
- **Évaluation directe** : Il s'agit de l'attribution d'un score d'appréciation, typiquement sur l'échelle de Likert allant de 1 à 5 (où 1=le pire score et 5=le score parfait) (Bojar *et al.*, 2016), ou bien sur une échelle continue (Graham *et al.*, 2013). Ce score est unique pour toute la phrase, il a comme objectif d'englober tous les aspects pertinents (qui peuvent varier selon le cas d'usage) tels que la fluidité, la précision, l'exactitude terminologique, le style, etc. C'est la métrique utilisée dans la conférence WMT à partir de 2017 jusqu'à 2021. Le désavantage de cette technique est que les valeurs d'accord inter-annotateurs sont souvent basses. Ce type d'évaluation est appelée en anglais *direct assessment*.
  - **Scores basés sur la fluidité et l'adéquation** : Cela est une sous-catégorie de l'évaluation directe. Les évaluateurs donnent un score pour la fluidité (*fluency*) de la traduction et un autre pour l'adéquation (*adequacy*). Le score attribué l'est généralement sur une échelle de Likert de 1 à 5. Cette approche a été utilisée à WMT en 2006 et 2007.
- **Classement comparatif** : Il s'agit de comparer deux ou plusieurs sorties candidates. Les évaluateurs classent les traductions de la meilleure à la pire (classer deux sorties comme ayant le même niveau est d'habitude possible). Cette métrique a été utilisée à WMT de 2007 à 2016. Elle a comme désavantage la croissance du nombre de phrases à lire, surtout en comparant plusieurs systèmes. Un autre désavantage de cette approche est que le score reste relatif et ne peut pas être interprété en valeur absolue (Bojar *et al.*, 2016). Ce type d'évaluation est appelé en anglais *relative ranking*.
- **Annotation de type *Multidimensional Quality Metrics* (MQM)** : Introduit par Lommel *et al.* (2014) et popularisé par Freitag *et al.* (2021a), cette approche consiste à annoter les erreurs dans le texte traduit. Chaque erreur est liée à un poids selon sa gravité. Le score final est calculé sur la base des erreurs repérées. L'avantage de cette approche est qu'elle permet de définir un ensemble d'erreurs ainsi que leurs poids selon le cas d'usage, ce qui rend l'évaluation plus spécifique puisqu'elle prend en compte les critères concrets du domaine et de l'application. Elle était utilisée à WMT en 2021 et 2022.

L'évaluation humaine est considérée la plus fiable. Cependant, durant ces dernières années, il s'est avéré que rien que le fait d'effectuer une évaluation avec des évaluateurs humains n'est pas la garantie d'une évaluation de qualité. En effet, elle est souvent subjective et de mauvais protocoles d'évaluation peuvent la rendre peu fiable. L'annonce de la parité de la traduction automatique et humaine de

(Hassan *et al.*, 2018) a généré un grand débat concernant l'évaluation humaine de la TA. De nombreux auteurs ont répondu en démontrant que cette annonce n'était qu'une fausse conclusion liée aux défauts du protocole d'évaluation humaine (Läubli *et al.*, 2018; Toral *et al.*, 2018; Graham *et al.*, 2019). Lors de l'évaluation humaine, les évaluateurs n'avaient pas accès au contexte des phrases à évaluer, ce qui a rendu certaines erreurs impossibles à détecter (Läubli *et al.*, 2018; Freitag *et al.*, 2021a). Également, certaines phrases utilisées comme phrases sources dans le jeu de test étaient elles-mêmes des traductions, ce qui rend la traduction plus facile dû à l'effet « translationese » (Graham *et al.*, 2019). Dans le contexte de nouvelles approches neuronales où les modèles devenaient de plus en plus robustes, cette évaluation n'était plus suffisante et les organisateurs de WMT ont annoncé qu'il faudrait trouver de nouveaux moyens pour évaluer (Bojar *et al.*, 2018). À partir de 2021, la conférence a changé de protocole d'évaluation et a adopté l'approche d'annotation MQM telle que proposée dans Freitag *et al.* (2021a).

L'idée principale derrière cette approche est qu'implicitement, l'attribution de score par un évaluateur se fait à travers l'analyse des erreurs dans la traduction. C'est pourquoi les auteurs proposent de rendre explicite cette analyse lors de l'évaluation grâce à l'annotation MQM. Ils ont rendu publiques des annotations MQM des jeux de test issus de la conférence WMT pour permettre de futurs travaux [1]. Certaines implémentations de l'évaluation MQM introduisent des arbres de décision pour guider l'évaluateur, rendant ainsi l'évaluation un processus plus logique et conscient, ce qui permet d'obtenir des résultats plus standardisés et moins subjectifs.

Dans un effort de compréhension de la significativité statistique de l'évaluation humaine, Wei *et al.* (2022) ont analysé un ensemble de 1728 campagnes d'évaluation provenant de l'évaluation interne chez Microsoft (Kocmi *et al.*, 2021) et sont arrivés à la conclusion que dans la majorité des cas, les résultats ne sont pas statistiquement significatifs, c'est-à-dire que l'évaluation ne permet pas de déterminer si un système est plus performant qu'un autre. Ils concluent que la différence entre les modèles étant petite, les échantillons évalués sont insuffisants et un nombre d'échantillons (et donc un budget) beaucoup plus élevé serait nécessaire pour aboutir à une évaluation statistiquement significative. Nous pouvons nous poser la question si une telle analyse statistique sur l'évaluation MQM donnerait des conclusions similaires. Freitag *et al.* (2021b) ont analysé les annotations MQM et ont trouvé que malgré l'accord inter-annotateur plutôt bas, les classements finaux des systèmes restent cohérents parmi les évaluateurs.

Nous voyons alors que l'évaluation humaine a quelques défauts. Ceci est d'autant plus problématique vu que, comme nous allons le présenter dans la section suivante, l'évaluation humaine sert de référence de qualité (*gold standard*) pour évaluer et, selon les cas, entraîner les approches automatiques. Dans la section suivante, nous présentons les approches d'évaluation automatique.

3 Évaluation automatique

Une grande partie de l'évaluation de la TA est faite par des méthodes automatiques, c'est-à-dire qui ne nécessitent pas d'évaluateurs humains pour obtenir un jugement sur la qualité du système de traduction. Ces méthodes sont souvent appelées métriques, pourtant cela n'est pas dans le sens mathématique du terme. Leur avantage principal est leur coût réduit (surtout grâce aux ensembles de tests publics et par rapport à l'évaluation humaine où les évaluateurs sont payés) et leur rapidité. Ces méthodes sont largement utilisées pendant la phase de développement des systèmes. Nous allons

1. github.com/google/wmt-mqm-human-evaluation

diviser les approches en deux catégories : (1) les méthodes qui évaluent la traduction isolée de son contexte et (2) les méthodes au niveau du document, qui cherchent à évaluer la traduction dans son contexte. Il existe un grand nombre de méthodes d'évaluation de la génération de texte. Même si la traduction automatique fait partie des tâches de la génération de texte, dans le cadre de ce travail, nous allons nous contenter de présenter les méthodes les plus utilisées en évaluation de la traduction automatique.

3.1 Évaluation automatique au niveau de la phrase

La majorité des métriques automatiques existantes opèrent au niveau de la phrase. C'est-à-dire que le contexte (les phrases précédentes ou suivantes) n'est pas pris en compte lors de l'évaluation de la traduction. Nous pouvons distinguer deux types : des métriques superficielles basées sur les chaînes de caractères et des métriques apprises.

3.1.1 Métriques superficielles basées sur les chaînes de caractères

Ces métriques opèrent au niveau superficiel de la phrase. Typiquement, elles comparent la traduction automatique (également appelée la traduction candidate) à une traduction de référence. Plus ces deux phrases se ressemblent, plus la traduction est jugée correcte. Voici une description des métriques les plus utilisées.

Les métriques basées sur la comparaison des n-grammes

ChrF (Popović, 2015) : cette métrique qui est basée sur les caractères, compare les n-grammes de caractères (et non pas de mots) entre la traduction automatique et la traduction de référence et calcule la F-mesure des correspondances.

BLEU (Papineni *et al.*, 2002) : compare les n-grammes de mots de la phrase candidate aux n-grammes de mots de la phrase de référence. Le score final est le produit de la précision des n-grammes et de la pénalité de brièveté qui pénalise les traductions plus courtes que la référence. De nombreuses implémentations sont disponibles, celle qui est recommandée est sacreBLEU (Post, 2018).

METEOR (Denkowski & Lavie, 2011) : prend en compte non seulement les formes exactes des mots, mais également les lemmes, les synonymes et les paraphrases. La métrique attribue un poids différent selon la partie du discours du mot. Comme cela requiert des ressources externes, un nombre limité de langues est supporté. Une méthode pour adapter l'approche au multilinguisme a été proposée (Elloumi *et al.*, 2015).

NIST (Doddington, 2002) : son fonctionnement est similaire au score BLEU, mais cette métrique calcule le score avec l'aide d'un poids qui est plus important quand la séquence de mots correcte est moins probable d'apparaître dans un texte, donnant ainsi plus de poids aux mots porteurs d'information.

D'autres métriques ont été proposées, nous pouvons encore citer ROUGE (Lin & Och, 2004) basée sur le calcul de rappel qui a été proposé pour l'évaluation de résumé automatique, GTM (Melamed *et al.*, 2003) basée sur la F-mesure, BLANC (Lita *et al.*, 2005), WNM (Babych & Hartley, 2004), CDER (Leusch *et al.*, 2006) et SIA (Liu & Gildea, 2006).

Les métriques basées sur la distance d'édition

CharacTER (Wang *et al.*, 2016) : cette approche est très similaire à TER, mais au lieu d'opérer au niveau des mots, elle opère au niveau des caractères.

TER (Snover *et al.*, 2006) : calcule le nombre d'éditions de mots (insertions, suppressions, déplacements et substitutions) nécessaires pour passer de la traduction candidate à la traduction de référence.

Parmi d'autres approches, nous pouvons citer WER (Nießen *et al.*, 2000) qui est similaire à TER, mais qui ne compte pas le déplacement comme une édition ou PER (Tillmann *et al.*, 1997) qui ignore l'ordre des mots et les seules éditions sont l'insertion et la suppression.

Le défaut principal de ces métriques est qu'elles sont limitées par la forme superficielle de la phrase. Cela veut dire que la traduction et la référence sont comparées sur la base de la forme uniquement. Dès que la forme d'un mot n'est pas identique à la forme de la référence ceci est considéré comme une erreur. Certaines métriques utilisent des ressources externes (comme des dictionnaires de synonymes ou de la lemmatisation) pour surmonter cette limite. Cela permet de considérer les synonymes ou les formes fléchies du mot comme une traduction correcte malgré leur forme différente. Une deuxième approche pour surmonter la difficulté des formes fléchies des mots consiste à utiliser des métriques qui opèrent au niveau des caractères. Typiquement, la flexion concerne une sous-partie du mot (comme les suffixes), c'est pourquoi comparer les caractères se montre une approche plus flexible que de comparer les mots entiers.

3.1.2 Métriques apprises

Les métriques apprises se distinguent des métriques superficielles par le fait qu'elles sont basées sur l'apprentissage automatique. Souvent, elles exploitent les plongements de mots issus des modèles pré-entraînés et ainsi elles sont plus robustes par rapport aux changements de la forme. Dans la suite nous présentons les approches les plus utilisées et/ou les mieux classées. Ces métriques sont aussi appelées *métriques neuronales* parce qu'elles sont basées sur des modèles de langues neuronaux (ceci est vrai pour toutes les métriques citées dans la suite sauf BEER).

Métriques non-supervisées

La majorité de ces approches reposent sur les calculs de similarité entre les plongements de mots des modèles pré-entraînés. Il s'agit d'apprentissage non-supervisé, parce que ces approches n'ont pas besoin de données annotées en jugements humains et elles reposent sur la logique que la distance entre les représentations de deux phrases dans l'espace vectoriel correspond à leur similarité sémantique.

MEANT (Lo, 2017) : utilise les plongements de mots (Mikolov *et al.*, 2013) et les analyses sémantiques peu superficielles qui déterminent la structure prédicat - argument entre les mots de chaque phrase. La métrique calcule la similarité lexicale et structurale entre la phrase candidate et la phrase de référence.

YISI-1 (Lo, 2019) : calcule la similarité sémantique de la phrase candidate et de référence en utilisant des plongements de mots contextuels de BERT (Devlin *et al.*, 2018). Un analyseur sémantique peut aussi être utilisé pour exploiter les structures sémantiques des deux phrases.

BERTscore (Zhang *et al.*, 2019) : utilise des plongements de mots contextuels de BERT (Devlin *et al.*, 2018) pour calculer la distance cosinus entre les vecteurs des mots de la traduction et de la référence.

Il y a également des approches non-supervisées qui ne reposent pas sur le calcul de similarité.

Prism (Thompson & Post, 2020a) : cette approche utilise un modèle de génération de paraphrases (Thompson & Post, 2020b) pour produire un score de la traduction automatique par rapport à la phrase de référence. Le modèle de paraphrases est multilingue (entraîné sur 39 langues), il peut évaluer la traduction vers toutes ces langues. Son avantage est qu'il n'a pas besoin de jugements humains pour son entraînement.

Métriques supervisées

Ces types d'approche sont entraînés à prédire les jugements humains qui sont fournis en données d'entraînement. La majorité des approches sont construites sur la base de modèles de langues, elles nécessitent alors un modèle pour la langue en question (ou un modèle multilingue) et des données annotées en jugements humains.

BEER (Stanojević & Sima'an, 2014) : cette approche est basée sur un modèle linéaire qui combine les caractéristiques linguistiques de la similarité (comme la précision, le rappel et la F-mesure de mots et de caractères) entre la traduction candidate et de référence avec les caractéristiques d'arbres de permutation (Zhang & Gildea, 2007) qui prennent en compte l'ordre de mots pour prédire le score.

COMET (Rei *et al.*, 2020) [2] : cette métrique est construite sur la base du modèle multi-langues XLM-R (Conneau *et al.*, 2019) et elle a été entraînée en utilisant les jugements d'évaluation humaine de type évaluation directe de WMT des années 2017 à 2020. Elle prend en entrée non seulement les deux traductions (candidate et référence), mais également la phrase source.

BLEURT (Sellam *et al.*, 2020) : cette métrique a été développée pour l'évaluation de la génération du langage naturel. Elle exploite le modèle anglais BERT (Devlin *et al.*, 2018), qui est d'abord affiné sur des données synthétiques et ensuite affiné une deuxième fois sur les jugements humains de l'évaluation de la TA de type évaluation directe provenant de WMT.

UNITE (Wan *et al.*, 2022) : à la différence des autres approches proposées, UNITE est un modèle qui peut servir pour l'évaluation de la TA avec (1) la source uniquement, (2) la référence uniquement, (3) la source et référence combinées. Il repose sur des modèles de langue pré-entraînés qui sont affinés sur des jugements humains.

MATESE (Perrella *et al.*, 2022) : cette métrique repose sur l'approche de l'annotation humaine MQM. Elle utilise des modèles multi-langues comme Conneau *et al.* (2019) et Liu *et al.* (2020a) pour annoter les erreurs de la phrase candidate et les classifier en erreurs majeures et mineures. Le score final est calculé selon les poids tels que défini dans le protocole MQM.

Les métriques décrites ci-dessus nécessitent une traduction de référence. Certains travaux ont également exploré la possibilité d'une évaluation sans besoin de référence (en anglais cette approche est appelée *quality estimation*) et ont proposé des variantes de leur métrique principale. Ceci est le cas pour COMET-QE (Rei *et al.*, 2021), PRISM-src (Thompson & Post, 2020a) et YISI-2 (Lo, 2019). UNITE est la seule métrique où le même modèle peut servir pour évaluer avec ou sans référence.

Les métriques automatiques doivent être évaluées pour savoir si leur score est fiable. La sous-section 3.3 décrit comment cela est fait.

Les métriques automatiques classiques (dont le score BLEU) présentent une faiblesse en évaluant des moteurs de qualité élevée dont elles sont incapables de capturer les différences subtiles (Fomicheva

2. Plus précisément le modèle wmt22-comet-da.

& Specia, 2019; Mathur *et al.*, 2020a). Des travaux plus récents aboutissent à la conclusion que les métriques neuronales ont les meilleures performances et recommandent l'usage de COMET comme métrique principale (Freitag *et al.*, 2022; Kocmi *et al.*, 2021).

Les métriques apprises montrent des résultats encourageants, pourtant elles relèvent de l'effet *boîte noire* : leur score est difficile à expliquer et elles peuvent contenir des biais dont nous ne sommes pas encore conscients. Les métriques basées sur les chaînes de caractères sont moins problématiques de ce point de vue là parce que leur score est facilement explicable.

Les travaux de Kocmi *et al.* (2021) se sont intéressés à étudier si les métriques apprises présentent des biais selon leur mode d'entraînement. Ils ont analysé les performances des métriques selon différents scénarios tel qu'en fonction du couple de langues, de l'alphabet utilisé et du domaine des documents sans pourtant découvrir de biais. Il faut noter que ces analyses ont été effectuées avec la mesure de précision par paire (présentée dans la sous-section 3.3) qui s'est montrée peu performante. Une nouvelle analyse des biais avec une mesure plus fiable serait alors intéressante. Les auteurs préviennent que la pluralité des métriques pourrait entraîner des initiatives malhonnêtes à simplement choisir celle qui donne le résultat le plus opportun, et proposent d'utiliser COMET comme standard.

Aujourd'hui, nous observons une période de transition où le BLEU est encore très utilisé, mais aussi fortement critiqué. Les années d'usage du score BLEU ont fait que ses faiblesses sont plutôt bien connues, même si l'interprétation de petites différences de scores est restée peu claire (Popescu-Belis, 2019). Il a été démontré que la communauté se fiait pendant de nombreuses années sur la comparaison des scores BLEU qui en réalité n'étaient pas comparables (Marie *et al.*, 2021), un changement est alors nécessaire. La métrique COMET commence à se montrer comme la nouvelle métrique standard, sans pourtant encore être complètement adoptée par tous les acteurs. À part le choix de la métrique, il est indispensable d'utiliser des tests de significativité statistique avant de tirer des conclusions des scores obtenus et un des tests très utilisé est le ré-échantillonnage bootstrap (*bootstrap resampling*) (Efron & Tibshirani, 1994).

3.2 Évaluation au niveau de document

Dans le domaine de la TA, une lignée importante de travaux se focalise sur la traduction au niveau du document. Ce type de traduction a pour but de franchir la limite de la phrase et de traduire des segments plus longs, assurant ainsi la prise en compte du contexte. Certains phénomènes linguistiques ont besoin d'un contexte plus large pour permettre une bonne compréhension du texte et par conséquent une bonne traduction. Parmi ces phénomènes, nous pouvons citer par exemple :

- *les pronoms anaphoriques* : le pronom doit respecter l'accord avec son antécédent qui peut se trouver dans les phrases précédentes.
- *les noms polysémiques* : le sens des mots polysémiques dépend du contexte.
- *la cohérence lexicale et terminologique* : certains termes doivent être traduits de la même façon au sein du document. C'est le cas des contrats où au début apparaissent les définitions des parties avec leurs appellations qui doivent être respectées tout au long du document.
- *la structure discursive* : la traduction des connecteurs logiques entre les phrases nécessite la connaissance de la relation logique entre les phrases.
- *la richesse lexicale* : pour ne pas répéter les mêmes expressions, l'accès aux phrases précédentes est nécessaire pour choisir des nouvelles formulations.

Les métriques d'évaluation traditionnelles ne sont pas adaptées pour évaluer les améliorations

obtenues et cela surtout parce qu'elles évaluent la traduction isolée. Aussi, la contextualisation de la traduction touche à une petite proportion du texte, les erreurs ou améliorations de la traduction de ces éléments problématiques ne sont reflétées que très peu dans le score final. De plus, lors de la traduction des pronoms anaphoriques, plus que l'accord avec la traduction de référence, c'est l'accord avec son antécédent qui compte, comme montré dans le tableau 1. Tout cela appauvrit l'évaluation et la rend moins fiable ; d'autant plus quand l'objectif est de développer et d'évaluer un système de traduction sensible au contexte.

Phrase source :	Look, a stone. I'll throw it.
Traduction automatique :	Regarde, un caillou. Je vais **le** lancer.
Traduction de référence :	Regarde, une pierre. Je vais **la** lancer.

TABLE 1 – Exemple d'une traduction correcte même si différente de la référence.

Les travaux proposant des méthodes d'évaluation au niveau du document peuvent être classés en deux catégories. Pour mesurer les améliorations sur certaines erreurs, quelques auteurs utilisent des méthodes basées sur les références (appelées *ground truth*). Par exemple, la test-suites Protest (Guillou & Hardmeier, 2016) cible l'évaluation de la traduction des pronoms anaphoriques. Wong *et al.* (2020) proposent un jeu de test pour les pronoms cataphoriques. La faiblesse de ces approches repose sur le fait qu'elles comparent la traduction à la référence, pourtant comme nous l'avons vu dans le tableau 1, cette méthode n'est pas adaptée.

La deuxième approche est constituée des méthodes basées sur les paires contrastives (en anglais *contrastive pairs*). Il s'agit de test-suites qui ciblent certains phénomènes spécifiques et où pour chaque phrase source, il est associé son contexte, une traduction correcte et une ou plusieurs traductions incorrectes. Le système est évalué sur sa capacité à donner une probabilité plus élevée à la traduction correcte (Rios *et al.*, 2018). Pour pouvoir utiliser cette approche, il est nécessaire d'avoir accès aux probabilités du modèle. Il faut noter que l'approche est basée uniquement sur la comparaison des probabilités à générer la phrase correcte et celle incorrecte. Pourtant, ceci n'est pas vraiment une évaluation de la traduction, parce que rien n'assure que le moteur générerait une de ces phrases (Popescu-Belis, 2019; Post & Junczys-Dowmunt, 2023).

Nous pouvons citer la test-suite de Lopes *et al.* (2020) pour l'évaluation de la traduction de pronoms de l'anglais vers le français, Müller *et al.* (2018) pour la même tâche pour la paire de langues de l'anglais vers l'allemand ou encore Bawden *et al.* (2018) pour évaluer la cohérence lexicale. Rios *et al.* (2018) ont proposé un ensemble pour évaluer la désambiguïsation des mots polysémiques pour les paires de langues allemand-anglais et français-anglais. De la même manière, Alves *et al.* (2022) ont proposé SMAUG, une test-suite contrastive qui cible les erreurs graves en anglais-portugais, espagnol-anglais et portugais-anglais. Il ne se focalise pas sur les erreurs liées au contexte, mais utilise la même logique que celle des paires contrastives.

Les deux approches décrites ci-dessus ne peuvent pas être utilisées comme seule métrique automatique. Leur caractère reste complémentaire et elles sont caractérisées par la faiblesse d'être coûteuses et chronophages à l'élaboration. Elles ne traitent que quelques problèmes bien précis et uniquement pour certaines langues. De nouvelles approches, à vocation plus globale sont apparues. Vernikos *et al.* (2022) proposent d'adapter des métriques neuronales existantes en encodant la phrase à évaluer avec son contexte . Il s'agit d'une extension possible à effectuer avec n'importe quelle métrique neuronale, avec uniquement l'entrée qui est modifiée (en intégrant le contexte) sans nouvel entraînement de la métrique. Les auteurs ont exemplifié leur approche avec BERTScore, Prism, COMET et COMET-src.

Liu *et al.* (2020b) ont utilisé le score BLEU en calculant les correspondances des n-grammes de mots sur l'ensemble du document et appellent cette version de la métrique *d-bleu* (*document-bleu*). Pourtant, comme l'indiquent Post & Junczys-Dowmunt (2023), il y a encore besoin d'une méthode automatique au niveau du document globale qui soit robuste et fiable.

3.3 Méta-évaluation des métriques automatiques

Une méta-évaluation est nécessaire pour connaître la qualité des métriques et savoir lesquelles donnent les scores les plus fiables. Les métriques sont évaluées en comparant les scores automatiques avec les résultats de l'évaluation humaine. L'évaluation humaine devient alors la référence d'or pour l'évaluation (et éventuellement l'entraînement) des métriques automatiques. La manière classique pour évaluer les métriques est de mesurer la corrélation entre les scores prédits par les métriques et ceux attribués par les évaluateurs humains.

Un des acteurs majeurs dans le développement et l'évaluation de nouvelles métriques automatiques de TA est la conférence WMT. Au cours des premières années de la conférence, à partir de 2007, la corrélation de rang de Spearman était initialement utilisée, mais il a été montré que cette corrélation pénalisait très fortement les désaccords même si la différence de qualité de deux systèmes de traduction était petite (Mathur *et al.*, 2020a). C'est pourquoi, à partir de 2014, elle a été remplacée par la corrélation de Pearson. Le test de Williams (Graham & Baldwin, 2014) est utilisé pour vérifier si la différence de performance de deux métriques est statistiquement significative.

Cette façon d'évaluer a récemment été mise en doute. Mathur *et al.* (2020a) démontrent que la présence des modèles dont la qualité mesurée est très distante de celle des autres modèles change considérablement la valeur de la corrélation. Ces modèles sont très faciles à distinguer des autres par les métriques, ce qui se traduit par une corrélation gonflée et donne une fausse confiance dans la fiabilité des métriques. Les auteurs recommandent de recalculer la corrélation après avoir identifié et supprimé ces systèmes de qualité très différente qui sont considérés comme « donnée aberrante ».

Kocmi *et al.* (2021) ont analysé les scores de l'évaluation des systèmes classés en trois groupes selon la paire de langues : 1. l'anglais comme langue cible, 2. l'anglais comme langue source, 3. les paires où ni la langue cible ni la langue source n'est l'anglais. Ils se sont aperçus que les scores des métriques sont sur des échelles différentes selon les paires de langues. Ils argumentent que cela rend l'usage de la corrélation inutilisable. Comme les échelles ne sont pas les mêmes, il n'est pas adéquat de calculer la moyenne des corrélations à travers les différents scénarios pour déterminer quelle métrique a le taux de corrélation le plus élevé. Les auteurs proposent une nouvelle mesure qu'ils appellent « pair-wise accuracy », la *précision par paire* qui est définie comme le pourcentage des paires de systèmes pour lesquels la métrique automatique a assigné le score le plus élevé au même système que l'évaluation humaine.

Cette méthode a été adoptée en 2021 à la conférence WMT (Freitag *et al.*, 2021b), pourtant elle s'est révélée peu discriminante, car elle ne montrait pas assez de différences significatives entre les métriques. Ainsi, plusieurs métriques ont été classées sur la même place. L'année suivante, WMT a alors proposé une nouvelle méthode (Freitag *et al.*, 2022) qui calcule le rang moyen de chaque métrique basée sur la corrélation à travers les différents scénarios, comme : la paire de langues, le domaine, le niveau d'évaluation (niveau du segment ou du système), le coefficient de corrélation utilisé (la précision par paire, le coefficient de Pearson et de Kendall), la méthode de calcul de moyennes. Le classement final est une moyenne des classements à travers les scénarios cités.

4 Discussion

Avec le nombre grandissant de métriques automatiques, la question principale est de savoir laquelle choisir et pour cela, la méta-évaluation doit apporter la réponse. Comme nous l'avons vu, la méta-évaluation est étroitement liée à (1) la méthode de mesure de corrélation et (2) les jugements d'évaluation humaine. Kocmi *et al.* (2021) ont présenté un classement de performance des métriques automatiques. Comme référence, ils ont utilisé les jugements d'évaluation directe collectés lors des évaluations internes chez Microsoft. Comme méthode de mesure de corrélation, ils ont introduit et utilisé la précision par paire. Pourtant, il s'est montré que les jugements humains ont des défauts de significativité statistique (Wei *et al.*, 2022) et que la mesure de la précision par paire n'arrive pas à bien distinguer entre les métriques (Freitag *et al.*, 2022). Tout cela montre que malgré les avancées en matière d'évaluation humaine, la procédure de la méta-évaluation des métriques présente des difficultés et doit être davantage étudiée et développée.

Même s'il semble que l'évaluation MQM est la plus fiable, la quantité de cette donnée reste petite par rapport à l'historique d'évaluation directe. En plus, cette approche d'évaluation est plus chronophage, ce qui rend son utilisation peu pratique. Cela fait que dans certains cas, les organisateurs d'évaluations préfèrent d'autres approches, comme Kocmi *et al.* (2022) qui ont combiné l'évaluation directe en échelle continue avec l'évaluation sur l'échelle de Likert qui corrèle bien avec l'évaluation MQM. Au mieux de notre connaissance, aucune analyse statistique de grande échelle telle que celle de (Wei *et al.*, 2022) n'a été faite sur les annotations MQM, ce qui se révèlerait fortement intéressant.

Du point de vue de l'évaluation humaine comme donnée d'entraînement pour les métriques automatiques, les données les plus abondantes restent les jugements d'évaluation directe. En effet, les nouvelles métriques de la famille COMET commencent à utiliser les annotations MQM pour l'entraînement, mais l'évaluation directe reste la source de données la plus riche. Pourtant, il a été démontré que ces jugements corrèlent très peu avec les jugements d'annotation MQM (considérées comme la référence la plus exacte et véridique) (Freitag *et al.*, 2021a). Nous pouvons alors supposer qu'avec la croissance d'ensemble de données annotées en MQM, les métriques automatiques à leur tour vont être améliorées.

Une vraie carence reste l'évaluation automatique sensible au contexte. L'approche des test-suites est insuffisante pour deux raisons : 1) elle n'est qu'une approche complémentaire à utiliser avec une autre métrique, 2) les ensembles de test ne ciblent que certains cas bien précis et en nombre limité de langues. L'approche de Vernikos *et al.* (2022) propose une solution, cependant elle ne fait que modifier l'entrée du modèle. Une métrique apprise entraînée avec des données annotées au niveau du document est souhaitable et peut faire l'objet de futurs travaux.

Un autre aspect à prendre en compte est l'impact des traductions de référence sur la performance des métriques, puisque la plupart des métriques les plus performantes ont besoin de ces traductions. La qualité de ces références impacte largement leur performance et il n'est pas encore clair de savoir quelles sont les caractéristiques d'une bonne référence (Freitag *et al.*, 2022). Les métriques sans références qui deviennent de plus en plus robustes (Rei *et al.*, 2021) pourraient être une solution à ce problème.

Une autre piste de travaux possibles est d'investiguer davantage si les différentes métriques présentent un biais par rapport à certaines familles de modèles de traduction (selon leur architecture ou langues). Avec le développement rapide de grands modèles de langues (*large language models*), il serait intéressant de tester la fiabilité des métriques d'évaluation sur les traductions produites par ces

modèles.

5 Conclusion

En conclusion, cet article a présenté l'état de l'art de l'évaluation de la traduction automatique, en soulignant que l'évaluation de la qualité de la traduction automatique est subjective et difficile. Nous avons présenté les différentes approches de l'évaluation humaine avec leurs avantages et inconvénients. Ensuite, nous avons présenté les approches d'évaluation automatique les plus utilisées dans le domaine de la traduction. D'après les études récentes, il est impératif d'arrêter l'usage du score BLEU comme métrique principale. Les métriques neuronales se montrent plus fiables. Notamment le score COMET a attiré beaucoup d'attention, il a été classé comme métrique la plus performante dans plusieurs études et pour cela commence à devenir la nouvelle métrique standard. Nous avons montré qu'il y a un besoin pressant de développer des métriques fiables pour évaluer la traduction tout en prenant compte de son contexte.

L'évaluation humaine est essentielle pour le développement des approches automatiques, elle est utilisée pour la méta-évaluation et (selon l'approche) pour l'entraînement. Il faut mentionner également que la méta-évaluation qui consiste à calculer la corrélation entre les scores automatiques et humains ne fait pas l'unanimité et de nouvelles méthodes ont été proposées sans encore définir de protocole. En somme, nous avons souligné l'importance de continuer à développer des protocoles d'évaluation humaine et de méta-évaluation fiables pour ainsi créer une base solide pour le développement des métriques automatiques. Une bonne évaluation étant la condition essentielle pour un développement optimal dans le domaine de la traduction automatique, cela contribuera à l'avancement de manière globale.

Remerciements

Ce travail était effectué dans le cadre d'une convention CIFRE, gérée par l'Association Nationale de la Recherche Technique, et établie entre le Laboratoire d'Informatique de Grenoble et la société Lingua Custodia. Nous tenons à remercier les encadrants Emmanuelle Esperança-Rodier, Marco Dinarelli, Hervé Blanchon et Raheel Qader pour leurs conseils et commentaires pertinents. Nous remercions également les relecteurs anonymes.

Références

ALVES D., REI R., FARINHA A. C., DE SOUZA J. G. & MARTINS A. F. (2022). Robust mt evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 469–478.

BABYCH B. & HARTLEY T. (2004). Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 621–628.

BAWDEN R., SENNRICH R., BIRCH A. & HADDOW B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers), p. 1304–1313, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1118.

BOJAR O., FEDERMANN C., FISHEL M., GRAHAM Y., HADDOW B., HUCK M., KOEHN P. & MONZ C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, volume 2, p. 272–307.

BOJAR O., FEDERMANN C., HADDOW B., KOEHN P., POST M. & SPECIA L. (2016). Ten years of wmt evaluation campaigns : Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, p. 27–34.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv :1911.02116*.

DENKOWSKI M. & LAVIE A. (2011). Meteor 1.3 : Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, p. 85–91.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, p. 138–145.

EFRON B. & TIBSHIRANI R. J. (1994). *An introduction to the bootstrap*. CRC press.

ELLOUMI Z., BLANCHON H., SERASSET G. & BESACIER L. (2015). METEOR for multiple target languages using DBnary. In *Proceedings of Machine Translation Summit XV : Papers*, Miami, USA.

FOMICHEVA M. & SPECIA L. (2019). Taking mt evaluation metrics to extremes : Beyond correlation with human judgments. *Computational Linguistics*, **45**(3), 515–558.

FREITAG M., FOSTER G., GRANGIER D., RATNAKAR V., TAN Q. & MACHEREY W. (2021a). Experts, errors, and context : A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, **9**, 1460–1474.

FREITAG M., REI R., MATHUR N., LO C.-K., STEWART C., AVRAMIDIS E., KOCMI T., FOSTER G., LAVIE A. & MARTINS A. F. (2022). Results of wmt22 metrics shared task : Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 46–68.

FREITAG M., REI R., MATHUR N., LO C.-K., STEWART C., FOSTER G., LAVIE A. & BOJAR O. (2021b). Results of the wmt21 metrics shared task : Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, p. 733–774.

GRAHAM Y. & BALDWIN T. (2014). Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 172–176, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1020.

GRAHAM Y., BALDWIN T., MOFFAT A. & ZOBEL J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop*

and Interoperability with Discourse, p. 33–41, Sofia, Bulgaria : Association for Computational Linguistics.

GRAHAM Y., HADDOW B. & KOEHN P. (2019). Translationese in machine translation evaluation. *arXiv preprint arXiv :1906.09833*.

GUILLOU L. & HARDMEIER C. (2016). Protest : A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 636–643.

HASSAN H., AUE A., CHEN C., CHOWDHARY V., CLARK J., FEDERMANN C., HUANG X., JUNCZYS-DOWMUNT M., LEWIS W., LI M. *et al.* (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv :1803.05567*.

KOCMI T., BAWDEN R., BOJAR O., DVORKOVICH A., FEDERMANN C., FISHEL M., GOWDA T., GRAHAM Y., GRUNDKIEWICZ R., HADDOW B. *et al.* (2022). Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 1–45.

KOCMI T., FEDERMANN C., GRUNDKIEWICZ R., JUNCZYS-DOWMUNT M., MATSUSHITA H. & MENEZES A. (2021). To ship or not to ship : An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv :2107.10821*.

KOPONEN M. (2016). *Machine Translation Post-editing and Effort : Empirical Studies on the Post-editing Process*. Thèse de doctorat, University of Helsinki, Finland.

LÄUBLI S., SENNRICH R. & VOLK M. (2018). Has machine translation achieved human parity ? a case for document-level evaluation. *arXiv preprint arXiv :1808.07048*.

LEUSCH G., UEFFING N. & NEY H. (2006). Cder : Efficient mt evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, p. 241–248.

LIN C.-Y. & OCH F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 605–612.

LITA L. V., ROGATI M. & LAVIE A. (2005). Blanc : Learning evaluation metrics for mt. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 740–747.

LIU D. & GILDEA D. (2006). Stochastic iterative alignment for machine translation evaluation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, p. 539–546.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020a). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020b). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LO C.-K. (2017). Meant 2.0 : Accurate semantic mt evaluation for any output language. In *Proceedings of the second conference on machine translation*, p. 589–597.

LO C.-K. (2019). Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, p. 507–513.

LOMMEL A., BURCHARDT A. & USZKOREIT H. (2014). Multidimensional quality metrics (mqm) : A framework for declaring and describing translation quality metrics. *Tradumàtica : tecnologies de la traducció*, **0**, 455–463. DOI : 10.5565/rev/tradumatica.77.

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

MARIE B., FUJITA A. & RUBINO R. (2021). Scientific credibility of machine translation research : A meta-evaluation of 769 papers. *arXiv preprint arXiv :2106.15195*.

MATHUR N., BALDWIN T. & COHN T. (2020a). Tangled up in bleu : Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv :2006.06264*.

MATHUR N., WEI J., FREITAG M., MA Q. & BOJAR O. (2020b). Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, p. 688–725.

MELAMED I. D., GREEN R. & TURIAN J. (2003). Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, p. 61–63.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MÜLLER M., RIOS A., VOITA E. & SENNRICH R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv :1810.02268*.

NIESSEN S., OCH F. J., LEUSCH G., NEY H. *et al.* (2000). An evaluation tool for machine translation : Fast evaluation for mt research. In *LREC*.

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.

PERRELLA S., PROIETTI L., SCIRÈ A., CAMPOLUNGO N. & NAVIGLI R. (2022). Matese : Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 569–577.

POPESCU-BELIS A. (2019). Context in neural machine translation : A review of models and evaluations. *arXiv preprint arXiv :1901.09115*.

POPOVIĆ M. (2015). chrF : character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

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*.

REI R., FARINHA A. C., ZERVA C., VAN STIGT D., STEWART C., RAMOS P., GLUSHKOVA T., MARTINS A. F. T. & LAVIE A. (2021). Are references really needed ? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, p. 1030–1040, Online : Association for Computational Linguistics.

REI R., STEWART C., FARINHA A. C. & LAVIE A. (2020). Comet : A neural framework for mt evaluation. *arXiv preprint arXiv :2009.09025*.

RIOS A., MÜLLER M. & SENNRICH R. (2018). The word sense disambiguation test suite at wmt18. In *Proceedings of the Third Conference on Machine Translation : Shared Task Papers* : Association for Computational Linguistics.

SELLAM T., DAS D. & PARIKH A. (2020). BLEURT : Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7881–7892, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.704.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas : Technical Papers*, p. 223–231.

STANOJEVIĆ M. & SIMA'AN K. (2014). Beer : Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 414–419.

THOMPSON B. & POST M. (2020a). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online : Association for Computational Linguistics.

THOMPSON B. & POST M. (2020b). Paraphrase generation as zero-shot multilingual translation : Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1 : Research Papers)*, Online : Association for Computational Linguistics.

TILLMANN C., VOGEL S., NEY H., ZUBIAGA A. & SAWAF H. (1997). Accelerated dp based search for statistical translation. In *Eurospeech*.

TORAL A., CASTILHO S., HU K. & WAY A. (2018). Attaining the unattainable ? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv :1808.10432*.

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**.

VERNIKOS G., THOMPSON B., MATHUR P. & FEDERICO M. (2022). Embarrassingly easy document-level mt metrics : How to convert any pretrained metric into a document-level metric. *arXiv preprint arXiv :2209.13654*.

VOJTĚCHOVÁ T., NOVÁK M., KLOUČEK M. & BOJAR O. (2019). Sao wmt19 test suite : Machine translation of audit reports. *arXiv preprint arXiv :1909.01701*.

WAN Y., LIU D., YANG B., ZHANG H., CHEN B., WONG D. & CHAO L. (2022). UniTE : Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8117–8127, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.558.

WANG W., PETER J.-T., ROSENDAHL H. & NEY H. (2016). CharacTer : Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, p. 505–510, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-2342.

WEI J. T.-Z., KOCMI T. & FEDERMANN C. (2022). Searching for a higher power in the human evaluation of mt. *arXiv preprint arXiv :2210.11612*.

WONG K., MARUF S. & HAFFARI G. (2020). Contextual neural machine translation improves translation of cataphoric pronouns. *arXiv preprint arXiv :2004.09894*.

ZHANG H. & GILDEA D. (2007). Factorization of synchronous context-free grammars in linear time. In *Proceedings of SSST, NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, p. 25–32.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.