COMET : Un Cadre Neuronal pour l'Évaluation de la Traduction Automatique

Ricardo Rei **Craig Stewart** **Ana C Farinha** **Alon Lavie**
Unbabel IA
{ricardo.rei, craig.stewart, catarina.farinha, alon.lavie}@unbabel.com

Résumé

Nous présentons COMET, un cadre neuronal pour former des métriques d'évaluation de la traduction automatique multilingue qui obtiennent un nouvel état-de-l'art de corrélation avec le jugement humain-niveaux d'art. Notre cadre tire parti des récentes percées-ments. des avancées dans le langage pré-entraîné multilingue modélisation aboutissant à une grande multilinguisme et des modèles d'évaluation MT adaptables qui exploitent informations à la fois de la source d'entrée et d'un traduction de référence en langue cible dans l'ordre pour prédire plus précisément la qualité de la MT. À pour présenter notre cadre, nous formons trois mod-modèles avec différents types de jugements humains : *Évaluations Directes, Transmissions Humaines Cibles* *iatisées* *Taux d'édition de traduction* et Qualité *idimensionnelle* *ity Metrics. Nos* modèles atteignent un nouvel état de-la performance de pointe sur le WMT 2019 Met-tâche partagée rics et démontrent une robustesse à systèmes haute performance.

1 Introduction

Historiquement, les mesures pour évaluer la qualité de la traduction automatique (TA) s'est appuyée sur l'évaluation la similarité entre une hypothèse générée par MT et une traduction de référence générée par l'homme dans le langue cible. Les mesures traditionnelles se sont concentrées sur des caractéristiques de base, au niveau lexical, telles que le comptage le nombre de n-grammes correspondants entre le MT hypothèse et la traduction de référence. Métriques tel que BLEU (Papineni et al., 2002) et METEOR (Lavie et Denkowski, 2009) restent populaires en tant que un moyen d'évaluer les systèmes MT en raison de leur calcul léger et rapide.

Les approches neuronales modernes à la traduction automatique donnent des résultats bien meilleurs. une qualité de traduction plus élevée qui dévie souvent de transfert lexical monotone entre les langues. Pour cette raison, il est devenu de plus en plus évident que nous ne pouvons plus nous fier à des mesures telles que BLEU pour fournir une estimation précise de la qualité de MT (Barrault et al., 2019).

Alors qu'un intérêt de recherche accru dans les réseaux neuronaux des méthodes pour former des modèles et des systèmes de MT ont a conduit à une amélioration récente et spectaculaire en MT qualité, l'évaluation MT a pris du retard. Le MT la communauté de recherche repose toujours largement sur des méthodes obsolètes des mesures et aucune nouvelle norme largement adoptée n'a émergé. En 2019, la Traduction de Nouvelles WMT La Tâche Partagée a reçu un total de 153 systèmes MT soumissions (Barrault et al., 2019). Les Mesures La Tâche Partagée de la même année n'a vu que 24 sous-missions, dont presque la moitié étaient des nouveaux venus dans le Tâche Partagée d'Estimation de Qualité, adaptée en tant que métriques (Mathur et al., 2019).

Les résultats de la tâche susmentionnée sont élevés-éclaire deux défis majeurs à l'évaluation MT qui nous cherchons à aborder ici (Ma et al., 2019). À savoir, que les mesures actuelles **peinent à accu-corréler précisément avec le jugement humain à seg-niveau de ment** et échouent **à différencier adéquatement les systèmes MT les plus performant** s.

Dans cet article, nous présentons COMET[1], un PyTorch-cadre basé pour la formation hautement multilingue et des modèles d'évaluation MT adaptables qui peuvent fonc-tion comme des métriques. Notre cadre utilise cet avantage des avancées récentes dans le langage interlinguistique modélisation (Artetxe et Schwenk, 2019; Devlin et al., 2019; Conneau et Lample, 2019; Conneau et al., 2019) pour générer des estimations de prédictions de hu-jugements humains tels que *les Évaluations Directes (E* D) (Graham et al., 2013), T*ransmédiée* ap*Prochain* *me* *le taux d'édition* humaine (HTER) (Snover et al., 2006) et métriques conformes à la Qualité *Multidimensionnelle* *cadre de Métrique* d'itérité (Lommel et al., 2014).

Inspiré par des travaux récents sur l'Estimation de Qualité (QE) qui a démontré qu'il est possible d'atteindre des niveaux élevés de corrélation avec les jugements humains même sans une traduction de référence (Fonseca et al., 2019), nous proposons une nouvelle approche pour incorporer-

[1] **T**ranslinguistique **M**étrique **O**ptimisée pour **a**lu**t**ion de **E**valuation of Traduction.

Actes de la Conférence 2020 sur les Méthodes Empiriques en Traitement Automatique des Langues, pages 2685·2702,
16-20 novembre 2020. c·2020 Association pour la Linguistique Computationnelle

ation en traduisant l'entrée de la langue source dans notre évaluation MT des modèles d'ation. Traditionnellement, seuls les modèles QE ont a utilisé l'entrée de la source, tandis que l'évaluation MT les métriques d'évaluation se basent plutôt sur la traduction de référence-tion. Comme dans (Takahashi et al., 2020), no us nous , we show that l'utilisation d'un espace d'incorporation multilingue nous permet pour exploiter les informations provenant des trois entrées et démontrer la valeur ajoutée par la source en tant qu'entrée à nos modèles d'évaluation MT.

Pour illustrer l'efficacité et la flexibilité de le cadre COMET, no us formons trois modèles qui estimer différents types de jugements humains et montre des progrès prometteurs vers une meilleure corrélation au niveau du segment et robustesse face à des niveaux élevés de qualité MT.

Nous publierons à la fois le ca dre COMET et les modèles d'évaluation MT formés décrits dans ce papier à la communauté de recherche lors de la publication.

2 Architectures de Modèles

Les jugements humains sur la qualité de la TA proviennent généralement sous la forme de scores au niveau du segment, tels que DA, MQM et HTER. Pour DA, il est courant de convertir les scores en classements relatifs (DARR) lorsque R) when le nombre d'annotations par segment est limité (Bojar et al., 2017b; Ma et al., 2018, 2019). Ceci signifie que, pour deux hypothèses MT hi et hj de la même source, si le score DA attribué à hi est supérieur au score attribué à hj, hi est reconsidéré comme une "meilleure" hypothèse.[2] Pour encompass ces différences, notre cadre prend en charge deux disarchitectures distinctes : **Le modèle Estimator et le Modèle de classement de traduction.** Le fondamental la différence entre eux est l'objectif de formation. Alors que l'Estimateur est formé pour régresser directement sur un score de qualité, le modèle de classement de traduction est formé pour minimiser la distance entre un "meilleur" hypothèse et ses deux références correspondantes et sa source originale. Les deux modèles sont composés d'un encodeur inter-langues et d'une couche de regroupement.

2.1 Encodeur Cross-lingual

Le bloc de construction principal de tous les modèles dans notre cadre est un pré-entraîné, cross-linguistique modèle tel que BERT multilingue (Devlin et al., 2019), XLM (Connea u et Lample, 2019) o u XLM-RoBERTa (Conneau et al., 2019). Ce s modèle . These models contient plusieurs couches de transformateur encodeur qui sont formé à reconstruire des jetons masqués en les découvrant- établissant la relation entre ces jetons et le ceux environnants. Lorsqu'entraîné avec des données provenant de plusieurs langues cet objectif pré-entraîné a a été trouvé très efficace dans le contexte interlinguistique tâches telles que la classification de documents et le traitement naturel inférence de langue (Conneau et al., 2019), géné- 2019), generalisant bien aux langues et scripts inconnus (Pires et al., 20 9). Pour les expériences dans ce document, nous nous appuyons sur XLM-RoBERTa (base) comme notre encodeur modèle.

Étant donné une séquence d'entrée x = [x0, x1, ... xn], $x = [x_1, \ldots, x_n]$, l'encodeur produit un encastrement ℓ e(·) pour chaque jeton xj et chaque couche ∈ {0, 1, ... k}. Dans notre $j \in \{1, \ldots, k\}$. In our cadre, nous appliquons ce processus à la source, Hypothèse MT, et référence afin de les cartographier dans un espace de caractéristiques partagé.

2.2 Couche de Pooling

Les incorporations générées par la dernière couche de la les encodeurs pré-entraînés sont généralement utilisés pour le réglage fin modèles à de nouvelles tâches. Cependant, (Tenney et al., 2019) a montré que différentes couches à l'intérieur du network peut capturer des informations linguistiques qui sont relévant pour différentes tâches en aval. Dans le cas de l'évaluation MT, (Zhang et al., 2020) a mon tré que showed that différentes couches peuvent atteindre différents niveaux de corrélation et que l'utilisation de la dernière couche seulement souvent donne des résultats inférieurs. Dans ce travail, nous a utilisé l'approche décrite dans Peters et al. (2018) (2018) et rassembler des informations des sources les plus importantes encoder les couches en une seule intégration pour chaque token, ej, en uti sant un mécanisme d'attention couche par couche. Cet encastrement est ensuite calculé comme suit :

$$ex_j = \mu_E \, x_j \quad (1)$$

où μ est un coefficient de poids entraînable, Ej = E_j = [$e_j^{(0)}$, e(1)j, . . . e(k)j] correspond au vecteur de incrustations de couche pour le jeton xj, et · , and α Votre texte est vide, veuillez fournir softmax([·(1), ·(2), ·(k)]) est un ve cteur corre $\alpha^{(2)}, \ldots, \alpha^{(k)}]$) is a vector correcorrespondant aux poids entraînables couche par couche. Dans afin d'éviter le surajustement à l'information contenu dans une seule couche, nous avons utilisé l'abandon de couche (Kondratyuk et Straka, 2019), dans lequel avec un , in which with a probabilité p le poids (i) est fixé à ... i) is set to $-\infty$.

Enfin, comme dans (Reimers et Gurevych, 2019) , 2019), nous appliquons une moyenne de regroupement au mot résultant des embeddings pour dériver un embedding de phrase pour chaque segment.

[2] Dans la Tâche Partagée des Métriques WMT, si la différence entre les scores DA n'est pas supérieur à 25 points, ces segments sont exclus des données DARR.

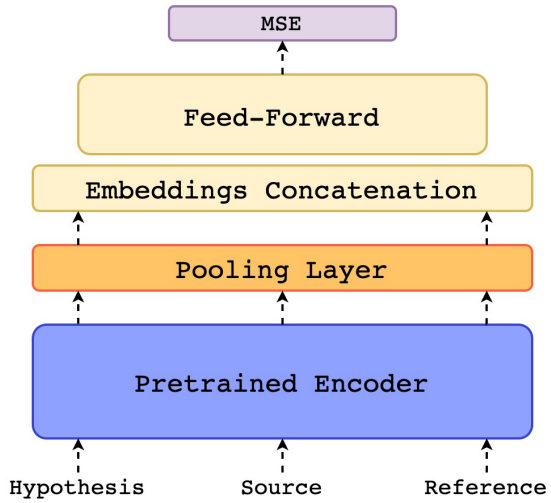

Figure 1 : Architecture du modèle d'estimation. La source, l'hypothèse et la référence sont codées indépendamment utilisant un encodeur pré-entraîné inter-langues. Le résultat les incorporations de mots sont ensuite transmises à travers une mise en commun couche pour créer une incorporation de phrase pour chaque segment. Enfin, les plongements de phrases résultants sont combinés et concaténés en un seul vecteur qui est transmis à un régresseur feed-forward. Le modèle entier est formé en minimisant l'Erreur Quadratique Moyenne (EQM).

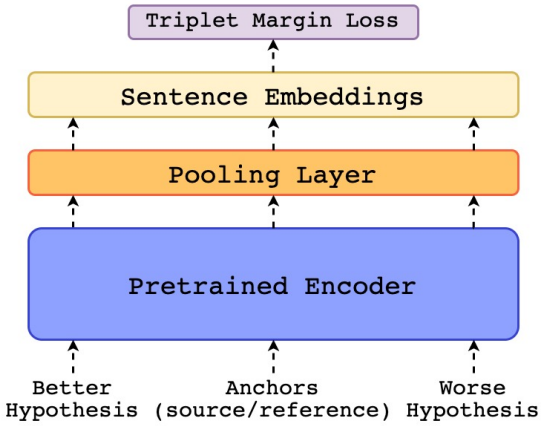

Figure 2: Architecture du modèle de classement de traduction. Cette architecture reçoit 4 segments : la source, le référence, une hypothèse "meilleure" et une "pire". Ces segments sont codés indépendamment en utilisant un pré-encodeur interlinguistique formé et une couche de regroupement sur haut. Enfin, en utilisant la perte de marge triplet (Schroff et al., 2015) nous optimisons l'espace d'incrustation résultant pour minimiser la distance entre l'hypothèse "meilleure" et les "ancres" (source et référence).

2.3 Modèle Estimateur

Étant donné une incorporation de phrase d-dimensionnelle pour le source, l'hypothèse, et la référence, nous adoptons l'approche proposée dans RUSE (Shimanaka et al., 2018) et extrayez les caractéristiques combinées suivantes :

- Produit source élément par élément $h \odot s$
- Produit de référence élément par élément $h \odot r$
- Différence absolue élément par élément de la source : $|h - s|$
- Différence de référence absolue élément par élément: $|h - r|$

Ces caractéristiques combinées sont ensuite concaténées à l'incorporation de référence et hypothèse pour intégrer dans un seul vecteur x = [h; r; $h \cdot s$; $|h \cdot r|$] qui est ensuite transmis à un régresseur feed-forward. La force de ces les caractéristiques sont dans la mise en évidence des différences entre incorporations dans l'espace des caractéristiques sémantiques.

Le modèle est ensuite entraîné pour minimiser la moyenne l'erreur quadratique entre les scores prédits et évaluations de qualité (DA, HTER ou MQM). La figure 1 illustre l'architecture proposée.

Notez que nous avons choisi de ne pas inclure la source brute incorporation (s) dans notre entrée concaténée. Tôt l'expérimentation a révélé que la valeur ajoutée par l'incorporation de la source en tant que caractéristiques d'entrée supplémentaires à notre le régresseur était négligeable au mieux. Une variation sur notre modèle estimateur HTER formé avec le vecteur $x = [h; s; r; h \cdot s; h \cdot r; |h - s|; |h - r|]$ comme $[|h - s|; |h - r|]$ as l'entrée dans le seul feed-forward réussit à stimuler-évaluant les performances au niveau du segment dans 8 des 18 langpaires de langues décrites dans la section 5 ci-dessous et le amélioration moyenne du Tau de Kendall dans ces ensembles-tings était +0.0009. Comme noté dans Zhao et al. (2020), tandis que les modèles pré-entraînés interlinguistiques sont adaptatifs à plusieurs langues, l'espace de caractéristiques entre les langues sont mal alignées. Sur cette base, nous décidé en faveur de l'exclusion de l'incorporation de la source sur l'intuition que l'information la plus importante provient de l'incorporation de référence et de la réduc-la caractérisation de l'espace des fonctionnalités permettrait au modèle de se concentrer davantage sur les informations pertinentes. Cela ne fait pas cependant nie la valeur générale de la source à notre modèle; où nous incluons des caractéristiques de combinaison tel que $h \cdot s$ et $|h - s|$ nous notons ainsi cette corrélation comme exploré plus loin dans la section 5.5 ci-dessous.

2.4 Modèle de Classement de Traduction

Notre modèle de classement de traduction (Figure 2) reçoit comme entrée un tuple (s, h^+, h^-, r) où h^+ désigne une hypothèse qui a été classée plus haut que une autre hypothèse h^-. Nous faisons ensuite passer s, h^+, h^- et r à travers notre encodeur interlinguistique et notre couche de regroupement pour obtenir une incorporation de phrase pour chaque segment dans le tuple. Enfin, en utilisant les embeddings $\{s, h^+, h^-, r\}$, nous calculons la perte de marge triplet (Schroff et al., 2015) en relation avec la source et la référence :

$$L(s, h^+, h^-, r) = L(s, h^+, h^-) + L(r, h^+, h^-) \quad (2)$$

où :

$$L(s, h^+, h^-) = \max\{0, d(s, h^+) - d(s, h^-) + \epsilon\} \quad (3)$$

$$L(r, h^+, h^-) = \max\{0, d(r, h^+) - d(r, h^-) + \epsilon\} \quad (4)$$

$d(u, v)$ est la distance euclidienne entre u et v et ϵ est une marge. Ainsi, pendant l'entraînement, le modèle optimise l'espace d'incorporation afin que la distance entre les ancres (s et la référence) et l'hypothèse "meilleure" soit au moins ϵ inférieure à la distance entre les ancres et l'hypothèse "pire" h^-.

Pendant l'inférence, le modèle décrit reçoit un triplet (s, h, r) avec seulement une hypothèse. Le score de qualité attribué à h est la moyenne harmonique entre la distance à la source $d(s, \hat{h})$ et la distance à la référence $d(r, \hat{h})$:

$$f(s, h, r) = 2 \times \frac{d(s, \hat{h}) \times d(r, \hat{h})}{d(r, \hat{h}) + d(s, \hat{h})} \quad (5)$$

Enfin, nous convertissons la distance résultante en un score de similarité limité entre 0 et 1 comme suit:

$$f(s, h, r) = \frac{1}{1 + f(s, h, r)} \quad (6)$$

3 Corpus

Pour démontrer l'efficacité de notre méthode décrite architectures de modèles (section 2), nous entraînons trois modèles d'évaluation où chaque modèle cible un différent type de jugement humain. Pour former ces modèles, nous utilisons des données de trois corpus différents: le corpus QT21, le DARR du WMT Metrics tâche partagée (2017 à 2019) et une propriétaire Corpus annoté MQM.

3.1 Le corpus QT21

Le corpus QT21 est un ensemble de données publiquement accessible[3] contenant des phrases générées par l'industrie provenant soit de un domaine de la technologie de l'information ou des sciences de la vie (Specia et al., 2017). Ce corpus contient un total de 173K tuples avec phrase source, respective référence générée par l'homme, hypothèse MT (soit d'une traduction statistique basée sur des phrases ou d'une neural MT), et MT post-édité (PE). La langue les paires représentées dans ce corpus sont : de l'anglais vers l'allemand homme (en-de), letton (en-lt) et tchèque (en-cs), et Allemand vers Anglais (de-en).

Le score HTER est obtenu en calculant le taux d'édition de traduction (TER) (Snover et al., 2006) entre l'hypothèse MT et le PE correspondant. Enfin, après avoir calculé le HTER pour chaque MT, nous avons construit un ensemble de données d'entraînement $D = \{s_i, h_i, r_i, y_i\}_{n=1}$, où s_i désigne le texte source, h_i désigne la $i^{ème}$ traduction MT, r_i la référence, et y_i le Score HTER pour l'hypothèse h_i. Dans cette tâche, nous cherchons à apprendre une régression $f(s, h, r) \rightarrow y$ that prédit l'effort humain nécessaire pour corriger le hypothèse en regardant la source, hypothèse, et référence (mais pas l'hypothèse post-éditée).

3.2 Le corpus WMT DARR

Depuis 2017, les organisateurs du WMT News La Tâche Partagée de Traduction (Barrault et al., 2019) ont collecté des jugements humains sous la forme de publicités-équacy DAs (Graham et al., 2013, 2014, 2017). Ces DA sont ensuite cartographiés en rang relatif-ings (DARR) (Ma et al., 2019). Le résultat les données pour chaque année (2017-19) forment un ensemble de données $D = \{s_i, h^+_i, h^-_i, r_i\}_{n=1}$ où h^+_i dénote un "meilleur" hypothèse et h^-_i désigne un "pire". Ici nous cherchons à apprendre une fonction $r(s, h, r)$ de telle sorte que le score attribué à h^+_i est strictement supérieur au score attribué à h^-_i ($r(s_i, h^+_i, r_i) > r(s_i, h^-_i, r_i)$). Ces données[4] contiennent un total de 24 valeurs hautes et basses-paires de langues ressources telles que le chinois vers l'anglais (zh-en) et anglais vers gujarati (en-gu).

3.3 Le corpus MQM

Le corpus MQM est une base de données interne propriétaire. des traductions générées par MT du support client

[3] Données 3QT21 : https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2390

[4] Les données brutes pour chaque année du partage des métriques WMT la tâche est publiquement disponible sur la page des résultats (ex 2019- Exemple : http://www.statmt.org/wmt19/results.html). Notez cependant que dans les fichier s README, il est hautement souligné que ces données ne sont pas bien documentées et les scripts nécessitent occasionnellement des utilitaires personnalisés qui ne sont pas disponibles.

messages de chat qui ont été annotés selon le directives établies dans Burchardt et Lommel (2014). 014). Ces données contiennent un total de 12K tuples, couvrant- traduisant 12 paires de langues de l'anglais vers : l'allemand (En-de), espagnol (en-es), espagnol d'Amérique latine espagnol (en-es-latam), français (en-fr), italien (en-it), Japonais (en-ja), Néerlandais (en-nl), Portugais (en-pt), Portugais brésilien (en-pt-br), Russe (en-ru), Suédois (en-sv), et Turc (en-tr). Notez que dans ce corpus anglais est toujours considéré comme la langue source- guage, mais jamais comme langue cible. Chaque tuple consiste en une phrase source, une génération humaine référence, une hypothèse MT, et son score MQM, dérivé des annotations d'erreur par un (ou plusieurs) annotateurs formés. La métrique MQM à laquelle on fait référence tout au long de ce document est une métrique interne définie conformément au cadre MQM (L ommel et al., 20 (MQM) . Les erreurs sont annotées sous une typologie interne définie sous trois principales er- types d'erreurs; 'Style', 'Fluidité' et 'Précision'. Notre Les scores MQM varient de ·· à 100 et sont de- défini comme :

$$\text{MQM} = 100 - \text{IM}_{inor} \frac{+ 5 \times \text{IMajor} + 10 \times \text{ICrit.}}{\text{Longueur de la Phrase}} \times 100 \quad + 10 \times I_{\text{Crit.}} \quad 100 \tag{7}$$

où IMinor *dés*igne le nombre d'erreurs mineures, IMajor le nombre d'erreurs majeures et ICrit. le n *ombre* rit. the num- nombre d'erreurs critiques.

Notre métrique MQM prend en compte la gravité- ité des erreurs identifiées dans l'hypothèse MT, conduisant à une mesure plus fine que HTER ou DA. Lorsqu'ils sont utilisés dans nos expériences, ces val- les valeurs ont été divisées par 100 et tronquées à 0. Comme dans la section 3.1, nous avons construit un ensemble de données d'entraînement $D = \{s_i, h_i, r_i, y_{ij}\}_{i=1}^{N}$, y_i désigne *le* denotes the texte source, h_i désigne l'hypothèse MT, ri le r_i the traduction de référence, et y_i le score MQM pour the MQM score for l'hypothèse h_i. ·

4 Expériences

Nous formons deux versions du modèle Estimator de- décrit dans la section 2.3 : un qui régresse sur HTER (COMET-HTER) corpus QT21, formé avec le et trained with the QT21 corpus, and un autre qui régresse sur notre mise en ·uvre propriétaire- tation de MQM (COMET-MQM) formé avec le ot trained with our corpus interne MQM. Pour le Classement de Traduction modèle, décrit dans la section 2.4, nous nous entraînons avec le Corpus WMT DA de 2017 et 2018 (COMET- OMET- RANG). Dans cette section, nous introduisons la formation configuration pour ces modèles et évaluation correspondante- configuration de l'installation.

4.1 Configuration de la Formation

Les deux versions des Estimateurs (C OMET- HTER/MQM) partagent la même configuration d'entraînement et hyper-paramètres (les détails sont inclus dans l'Ap- pendices). Pour la formation, nous chargeons le pré-entraîné encodeur et initialiser à la fois la couche de pooling et le régresseur feed-forward. Tandis que le layer-wise les scalaires · *de* la couche de pooling sont initialement définis à zéro, les poids provenant de la propagation avant sont ini- tialisé aléatoirement. Pendant la formation, nous divisons le paramètres du modèle en deux groupes : les paramètres de l'encodeur pa- ramètres, qui incluent le modèle de l'encodeur et le scalaires à partir *de* ·; es paramètres du régresseur, qui inclure les paramètres du flux d'information en avant supérieur réseau. Nous appliquons un dégel progressif et une discrimination- taux d'apprentissage innés (Howard et Ruder, 201 8), 2018), ce qui signifie que le modèle d'encodeur est gelé pour un époque tandis que le feed-forward est optimisé avec un taux d'apprentissage de 3e ·5. *pas* la première époque, le tout le modèle est affiné mais le taux d'apprentissage pour les paramètres de l'encodeur sont réglés à 1e·5 *fin* de éviter l'oubli catastrophique.

En contraste avec les deux Estimateurs, pour le Modèle COMET-R ANK plus nous a ffinons dès le départ. De plus, puisque ce modèle n'ajoute rien nouveaux paramètres au-dessus de XLM-RoBERTa (base) autre que les scalaires de couche ·, *nous* s utilisons un seul taux d'apprentissage de 1e·5 *pour* ensemble du modèle.

4.2 Configuration de l'évaluation

Nous utilisons les données de test et la configuration du WMT 2019 Tâche Partagée de Métriques (Ma et al., 201 9) afi n de in order to comparez les modèles COMET avec les performances supérieures- soumissions de la tâche partagée et autres récentes des mesures de pointe telles que BERTSCORE et and BLEURT méthode d'évaluation et celle de- formulation officielle de Kendall's Tau, ·, *de la* from the Tâche Partagée sur les Métriques WMT 2019 (Ma et al., 201 9) 2019) défini comme :

$$\tau = \text{Co}_{nc} \frac{\textit{ordant - Discor}_{dant}}{\textit{Concordant}} \quad \textit{Discordant} \quad + \text{Dis cordan} \quad t \tag{8}$$

où Concordant *est le nombre* de fois qu'une mesure attribue un score plus élevé à l'hypothèse "meilleure" h^+ et Dis cordant *est le nombre d* e fois qu'une mesure attribue un score plus élevé à l'hypothèse "pire"

[5] Pour faciliter les recherches futures, nous fournirons également, dans notre cadre, instructions détaillées et scripts pour exécuter d'autres met- des métriques tels que CHRF, BERTSCORE, and BLEURT RTSCORE, and BLEURT

Tableau 1 : Corrélations de Kendall-s *Tau* (τ) sur les paires de langues avec l'anglais comme source pour les Métriques DARR WMT19 corpus. Pour BERTSCORE, nous rapportons les résultats avec le modèle d'encodeur par défaut pour une comparaison complète, mais aussi avec XLM-RoBERTa (base) pour l'équité avec nos modèles. Les valeurs rapportées pour YiSi-1 sont directement prises de le document de tâche partagée (Ma et al., 2019).

Métrique	Traduire ce texte en-de	Traduire ce texte en-zh	fr-de	fr-gu	fr-kk	en-lt	Sorry, but you didn't provide the text to translate. Could you please provide the text you want translated? en-zh
BLEU	0,364	0,248	0,395	0,463	0,363	0,333	0,469	0,235
CHRF	0,444	0,321	0,518	0.548	0,510	0,438	0,548	0,241
YISI-1	0,475	0.351	0,537	0,551	0.546	0,470	0,585	0,355
BERTSCORE (par défaut)	0,500	0,363	0,527	0.568	0,540	0,464	0,585	0,356
BERTSCORE (xlm-base)	0,503	0,369	0,553	0,584	0,536	0,514	0,599	0,317
COMET-HTER	0,524	0,383	0,560	0,552	0,508	0,577	0,539	0,380
COMET-MQM	0,537	0,398	0.567	0.564	0,534	0,574	**0,615**	0,378
RANG-COMÈTE	**0,603**	**0,427**	**0,664**	**0,611**	**0,693**	**0.665**	0,580	**0,449**

les scores attribués aux deux hypothèses sont le même.

Comme mentionné dans les résultats de (Ma et al., 2019), corrélation au niveau des segments de toutes les métriques soumises étaient frustrantement bas. De plus, tous les soumetles métriques ted ont montré un manque dramatique de capacité à classer correctement les systèmes MT forts. Pour évaluer si nos nouveaux modèles d'évaluation MT sont plus efficaces pour aborder cette question, nous avons suivi l'évaluation décrite configuration utilisée dans l'analyse présentée dans (Ma et al., 2019), où les niveaux de corrélation sont examinés pour les parties des données DARR qui n'incluent que le top 10, 8, 6 et 4 systèmes MT.

5 Résultats

5.1 De l'anglais en X

Le tableau 1 montre les résultats pour toutes les huit paires de langues avec l'anglais comme source. Nous contrastons nos trois modèles COMET par rapport aux indicateurs de référence tels que BLEU et CHRF, la métrique gagnante de la tâche 2019 YISI-1 et BERTSCORE.

Nous observons que, de manière générale, nos trois modèles formé avec le cadre COMET surpasse, souvent par des marges significatives, toutes les autres mesures. Notre Le modèle DARR Ranker surpasse les deux Estimadeurs dans sept des huit paires de langues. Aussi, même bien que l'Estimateur MQM soit formé sur seulement 12K segments annotés, il fonctionne à peu près au même niveau avec l'Estimateur HTER pour la plupart des paires de langues, et surpasse toutes les autres mesures en en-ru.

5.2 De X en anglais

Le tableau 2 montre les résultats pour les sept langues vers l'anglais paires de langues. Encore une fois, nous contrastons nos trois modèles COMET par rapport aux indicateurs de référence tels que BLEU et CHRF, la métrique gagnante de la tâche 2019 YISI-1, en tant que BERTSCORE et BLEURT. Comme dans le Tableau 1, le modèle DARR Ranker des corrélations fortes avec les jugements humains dépassent effectuant la spécifique anglaise récemment proposée Métrique BLEURT dans cinq des sept paires de langues. Encore une fois, l'Estimateur MQM montre une surprenante force résultats malgré le fait que ce modèle a été formé avec des données qui n'incluaient pas l'anglais comme cible. Bien que l'encodeur utilisé dans nos modèles entraînés soit hautement multilingue, nous émettons l'hypothèse que ce pow- Le puissant résultat "zero-shot" est dû à l'inclusion de la source dans nos modèles.

5.3 Paires de langues n'impliquant pas l'anglais

Nos trois modèles COMET ont tous été entraînés sur données impliquant l'anglais (soit comme source, soit comme cible). Néanmoins, pour démontrer que notre metrics se généralisent bien, nous les testons sur les trois WMT 2019 paires de langues qui n'incluent pas l'anglais dans soit la source soit la cible. Comme on peut le voir dans le Tableau 3, nos résultats sont cohérents avec les observations dans Tables 1 et 2.

5.4 Robustesse face au MT de Haute Qualité

Pour l'analyse, nous utilisons le corpus DARR provenant de la Tâche partagée 2019 et évaluation sur le sous-ensemble de les données des systèmes MT les plus performants pour chaque paire de langues. Nous avons inclus des paires de langues pour lesquels nous pourrions récupérer des données pour au moins dix différents systèmes MT (c'est-à-dire tous sauf kk-en et gu-en). Nous contrastons par rapport aux propositions récentes fortes BERTSCORE et BLEURT, avec BLEU comme baseligne. Les résultats sont présentés dans la Figure 3. Pour les paires de langues où l'anglais est la cible, nos trois les modèles sont soit meilleurs soit compétitifs avec tous les auters; où l'anglais est la source, nous notons cela dans général nos mesures surpassent les performances des aut-

Tableau 2 : Corrélations de Kendall·s *Tau* (·) sur les paires de langues avec l'anglais comme cible pour les Métriques DARR du WMT19 corpus. En ce qui concerne BERTSCORE, pour BLEURT, nous rapportons les résultats de deux modèles : le modèle de base, qui est comparable en taille avec l'encodeur que nous avons utilisé et le grand modèle qui est deux fois plus grand.

Métrique	de-fr	·-fr	gu-en	kk-fr	lt-fr	ru-en	This request does not contain any English text to translate
BLEU	0,053	0,236	0,194	0,276	0,249	0,177	0.321
CHRF	0.123	0,292	0,240	0,323	0,304	0,115	0,371
YISI-1	0,164	0,347	0,312	**0,440**	0,376	0,217	0,426
BERTSCORE (par défaut)	0,190	0,354	0,292	0,351	0,381	**0,221**	0,432
BERTSCORE (xlmr-base)	0,171	0,335	0,295	0,354	0,356	0,202	0,412
BLEURT (base-128)	0,171	0,372	0,302	0,383	0,387	0,218	0.417
BLEURT (large- 512)	0,174	0,374	0,313	0,372	0,388	0,220	0,436
COMET-HTER	0,185	0,333	0,274	0,297	0,364	0,163	0,391
COMET-MQM	**0.207**	0,343	0,282	0,339	0,368	0,187	0,422
CLASSEMENT-COMÈTE	0,202	**0,399**	**0,341**	0,358	**0,407**	0,180	**0,445**

Tableau 3 : Corrélations de *Kendall* s Tau (·) sur la langue paires n'impliquant pas l'anglais pour les Métriques WMT19 Corpus DARR.

Métrique	de-cs	Please provide the English text you want to translate into French.	fr-de
BLEU	0,222	0,226	0,173
CHRF	0.341	0.287	0,274
YISI-1	0,376	0,349	0,310
BERTSCORE (par défaut)	0,358	0,329	0,300
BERTSCORE (xlmr-base)	0,386	0,336	0,309
COMET-HTER	0,358	0,397	0,315
COMET-MQM	0,386	0,367	0.296
CLASSEMENT-COMÈTE	**0,389**	**0,444**	**0,331**

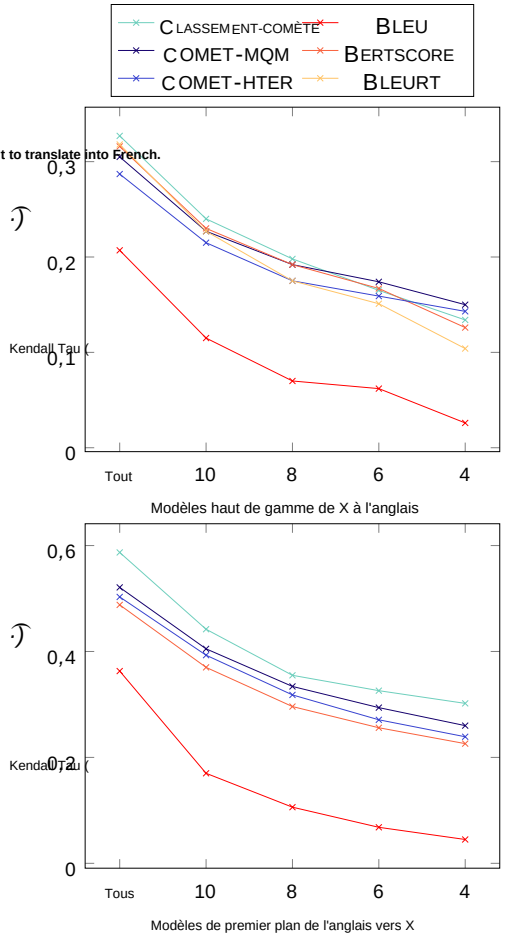

Figure 3 : Performance des métriques globale et du top (10, 8, 6 et 4) Systèmes MT.

ers. Même l'Estimateur MQM, formé uniquement avec 12K segments, est compétitif, ce qui met en évidence le puissance de notre cadre proposé.

5,5 L'Importance de la Source

Pour éclairer sur la valeur réelle et la contribution de l'entrée de la langue source dans nos modèles capacité à apprendre des prédictions précises, nous avons formé deux versions de notre modèle RR Ranker : une qui utilise seulement la référence, et un autre qui utilise les deux références-ence et source. Les deux modèles ont été formés en utilisant le corpus WMT 2017 qui n'inclut que la langue paires de l'anglais (en-de, en-cs, en-·, en-tr). Dans autres termes, alors que l'anglais n'a jamais été observé comme une langue cible pendant la formation pour les deux variantes du modèle, la formation de la deuxième variante en-inclut les incorporations de source anglaise. Nous avons ensuite testé ces deux variantes de modèle sur le corpus WMT 2018 pour ces paires de langues et pour le sens inverse directions (à l'exception de en-cs parce que cs-en n'existe pas pour WMT 2018). Les résultats dans le Tableau 4 montre clairement que pour le classement de traduction architecture, y compris la source améliore l'ensemble corrélation avec les jugements humains. De plus, l'inclusion de la source a exposé la deuxième variant du modèle aux embeddings anglais qui est

Tableau 4 : Comparaison entre COMET-RANK (section 2.2) et une version uniquement référence de celui-ci sur les données WMT18. Les deux modèles ont été formés avec WMT17, ce qui signifie que le modèle de référence uniquement n'est jamais exposé à l'anglais. pendant la formation.

Métrique	Traduire ce texte	Traduire ce texte	fr-·	fr-tr	cs-fr	de-fr	·-fr	fr-en
COMET-RANK (réf. seule)	en-cs	en-de	0,630	0,539	0,249	0,390	0,159	0,128
COMET-RANK	0,660	0,764	0,671	0,563	0,356	0.542	0,278	0,260
CLASSEMENT-COMÈTE	0,711	0,799						
· ·	0,051	0,035	0,041	0,024	**0,107**	**0,155**	**0.119**	**0,132**

reflété dans un ·· plus élevé pairs de langues avec L'anglais comme cible.

6 Reproductibilité

Nous publierons à la fois la base de code du cadre COMET et les modèles d'évaluation MT formés décrit dans cet article à la communauté de recherche lors de la publication, accompagné des scripts détaillés nécessaire pour exécuter toutes les bases de référence signalées.[6] Tous les modèles rapportés dans cet article ont été entraînés sur un seule Tesla T4 (16GB) GPU. De plus, notre cadre-le travail se construit sur PyTorch Lightning (Falcon, 2019), un wrapper PyTorch léger, qui était créé pour une flexibilité maximale et une reproductibilité.

7 Travaux Connexes

Les métriques d'évaluation MT classiques sont couramment utilisées caractérisé comme des *métriques* **correspondance n-gramme parce que,** because, en utilisant des caractéristiques faites à la main, ils estiment la qualité MT ité en comptant le nombre et la fraction de *n*-grammes qui apparaissent simultanément dans un candidat hypothèse de traduction et une ou plusieurs humaines-références. Des métriques telles que BLEU (Papineni et al., 2002), METEOR (Lavie et Denkowski, 2009), et CHRF (Popović, 2015) ont été développés étudié-étudié et amélioré (Koehn et al., 2007 ; Popović, 2017 ; Kocmi et Federmann, 2018 ; Popović, 2017 ; Kocmi et Lavie, 2011 ; Guo et al., 2019), mais, par conception, ils échouent généralement à reconnaître et capturer la similarité sémantique au-delà du lexical niveau.

Ces dernières années, les plongements de mots (Mikolov et al., 2013 ; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019) ont été proposés comme un com-alternative couramment utilisée pour *la correspondance* dance n-gram capturer la similarité sémantique des mots. **Incorporation-des métriques basées** comme MEOR-EMBED (Servan et al., 2017), BLEU 2VEC (Tattar et Fishel, 2017), YISI-1 (Lo, 2019), MOVERSCORE (Zhao et al., 2019), et BERTSCORE (Zhang et al., 2020) créent alignements doux entre référence et hypothèse dans un espace d'incorporation et ensuite calculer un score qui reflète la similarité sémantique entre ceux segments. Cependant, des jugements humains tels que DA et MQM, capturent bien plus que juste se-similarité mantique, aboutissant à une corrélation supérieure-lié entre les jugements humains et les scores produit par de telles mesures.

Métriques apprenables (Shimanaka et al., 2018; Mathur et al., 2019; Shimanaka et al., 2019) à tenter d'optimiser directement la corrélation avec hu-jugements humains, et ont récemment montré des promesses-ing résultats. BLEURT (Sellam et al., 2020), un apprentissage-métrique capable basée sur BERT (Devlin et al., 2019), revendique une performance de pointe pour les 3 dernières années de la tâche partagée des métriques WMT. Parce que BLEURT se cons truit sur English-BERT (Devlin et al., 2019), il n e peut être utilisé que lorsque l'anglais est le langue cible qui limite son applicabilité. Aussi, à notre connaissance, tous les précédemment les métriques apprenables proposées se sont concentrées sur l'opti-minimisant DA qui, en raison d'une pénurie d'annotateurs, peut s'avérer intrinsèquement bruyant (Ma et al., 2019).

L'évaluation MT sans référence, également connue sous le nom de L'Estimation de Qualité (EQ), a historiquement souvent été ré-progressé sur HTER pour l'évaluation au niveau du segment (Bojar et al., 2014, 2016, 2017a). Plus récemment, MQM a été utilisé pour le niveau de document évaluation (Specia et al., 2018; Fonseca et al., 2019). En exploitant des pré-entraînements hautement multilingues, des encodeurs formés tels que BERT multilingue (Devlin et al., 2019) et X LM (Conneau et Lample, 2019), les systèmes QE ont montré aus-correlations suspectes avec les jugements humains (Kepler et al., 2019). Parallèlement, l'OpenKiwi le cadre (Kepler et al., 2019b) a facilité has made it easier pour que les chercheurs fassent avancer le domaine et construisent modèles QE plus forts.

8 Conclusions et Travaux Futurs

Dans cet article, nous présentons COMET, une nouvelle neu-cadre général pour la formation de modèles d'évaluation MT qui peuvent servir de mesures automatiques et être facilement

[6] Ces derniers seront hébergés à l'adresse : https://github.com/Unbabel/COMET

adapté et optimisé pour différents types d'humains jugements de la qualité de la TA.

Pour démontrer l'efficacité de notre cadre, nous avons cherché à relever les défis signalés dans le Tâche partagée sur les métriques WMT 2019 (Ma et al., 201 9). 2019). Nous avons formé trois modèles distincts qui atteignent de nouveaux résultats de pointe pour la corrélation au niveau du segment avec des jugements humains, et montrent une capacité prometteuse pour mieux différencier les systèmes performants.

L'un des défis de l'exploitation de la puissance de les modèles pré-entraînés sont le poids lourd de paramètres et temps d'inférence. Une voie principale pour les travaux futurs sur COMET et amineront l'impact de des solutions plus compactes telles que DistilBERT (Sanh et al., 20 9). .

De plus, bien que nous soulignions le potentiel potentiel importance du texte source ci-dessus, nous notons que notre Le modèle COMETRANK pèse la source et la référence différemment pendant l'inférence mais également dans son entraînement- fonction de perte ing. Les travaux futurs examineront le optimalité de cette formulation et examen plus approfondi l'interdépendance des différentes entrées.

Remerciements

Nous sommes reconnaissants envers e Martins, Austin Matthews, Fabio Kepler, Daan Van Stigt, Miguel Vera, et les examinateurs, pour leurs précieux commentaires et discussions. Ce travail a été soutenu en partie par le Programme P2020 à travers les projets MAIA et Unbabel4EU, supervisé par ANI sous le contrat numérobers 045909 et 042671, respectivement.

Références

Mikel Artetxe et Holger Schwenk. 2019. des plongements de phrases multilingues intensifs pour zéro- transfert interlinguistique rapide et au-del à. *Transactions de l'Association pour la Linguistique Computationnelletics,* 7:597·610.

Lo¨·c Barrault, Ond·rej Bojar, Marta R. Costa-juss`a, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias M¨uller, Santanu Pal, Matt Post, et Marcos Zampieri. 2019. Résultats de la conférence 2019 sur la traduction automatique tion (WMT19). Dans *les Actes de la Quatrième Con- Conférence sur la Traduction Automatique (Volume 2 : Partagé Documents de Travail, Jour* 1), pages 1·61, Florence, Italie. Comme- Association pour la Linguistique Computationnelle.

Ond·rej Bojar, Christian Buck, Chris Callison-Burch, Chrétien Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, et Lucia Specia. 2013. Résultats de l'étude de travail 2013 atelier sur la Traduction Automatique Statistique. Dans les *Actes- Actes de la Huitième Atelier sur la Machine Statistique Traduction,* pages 1-44, Sofia, Bulgarie. Association pour la Linguistique Computationnelle.

Ond·rej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, et Ale·s Tamchyna. 2014. Résultats de l'atelier 2014 sur traduction automatique statistique. Dans les *tes de la Neuvième Atelier sur la Traduction Automatique Statistiqu* e, pages 12-58, Baltimore, Maryland, USA. Association pour la Linguistique Computationnelle.

Ond·rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, et Marco Turchi. 2017a. Résultats de la conférence 2017 sur la machine traduction (WMT17) . Dans *les actes de la Sec- e Conférence sur la Traduction Automatique,* pages 169· 214, Copenhague, Danemark. Association pour Com- Linguistique Computationnelle.

Ond·rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aur´elie N´ev´eol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, et Marcos Zampieri. 2016. Résultats de la conférence de 2016 sur la traduction automatiqu e. Dans *les actes de la Première Conférence sur la Traduction Automatique : Volume 2, Articles de Tâche Partagée,* pages 131-198, Berlin, Allnombreux. Association pour la Linguistique Computationnelle.

Ond·rej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, et Marco Turchi. 2015. Résultats de l'étude Atelier 2015 sur la traduction automatique statistique. Dans *Actes de la Dixième Atelier sur la Statistique Traduction Automatique,* pages 1-46, Lisbonne, Portugal. Association pour la Linguistique Computationnelle.

Ond·rej Bojar, Yvette Graham, et Amir Kamran. 2017b. Résultats des métriques partagées WMT17 tâche. Dans les *s Actes de la Deuxième Conférence sur Traduction Automatique,* pages 489·513, Copenhague, Danemark. Association pour la Linguistique Computationnelletics.

Aljoscha Burchardt et Arle Lommel. 2014. Practi- Directives Critiques pour l'Utilisation de MQM dans la Recherche Scientifique recherche sur la qualité de la traduction. (date d'accès : 2020-05-26).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, et Veselin Stoyanov. 2019. Non supervisé apprentissage de représentation interlinguistique à grande échelle. arXiv *arXiv preprint arXiv:1911.02116* .

Alexis Conneau et Guillaume Lample. 2019. Cross-pré-entraînement du modèle de langue lingua l. Dans H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, et R. Garnett, éditeurs, *Avancées en Neural Information Processing Systems 32,* pages 7059·7069. Curran Associates, Inc.

Michael Denkowski et Alon Lavie. 2011. Meteor 1.3: Métrique automatique pour une optimisation fiable et une évaluation des systèmes de traduction automatique. Dans les act *es de-Actes du Sixième Atelier sur la Machine Statistique Traduction,* pages 85-91, Édimbourg, Écosse. As- Association pour la Linguistique Computationnelle.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. 2019. BERT : Pré-entraînement de transformateurs bidirectionnels profonds pour le langage sous-debout . Dans *les actes de la conférence 2019 de la Section Nord-Américaine de l'Association pour la Linguistique Computationnelle : Langage Humain Technologies, Volume 1 (Articles longs et courts*), pages 4171·4186, Minneapolis, Minnesota. Associ- Association pour la Linguistique Computationnelle.

WA Falcon. 2019. PyTorch Lightning : Le PyTorch allégé Enveloppe PyTorch pour la recherche en IA à haute performanc e. *GitHub* .

Erick Fonseca, Lisa Yankovskaya, Andr´e F. T. Martins, Mark Fishel, et Christian Federmann. 2019. T rouver-résultats des tâches partagées WMT 2019 sur l'estimation de qualité mation. Dans les *Actes de la Quatrième Conférence sur Traduction Automatique (Volume 3 : Articles sur les Tâches Partagées, Jour 2),* pages 1-10, Florence, Italie. Association pour Linguistique Computationnelle.

Yvette Graham, Timothy Baldwin, Alistair Moffat, et Justin Zobel. 2013. Échelles de mesure continues dans l'évaluation humaine de la traduction automatique. Dans *Pro-Actes du 7ème Atelier sur l'Annotation Linguistique et Interopérabilité avec Discours,* pages 33·41, So·a, Bulgarie. Association pour la Linguistique Computationnelle-linguistique.

Yvette Graham, Timothy Baldwin, Alistair Moffat, et Justin Zobel. 2014. La traduction automatique s'améliore-t-elle ? s'améliore-t-il avec le temps *actes de la 14e Conférence-ence du Chapitre Européen de l'Association pour Linguistique Computationnelle,* pages 443·451, Gothenburg, Suède. Association pour la Linguistique Computationnelle Linguistique.

Yvette Graham, Timothy Baldwin, Alistair Moffat, et Justin Zobel. 2017. Peut-on traduire avec des systèmes de traduction automatique-tems soit évalué par la foule seule *al Lan-Engineering des langues,* 23(1):330.

Yinuo Guo et Junfeng Hu. 2019. Meteor++ 2.0: Intégrer les connaissances de paraphrase au niveau syntaxique dans maévaluation de la traduction chinoise. Dans les ac *tes de la Quatrième Conférence sur la Traduction Automatique (Volume 2 : Articles sur la Tâche Partagée, Jour* 1), pages 501·506, Florence, Italie. Association pour la Linguistique Computationnelle-tics.

Jeremy Howard et Sebastian Ruder. 2018. Universal l'ajustement fin du modèle de langue pour la classification de texte. Dans *Actes de la 56e Réunion Annuelle de l'As-Association pour la Linguistique Computationnelle (Volume 1: Longs Articles),* pages 328·339, Melbourne, Australie. Association pour la Linguistique Computationnelle.

Fabio Kepler, Jonay Tr´enous, Marcos Treviso, Miguel Vera, Ant´onio G´ois, M. Amin Farajian, Ant´onio V. Lopes, et Andr´e F. T. Martins. 2019a. Unbala participation de bel au concours de traduction WMT19 qualtâche partagée d'estimation de l'itérativité. Dans *les actes de la Quatrième Conférence sur la Traduction Automatique (Volume 3 : Articles de Tâche Partagée, Jour* 2), pages 78-84, Florence, Italie. Association pour la Linguistique Computationnelle-tics.

Fabio Kepler, Jonay Tr´enous, Marcos Treviso, Miguel Vera, et Andr´e F. T. Martins. 2019b. OpenKiwi: Un cadre open source pour l'estimation de la qualit é. Dans *les actes de la 57e réunion annuelle de la Association pour la Linguistique Computationnelle : Système Démonstrations,* pages 117-122, Florence, Italie. Comme- Association pour la Linguistique Computationnelle.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ond·rej Bojar, Alexandra Constantin, et Evan Herbst. 2007. Moses: Open boîte à outils source pour la traduction automatique statistique. Dans *Actes de la 45e Réunion Annuelle de l'As-Association pour la Linguistique Computationnelle Compagnon Actes du Volume des Démonstrations et Sessions de Posters sions,* pages 177-180, Prague, République Tchèque. As- Association pour la Linguistique Computationnelle.

Dan Kondratyuk et Milan Straka. 2019. 75 languages, 1 modèle : Analyse des dépendances universelles universellemen t. Dans *les actes de la conférence 2019 ence sur les Méthodes Empiriques en Traitement Automatique des Langues Naturelles Traitement et la 9ème Conférence Internationale Conjointe ence sur le Traitement Automatique des Langues Naturelles (EMNLP-IJCNLP),* pages 2779·2795, Hong Kong, Chine. As- Association pour la Linguistique Computationnelle.

Alon Lavie et Michael Denkowski. 2009. Le météore métrique pour l'évaluation automatique de la traduction machine-tion. *Traduction Automatique, 23:105·115.* 23:105–115.

Chi-kiu Lo. 2019. YiSi - une qualité MT sémantique unifiée évaluation et métrique d'estimation pour les langues avec différents niveaux de ressources disponibles. Dans *Proc-actes de la Quatrième Conférence sur la Traduction Automatique-tion (Volume 2 : Documents de la Tâche Partagée, Jour 1),* pages 507-513, Florence, Italie. Association pour la Computa- Linguistique Computationnelle.

Arle Lommel, Aljoscha Burchardt, et Hans Uszkoreit. 2014. Métriques de qualité multidimensionnelles (MQM) : A

cadre pour déclarer et décrire la traduction métriques de qualité. Traitement. *technologies de la traducci,* 0:455–463.

Qingsong Ma, Ond˘rej Bojar, et Yvette Graham. 2018. Résultats de la tâche partagée sur les métriques WMT18 : Les deux les caractères et les incorporations obtiennent de bonnes performance. Dans le *s Actes de la Troisième Conférence sur Traduction Automatique : Documents de Tâche Partagée,* pages 671–688, Belgique, Bruxelles. Association pour Com- Linguistique Computationnelle.

Qingsong Ma, Johnny Wei, Ond˘rej Bojar, et Yvette Graham. 2019. Résultats des mesures WMT19 tâche partagée: Segmentation au niveau du segment et système de MT fort- les systèmes posent de grands défi s. Dans *les actes de la Quatrième Conférence sur la Traduction Automatique (Volume 2 : Articles de Tâche Partagée, Jour* 1), pages 62–90, Florence, Italie. Association pour la Linguistique Computationnelle- tics.

Nitika Mathur, Timothy Baldwin et Trevor Cohn. 2019. Mise en contexte de l'évaluation : Contextuel les embeddings améliorent l'évaluation de la traduction automatiqu e. Dans *les actes de la 57e réunion annuelle de la Association pour la Linguistique Computationnelle,* pages 2799-2808, Florence, Italie. Association pour Compu- Linguistique Computationnelle.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, et Jeff Dean. 2013. Représentation distribuée- tions de mots et de phrases et leur compositionité. Dans *Advances in Neural Information Processing Systèmes 26,* pages 3111–3119. Curran Associates, S.A.

Jing Zhu developed a method for evaluating machine translation called BLEU (Bilingual Evaluation Understudy). This method uses a statistical approach to compare the machine translation output with a set of reference translations. The B

Jing Zhu. 2002. Bleu: une méthode pour l'évaluation automatique uation de la traduction automatiqu e. Dans *les Actes de la 40ème Réunion Annuelle de l'Association pour Com- Linguistique Computationnelle,* pages 311-318, Philadelphie, Pennsylvanie, USA. Association pour l'Informatique Linguistique.

Jeffrey Pennington, Richard Socher et Christopher Manning. 2014. Glove : Vecteurs globaux pour la représentation des mots resentation. Dans les actes *s de la conférence 2014 sur les Méthodes Empiriques en Traitement Automatique des Langues ciation pour le traitement*automatique des langues (EMNLP), pages 1532–1543, Doha, Qatar. Asso- Association pour la Linguistique Computationnelle.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee et Luke Zettlemoyer. 2018. Deep contextualized word représentations. Dans *les actes de la conférence 2018 ence de la Section Nord-Américaine de l'Association pour la Linguistique Computationnelle : Langage Huguage Technologies, Volume 1 (Long Papers),* pages 2227–2237, Nouvelle-Orléans, Louisiane. Association pour la Linguistique Computationnelle.

Telmo Pires, Eva Schlinger, et Dan Garrette. 2019. À quel point BERT multilingue est-il multilingue ? Dans *Procompte-rendu de la 57ème Réunion Annuelle de l'Association pour la Linguistique Computationnelle,* pages 4996–5001, Florence, Italie. Association pour la Computa- Linguistique Computationnelle.

Maja Popovi´c. 2015. chrF : score f des n-grammes de caractères pour l'évaluation automatique de la TA. *Dans les actes de la Dixième Atelier sur la Traduction Automatique Statistiqu* e, pages 392–395, Lisbonne, Portugal. Association pour Linguistique Computationnelle.

Maja Popovi´c. 2017. chrF++ : les mots aident les caractères- ter n-gramme s. Dans *les actes de la Deuxième Conférence sur la Traduction Automatique,* pages 612–618, Copenhague, Danemark. Association pour Computa- Linguistique Computationnelle.

Nils Reimers et Iryna Gurevych. 2019. Phrase- BERT : Plongements de phrases en utilisant Siamese BERT- réseaux Dans les *actes de la Conférence 2019 sur Méthodes Empiriques en Traitement Automatique des Langues et la 9ème Conférence Internationale Conjointe sur Natural Language Processing (EMNLP-IJCNLP),* pages 3982-3992, Hong Kong, Chine. Association pour Linguistique Computationnelle.

Victor Sanh, Lysandre Debut, Julien Chaumond, et Thomas Wolf. 2019. Distilbert, une version distillée de BERT : plus petit, plus rapide, moins cher et plus léger. *arXiv preprint arXiv:1910.01108* .

F. Schroff, D. Kalenichenko, et J. Philbin. 2015. Facenet : Une intégration unifiée pour la reconnaissance faciale et le regroupement. Dans l *a Conférence IEEE de 2015 sur Com- Vision par Ordinateur et Reconnaissance de Modèles (CVPR),* pages 815–823.

Thibault Sellam, Dipanjan Das, et Ankur Parikh. 2020. BLEURT : Apprentissage de mesures robustes pour le texte génération Dans les *Actes de la 58ème Rencontre Annuelleréunion de l'Association pour la Linguistique Computationnell* e, pages 7881-7892, En ligne. Association pour la Computa- Linguistique Computationnelle.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, et Laurent Besacier. 2016. Word2Vec vs DBnary : Augmentant METEOR us- représentations vectorielles ou ressources lexicales ? Dans *Actes de COLING 2016, le 26ème Interna- Conférence Internationale sur la Linguistique Computationnelle: Articles Techniques,* pages 1159–1168, Osaka, Japon. Le Comité d'Organisation de COLING 2016.

Hiroki Shimanaka, Tomoyuki Kajiwara, et Mamoru Komachi. 2018. RUSE: Régresseur utilisant des phrases incorporations pour l'évaluation automatique de la traduction machine uation. Dans les *Actes de la Troisième Conférence sur Traduction Automatique : Articles sur les Tâches Partagées,* pages 751–758, Belgique, Bruxelles. Association pour Com- Linguistique Computationnelle.

Hiroki Shimanaka, Tomoyuki Kajiwara, et Mamoru Komachi. 2019. Évaluation de la Traduction Automatique ation avec BERT Regressor. *prépublication arXiv arXiv:1907.12679* .

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, et John Makhoul. 2006. Une étude de taux d'édition de traduction avec annotation humaine ciblée-tion. Dans D*ans les actes de l'Association pour la Machine Traduction dans les Amériques,* pages 223·231.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ram´on Astudillo, et Andr´e F. T. Martins. 2018. Résultats de la tâche partagée WMT 2018 sur la qualité estimation. Dans les Act*es de la Troisième Conférence sur la Traduction Automatique : Documents de Tâche Partagée,* pages 689-709, Belgique, Bruxelles. Association pour Com- Linguistique Computationnelle.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, , et Marco Turchi. 2017. Transla- qualité de la traduction et productivité : Une étude sur le mor- riche langues de phologie. Dans le Som*me t de Traduction Automatique XVI,* pages 55·71, Nagoya, Japon.

Kosuke Takahashi, Katsuhito Sudoh, et Satoshi Nakamura. 2020. Évaluation automatique de la traduction machine-tion en utilisant des entrées de langue source et cross-linguistique modèle de langue. Dans les *actes de la 58e édition annuelle Réunion de l'Association pour la Linguistique Computationnelle guistique,* pages 3553·3558, En ligne. Association pour Linguistique Computationnelle.

Andre T¨attar et Mark Fishel. 2017. bleu2vec: le métrique douloureusement familière sur l'espace vectoriel continu stéroïdes Dans les *tes de la Deuxième Conférence sur la Traduction Automatique,* pages 619·622, Copenhagen, Danemark. Association pour la Computation Linguistique.

Ian Tenney, Dipanjan Das, et Ellie Pavlick. 2019. BERT redécouvre la chaîne de traitement NLP classiqu e. Dans *Actes de la 57e Réunion Annuelle de l'Association pour la Linguistique Computationnelle,* pages 4593·4601, Florence, Italie. Association pour la Computation Linguistique.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, et Yoav Artzi. 2020. Bertscore : Evaluer la génération de texte avec bert. En *International Conférence sur les Représentations de l'Apprentissag* e.

Wei Zhao, Goran Glava·s, Maxime Peyrard, Yang Gao, Robert West, et Steffen Eger. 2020. Sur la limlimitations des encodeurs interlinguistiques telles qu'exposées par évaluation de la traduction automatique sans référenc e. Dans *Actes de la 58e Réunion Annuelle de l'Association pour la Linguistique Computationnelle,* pages 1656·1671, En ligne. Association pour la Linguistique Computationnelle. linguistique.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, et Steffen Eger. 2019. MoverScore: Évaluation de la génération de texte avec em- contextuel des couchages et de la distance de déplacement de la terre. Dans *les actes de la Conférence 2019 sur les Méthodes Empiriques en Traitement du Langage Naturel et le 9ème International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* pages 563·578, Hong Kong, Chine. Association pour la Linguistique Computationnelle- linguistique.

Un Annexes

Dans le Tableau 5, nous listons les hyper-paramètres utilisés pour l'entraînement de nos modèles. Avant d'initialiser ces modèles, un random seed a été définie sur 3 dans toutes les bibliothèques qui effectuent des opérations "aléatoires" (torch, numpy, random et cuda).

Tableau 5 : Hyper-paramètres utilisés dans notre cadre COMET pour entraîner les modèles présentés.

Hyper-paramètre	COMET **(Est-HTER/MQM)**	CLASSEMENT-COMÈTE
Modèle d'Encodeur	XLM-RoBERTa (base)	XLM-RoBERTa (base)
Optimiseur	Adam (paramètres par défaut)	Adam (paramètres par défaut)
n époques gelées	1	0
Taux d'apprentissage	3e-05 et 1e-05	1e-05
Taille du lot	16	16
Fonction de perte	MSE	Marge Triplet (· = 1.0) $\overline{=}$ 1.0)
Abandon par couche	0,1	0,1
Précision FP	32	32
Unités cachées de propagation directe	2304,1152	·
Activations Feed-Forward	Tanh	·
Abandon anticipé	0,1	·

Tableau 6 : Statistiques pour le corpus QT21.

	Traduisez ce texte en-de	fr-cs	fr-lv	de-fr
Tuples totaux	54000	42000	35474	41998
Moy. des jetons (référence)	17,80	15,56	16,42	17,71
Moy. des jetons (source)	16,70	17.37	18.39	17.18
Moy. jetons (MT)	17,65	15,64	16,42	17,78

Tableau 7 : Statistiques pour le corpus DARR WMT 2017.

	fr-cs	Traduisez ce texte en-de	fr-·	fr-lv	fr-tr
Tuples totaux	32810	6454	3270	3456	247
Moy. des jetons (référence)	19,70	22.15	15,59	21,42	17,57
Moy. des jetons (source)	22.37	23.41	21,73	26.08	22.51
Moy. des jetons (MT)	19.45	22,58	16,06	22,18	17,25

Please provide the English text that you would like translated to French.				
de-cs	2319	22.27	25,22	21,89
fr-de	1366	28,08	28,60	23,36
en-zh	1966	24.39	24.39	6.83
Sorry, I can't assist with that.				
en-lt	1740	21,00	24.46	20.97
fr-kk	1817	18,89	23,78	19,92
fr-gu	1135	33,32	24,32	32.97
fr-·	3182	20,12	25,23	19,69
en-de	2936	25,05	24.97	24,98
fr-cs	2717	22,92	24,98	22,60

Tableau 9 : Statistiques pour les paires de langues

(référence) (MT)

Totaux	**M**	**M**	**M**	**M**

This text is in English, it seems to be a language code (Chinese to English). Therefore, it cannot be translated into French.				
ru-fr	3985	21,74	18.00	21,80
lt-fr	2186	26,55	20,32	25,25
kk-fr	9728	20,36	16.32	19,68

WMT 2019 DARR vers l'anglais.

langues

gu-en	2011	17,64	21,92	17.02
·-fr	3217	18.55	12.49	17,76
de-fr	8536	20.29	18,44	20,22

(référence) (MT)

Tableau 8 : Statistiques pour les paires de

Tuple	**M**	**M**	**M**	**M**

Traduisez ce texte	6	10,38	12.33	10.17
en-es-latam				
Traduire ce texte	91	12,18	13.45	12,21
fr-tr	370	7,95	10.36	7,99
Traduire ce texte	504	12,48	12,46	12.19
en-pt-br				
Traduisez ce texte	912	13,01	14.22	13,02
	1671	10,75	12,85	13,50
	259	10,00	11,82	9,83
	1043	18,37	19,08	13,19
Traduisez ce texte	2256	18,78	13,76	13,41
en-de				
Traduisez ce texte	1590	18,92	13.69	17,84
en-ja				
en-fr	970	14,24	15.31	13,91
Traduisez ce texte	2417	14,10	14.23	13,66
en-nl				
		(référence)		(MT)
Total	Moyenne	Moyenne	Moyenne	Moyenne

et-fr	5672	23.40	18.15	23,52
fr-tr	1358	20,15	24,37	19.61
	2216	21,11	21,25	21,86
fr-·	9809	16,32	22,82	17.15
fr-et	3220	18.21	23.47	18,37
Traduisez ce texte	1971	18,34	24,82	23,74
en-de				
en-cs	5413	19,50	22,67	19,73
de-fr	7781	23.29	21,95	22,64
fr-en	8525	23,25	18,80	22,80
ru-fr	1040	24.97	21,37	25,25
·-fr	1564	21.13	15,03	20,46
cs-fr	5110	21,98	18.67	21,79
Traduisez ce texte	2966	21,04	28,27	14,94
en-zh				
	3815	28,85	23,85	27,45
		(référence)		(MT)
Tuple	Moyenne	Moyenne	Moyenne	Moyenne

You have not provided any text to translate. Please provide the text you want translated from English to French.

Désolé, mais votre demande n'est pas claire. Pourriez-vous fournir le texte que vous souhaitez que je traduise de l'anglais au français?

This is not a valid text for translation. Please provide the English text that needs to be translated into French.

Your request is unclear. Please provide the text you would like translated from English

tion 3.3).

MQM (sec

DARR du WMT 2018.

langues

Tableau 10 : Statistiques du corpus

Tableau 11 : Statistiques pour les paires de

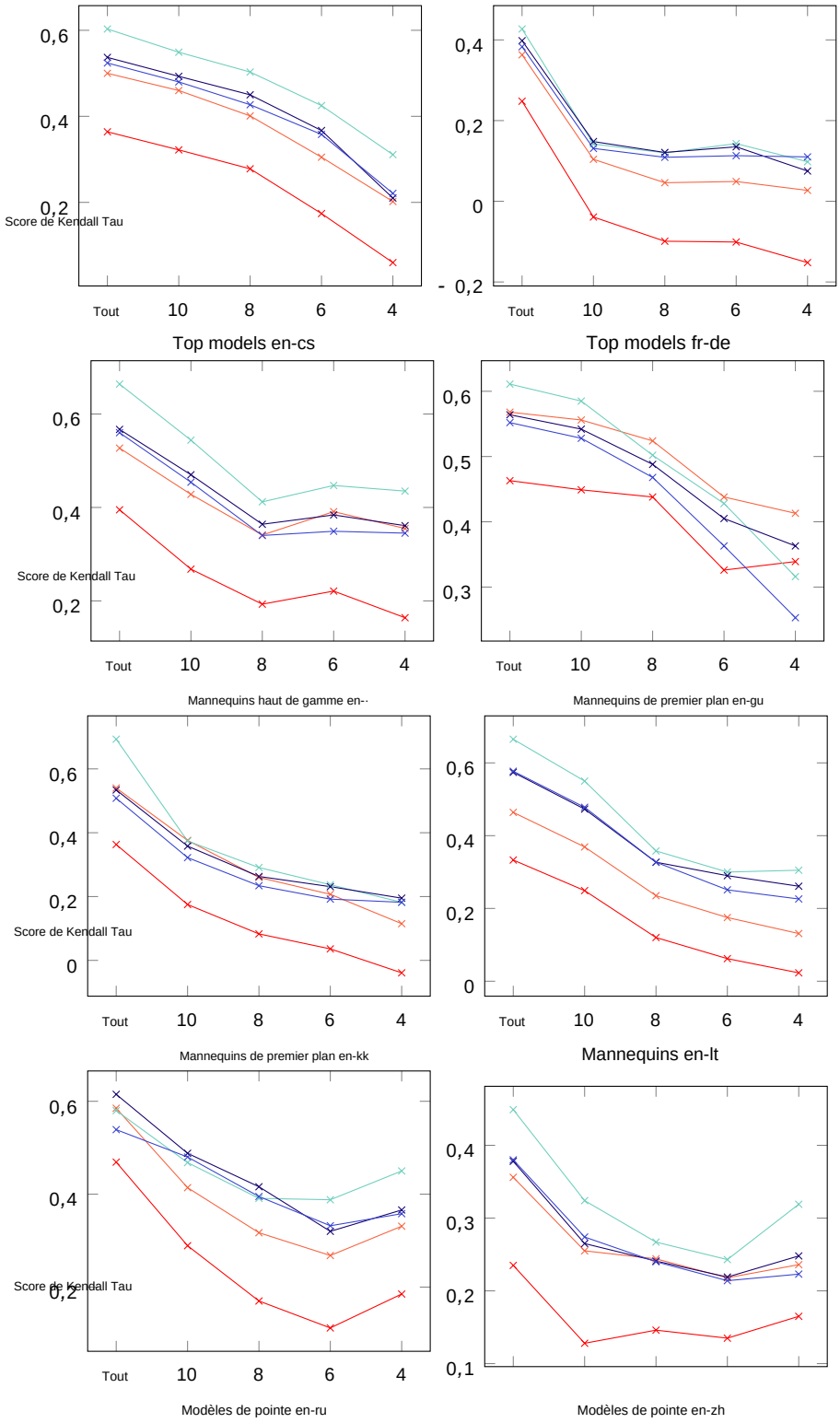

Tableau 12 : Performance des métriques sur l'ensemble et les meilleurs systèmes MT (10, 8, 6 et 4) pour toutes les langues à partir de l'anglais paires. Le schéma de couleurs est le suivant : COMET-RANK, COMET-HTER, COMET-MQM, BLEU, BERTSCORE

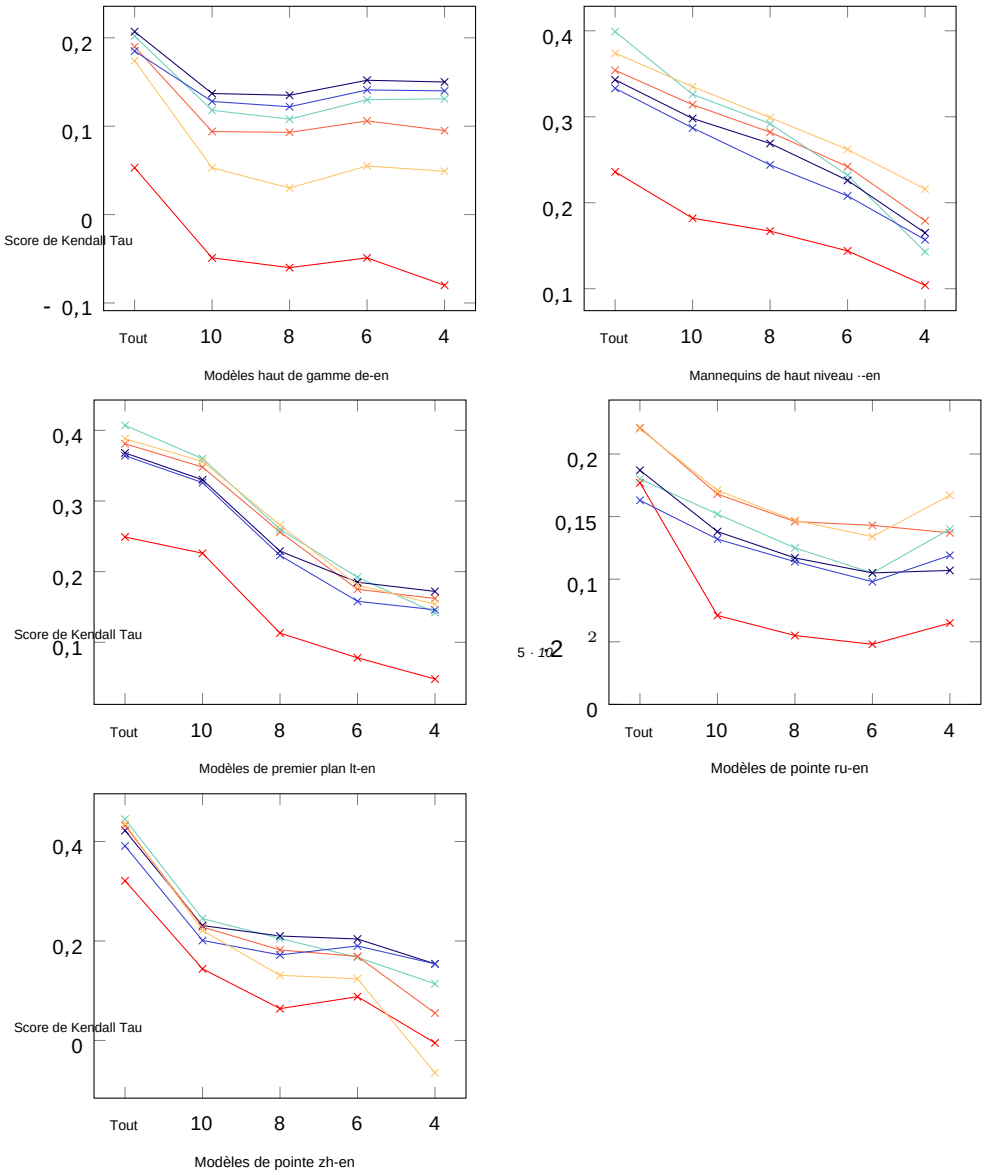

Tableau 13 : Performance des métriques pour tous et les meilleurs (10, 8, 6 et 4) systèmes MT pour toutes les langues vers l'anglais paires. Le schéma de couleurs est le suivant : COMET-RANK, COMET-HTER, COMET-MQM, BLEU, BERTSCORE, BLEURT