Evaluer, diagnostiquer et analyser la traduction automatique neuronale

François Yvon

Comprendre la traduction neuronale pour l'évaluation, le diagnostic et l'analyse.

par François Yvon*

Université Paris-Saclay, CNRS, LISN

Résumé Les outils de traduction automatique (TA) neuronale ont fait des progrès significatifs, qui les rendent utilisables pour un nombre croissant de domaines et de couples de langues. Cette évolution majeure des technologies de traduction invite à revisiter les méthodes de mesure de la qualité de la traduction, en particulier des mesures dites automatiques, qui jouent un rôle fondamental pour orienter les nouveaux développements de ces systèmes. Dans cet article, nous dressons un état des lieux des méthodes utilisées dans le cycle de développement des outils de traduction automatique, depuis les évaluations purement quantitatives jusqu'aux méthodologies récemment proposées pour analyser et diagnostiquer le fonctionnement de ces « boîtes noires »neuronales.

Abstract Neural machine translation (MT) technologies have made significant progress, making them useful for an increasing number of domains and language pairs. These major developments of translation technologies invite us to revisit our methods for measuring translation quality, in particular the so-called "automatic metrics", which play a fundamental role in guiding the new developments of MT systems. In this work, we review the methods used in the development cycle of machine translation tools, from purely quantitative evaluations to recently proposed methodologies aiming to analyse and diagnose the functioning of these neural "black boxes".

Mots-Clés : Traduction automatique neuronale ; Evaluation de la traduction automatique ; Métriques pour la traduction automatique

Keywords : Neural Machine Translation ; Machine Translation Evaluation ; Machine Translation Metrics

*Une version révisée de cet article est parue dans FORUM. *Revue internationale d'interprétation et de traduction*, Volume 20, Issue 2, Dec 2022, pp. 315–332 https://doi.org/10.1075/forum.00023.yvo.

Table des matières

1 Introduction

Dès les premières tentatives pour construire des outils de traduction automatique (TA) [Pierce et al., 1966], la question de l’évaluation objective des services réellement rendus par ces systèmes a été placée au centre des débats, et ces questions n’ont depuis jamais cessé d’animer les discussions entre les différents acteurs de ce domaine. Cette question ne peut être discutée qu’en relation avec les usages, les modes d’interaction et les attentes des utilisateurs de ces systèmes : traducteurs professionnels ou amateurs, lecteurs, auditeurs ou producteurs de documents dans langue étrangère, concepteurs de systèmes etc, et il est même douteux qu’un point de vue unifié sur ces questions puisse être dégagé [Hovy et al., 2002, Blanchon and Boitet, 2007, Castilho et al., 2018]. Soulignons également d’emblée qu’une évaluation complète des systèmes de TA ne peut être que multifactorielle de manière à prendre en compte par exemple la vitesse et le coût de traitement, la facilité de d’installation, d’adaptation et de personnalisation à des thèmes ou des des types de discours spécialisés, etc.

Dans cet article, ces aspects seront délibérément passés sous silence et nous nous focalisons principalement sur les problèmes d’évaluation de la

qualité de traduction qui se posent pour les concepteurs et développeurs de systèmes de TA, et qui, dans le cycle de conception et de développement, passent nécessairement par l'utilisation de méthodes automatiques d'évaluation. En effet, même s'il peut être tenu pour acquis [Freitag et al., 2021, Läubli et al., 2020] que seules des évaluations humaines, menées par des professionnels bien formés, permettra de recueillir des jugements précis et fiables sur la qualité des traductions ; il est également acquis que ces évaluations prennent du temps, qu'elles sont coûteuses à mettre en place, ce qui implique qu'il n'y soit fait appel que de manière parcimonieuse, avec une fréquence incompatible avec les besoins des développeurs. Des études récentes mettant en perspectives ces deux approches de l'évaluation en TA sont proposées dans [Castilho et al., 2018, Chatzikoumi, 2020, Balvet, 2020].

En effet, comme nous le rappelons à la section 2, la construction de systèmes neuronaux est principalement une question d'optimisation qui implique un ajustement itératif des multiples paramètres et méta-paramètres qui régulent le fonctionnement du système de TA, avec comme objectif d'atteindre la meilleure qualité de traduction possible. Au niveau le plus profond, l'optimisation des paramètres s'apparente à une procédure algorithmique d'essais-erreurs, guidée par de solides principes mathématiques, et demande, pour chaque réglage envisagé des paramètres, de calculer une mesure de qualité associée ce réglage. Ce cycle d'opérations d'ajustement et d'évaluation est réalisé des millions, voire des milliards de fois pour entraîner un système. Un second niveau d'optimisation consiste à comparer des configurations de méta-paramètres : par exemple de spécifier la taille du réseau neuronal, du vocabulaire qu'il considère, etc. Toutes ces opérations d'optimisation sont répétées quotidiennement. Dans la mesure où ces évaluations vont orienter les développements futurs dans les directions qui semblent les plus prometteuses, il importe qu'elles s'appuient sur des mesures qui évaluent correctement les propriétés désirées du système de TA : à quoi servirait une boussole qui n'indiquerait pas le Nord ?

Nous présentons et discutons dans cet article les protocoles et méthodes d'évaluation dont disposent les concepteurs de systèmes pour mesurer les performances de leurs modèles, pour les comparer avec d'autres modèles concurrents, pour enfin analyser leur fonctionnement. Comme nous le discutons ci-dessous, l'analyse diagnostique des systèmes neuronaux est rendue difficile par la complexité des calculs réalisés pour produire une traduction, et demande donc d'employer des outils idoines. Cette étude débute à la section 3 par la présentation des principales mesures d'évaluation globales, dont les plus usitées reposent sur une comparaison superficielle entre une traduction automatique et une ou plusieurs traductions humaines. Nous y présentons également les évolutions les plus récentes de ces métriques, avant d'introduire à la section 4 des approches plus diagnostiques, visant

à analyser les performances en utilisant des grilles d'analyse reposant sur des catégories linguistiques. À l'opposé de ces méthodes en boîte noire (*black box*), les approches en boîte blanche reposent sur une analyse des représentations internes du système de traduction. Les principales de ces approches sont présentées à la section 5.

2 La TA "Neuronale" : principes et concepts

2.1 Traduire par apprentissage

Les systèmes de traduction neuronale, apparus en 2014 [Cho et al., 2014] se sont rapidement imposés dans leurs différentes évolutions [Bahdanau et al., 2015, Gehring et al., 2017, Vaswani et al., 2017] comme les systèmes les plus performants pour une vaste gamme d'applications et d'usages. Ils s'inscrivent dans la continuité de la génération antérieure de systèmes, à base de segments [Koehn, 2010] et reposent principalement sur des méthodes d'apprentissage automatique. Dans leur version de base, ces modèles automatisent la traduction en la formalisant comme un algorithme, A, à qui on présente une phrase source F, et qui produit en réponse une phrase cible E, ce que l'on note $E = A(F)$. Le calcul réalisé par l'algorithme A est déterminé par une myriade de paramètres numériques, dénotés collectivement θ, qui se comptent souvent en millions, parfois en milliards. Cela signifie que chaque fois que l'on change la valeur d'un ou plusieurs paramètre(s) de θ, le résultat du calcul $A(F)$ est susceptible de changer. Pour expliciter cette dépendance, nous noterons dans la suite $E = A_\theta(F)$. L'entraînement du modèle consiste à trouver un réglage des paramètres qui conduit à des traductions correctes. Pour ce faire, on présente de manière répétée à A des exemples (F, E) de traductions humaines extraites de très grosses mémoires de traduction ou corpus parallèles. La sortie calculée $A_\theta(F)$ est alors comparée à la traduction attendue E, conduisant à un petit ajustement de θ, qui vise à corriger les écarts observés entre $A_\theta(F)$ (l'hypothèse) et E (la traduction correcte). En itérant cette procédure un grand nombre de fois, le comportement de A_θ se rapprochera du comportement souhaité, à savoir imiter les traductions humaines. Cette étape demande déjà de pouvoir évaluer la conformité de la traduction produite par rapport à la traduction souhaitée : dans la plupart des méthodes d'apprentissage, cette évaluation est stricte et l'on attend essentiellement de A_θ qu'il reproduise exactement, mot à mot les exemples proposés.

2.2 TAN : Traduction automatique numérique

Si l'on entre un peu plus au cœur du fonctionnement de A_θ, deux principes fondamentaux sont à l'oeuvre : * le calcul de $A_\theta(F)$ repose sur l'estimation

de scores numériques $S_\theta(F, E)$, eux aussi dépendants de θ, qui mesurent l'adéquation de chaque couple possible (E, F). $S_\theta(F, E)$ est grand si E est une traduction plausible de F, bas dans le cas contraire. L'algorithme A_θ consiste essentiellement à retrouver, parmi toutes les phrases possibles de la langue cible, la phrase E ayant le meilleur score pour un F donné. Retenons simplement que l'algorithme A_θ calcule une traduction et lui associe un score. * pour les besoins internes du calcul de A_θ, chaque mot source est converti dans un processus ''d'encodage"' en une représentation numérique (pour dire simple : un tableau de chiffres) ; de la même manière, chaque mot cible est également représenté sous une forme numérique et ce n'est qu'à la dernière étape du « décodage » qu'un mot linguistique est effectivement produit. Ce passage par des représentations internes numériques caractérise les approches neuronales, et est pour une large part responsable des gains de performances observés : il rend toutefois le fonctionnement de A_θ difficile à analyser puisqu'entre l'entrée et la sortie, tous les calculs réalisés portent sur des tableaux de chiffres qui n'ont plus rien à voir avec des entités linguistiques familières qu'il serait possible d'interpréter. D'autres principes sont importants pour bien comprendre ces modèles, par exemple le fait qu'ils traduisent les phrases mot par mot, de la gauche vers la droite, ou bien qu'ils s'appuient sur un pré-découpage rigide des mots source et cibles sous la forme de (petites) chaînes de caractères, constituant un vocabulaire d'unités fini, qui permet de décomposer (et traduire) y compris des mots inconnus. Nous ne reviendrons plus sur ces différents aspects, qui ne sont pas essentiels pour la suite de la présentation.

2.3 Configurer A_θ : le choix des méta-paramètres

Il existe une grande liberté pour concevoir et entraîner A_θ, et depuis 2014 plusieurs algorithmes ou familles d'algorithmes ont été successivement proposés et comparés. La famille d'algorithme la plus populaire repose sur le modèle Transformer [Vaswani et al., 2017], dont il est possible de spécifier de multiples variantes en choisissant le nombre de paramètres, la séquence de calculs impliqués par A_θ, la taille des représentations lexicales, le nombre d'entrées dans le vocabulaire, etc. La sélection des données et la préparation des corpus d'apprentissage font également partie des étapes critiques pour parvenir à bons résultats, et elles impliquent également de nombreuses décisions. Nous ferons référence à l'ensemble des éléments de configuration sous la dénomination générique de ''méta-paramètres". Plus que l'apprentissage lui-même, qui est essentiellement régi par des principes mathématiques qui le rendent complètement automatisable, ce sont des questions relatives à la conception et la sélection d'architectures algorithmiques pour implémenter A_θ, au réglage des méta-paramètres associés, ainsi qu'au choix et la préparation des corpus d'entraînement, qui occupe l'essentiel de l'activité des concepteurs de systèmes en TA. Pour effectuer

ce travail, il est essentiel de pouvoir facilement comparer la qualité des traductions produites par deux configurations ou deux algorithmes différents afin de sélectionner ou d'explorer plus avant l'alternative qui est la plus prometteuse.

En résumé, la performance observable des algorithmes de traduction automatique dépend du réglage d'un très grand nombre de paramètres numériques, ainsi que de multiples décisions relatives à la configuration de A_θ. Le fonctionnement de A_θ en est rendu totalement opaque et il devient bien plus difficile d'identifier de causes des succès ou des erreurs de la machine, comme de rechercher de nouvelles méthodes pour améliorer les performances. Pour comprendre les causes de succès comme les causes d'erreurs de la machine et rechercher des méthodes pour améliorer ses performances, il est donc nécessaire de développer de nouveaux outils d'analyse de ces algorithmes.

3 Évaluations automatiques : le rôle des références humaines

3.1

En TA comme dans d'autres domaines, la manière la plus naturelle d'évaluer la machine est de comparer ses productions avec des réalisations humaines. Une première manière d'évaluer ne s'intéresse ainsi qu'aux couples (entrée, sortie), pour vérifier leur conformité avec des traduction de référence.

3.2 Métriques automatiques : évaluations globales

Cette approche présuppose donc la capacité de comparer automatiquement $A_\theta(F)$ avec une traduction de référence réalisée par un professionnel $E^\star$ pour en déduire une mesure de qualité – ou "'métrique"' – $\Delta(A_\theta(F), E^*)$. La mise en {œ}uvre de cette idée a longtemps buté sur deux obstacles en apparence insurmontables. Le premier est la variabilité des traductions humaines, qui pose la question du choix particulier de $E^\star$: comment noter une machine qui imiterait parfaitement un traducteur, tout en se distinguant totalement d'un autre ? Le second est le choix de la fonction de comparaison, qui devra être telle que $\Delta(A_\theta(F), E^*)$ soit d'autant plus petite [1] que la qualité de la traduction est mauvaise. La métrique BLEU, introduite par [Papineni et al., 2002], essaie de répondre à ces questions. Elle s'appuie sur des comparaisons simplistes entre $A_\theta(F)$ et E^* (décompte du nombre de mots ou de groupes de mots communs), moyennées sur un grand nombre

1. Certaines métriques se fondent sur la similarité avec des des références, d'autres avec des dissimilarités. Nous nous plaçons dans le premier cas de figure.

de phrases ; elle autorise la prise en compte de plusieurs traductions de référence ; les mesures fournies semblent assez bien s'accorder avec des jugements de qualité humains. Une dernière propriété importante de BLEU est que son calcul est très rapide et n'implique aucune analyse linguistique des sorties de la TA : le même score peut donc être utilisé pour tous les couples de langues.

De multiples propositions alternatives, visant à corriger les (nombreux) travers de BLEU ont depuis été proposées : ainsi METEOR [Banerjee and Lavie, 2005], qui prend mieux en charge les possibles variations lexicales, ou TER (*Translation Edit Rate*) [Snover et al., 2006] qui cherche à mieux approximer le coût de la correction de $A_\theta(F)$. Il a en particulier été montré que METEOR, comme de nombreuses autres métriques, s'accorde mieux avec les évaluations humaines que BLEU quand on cherche à mesurer la qualité de phrases isolées.

La recherche de nouvelles métriques reste un domaine très actif : avec les progrès des technologies de TA, la capacité des métriques existantes à fournir des évaluations cohérentes doit être constamment revérifiée. Ces recherches ont récemment connu un renouveau avec l'apparition de métriques exploitant des représentations numériques calculées par des réseaux de neurones comme BertScore [Zhang et al., 2020], PRISM [Thompson and Post, 2020] ou COMET [Rei et al., 2020].

Malgré ces nouvelles propositions, le score BLEU est à ce jour indépassé et reste la méthode la plus utilisée pour présenter des résultats de recherche, pour orienter au quotidien les développements ou valider des choix techniques.

3.3 'Evaluer sans référence

L'utilisation de métriques à base de références humaines soulève plusieurs problèmes. Une première question pratique est celle de l'élaboration ces références et pour cela du choix des textes représentatifs de la tâche, ainsi que des traducteurs chargés de produire la traduction « idéale » à l'aune de laquelle les machines seront comparées. Aucun de ces choix ne va de soi, et influe grandement sur les mesures réalisées. Plus profondément, le calcul d'un score de performance indépendant des objectifs visés par la traduction semble très discutable : selon que la traduction doit être post-éditée, utilisée pour indexer des documents, ou pour soutenir une conversation en langue étrangère, le niveau de service ou d' '"utilisabilité" ' rendu par une même traduction sera très variable.

Une réponse à la première objection consiste à construire des mesures absolues de qualité qui se dispensent de référence, on parle également '"d'estimations de qualité" '. Il s'agit également d'un domaine de recherche

très actif, dont les objectifs sont toutefois très éloignés de la recherche de diagnostics. Nous renvoyons à [Specia et al., 2018] pour une introduction récente à ces approches.

La réponse à la seconde objection passe par la conception de protocoles et de métriques d'évaluation extrinsèques, fondés sur la tâche [Blanchon and Boitet, 2007], c'est-à-dire qui mesurent la manière dont une traduction automatique permettre d'atteindre l'objectif visé. Cette démarche ancienne est mise en avant par le groupe de travail EAGLES dont le rapport publié en 1996 [2] souligne « qu'il serait plus profitable de mesurer si une traduction est assez bonne pour un besoin spécifique, plutôt que d'essayer de définir une notion, nécessairement trop abstraite, de qualité d'une traduction en général. » (cité par [Blanchon and Boitet, 2007]). Dériver des métriques automatiques de ce principe demande de pouvoir formaliser la notion tâches et dévaluer automatiquement leur degré d'accomplissement. La métrique TER (Translation EDIT Rate) de Snover et al. [2006] illustre cette démarche et permet d'approximer une mesure du temps de post-édition. Avec les progrès des outils de traitement des langues, la mesure automatique d'autres tâches, en particulier de tâches de compréhension, semble envisageable [Scarton and Specia, 2016, Forcada et al., 2018, Krubiński et al., 2021].

4 À la recherche des failles de la TA

Les évaluations automatiques discutées à la section précédente, qu'elles utilisent ou non des traduction de référence, délivrent des mesures globales, qui ne permettent pas de diagnostiquer finement le comportement de la TA, ni de qualifier les erreurs qu'elle commet, en les associant à des catégories linguistiques interprétables : par exemple en distinguant les erreurs morphologiques aux erreurs syntaxiques. S'il existe des propositions solides pour établir ces diagnostics à partir de typologies d'erreurs [Kübler, 2008, Lommel et al., 2014], leur utilisation demande de mobiliser des experts humains pour annoter des traductions automatques et renvoie aux mêmes difficultés que toute évaluation humaine de la TA, la rendant impropre pour des évaluations internes.

Une autre manière de procéder, qui s'inspire des méthodes utilisées pour valider des logiciels, consiste à partir des difficultés supposées et à élaborer des banc d'essais (test suites) illustrant systématiquement chaque type de difficulté. L'analyse des erreurs par problème linguistique permet d'identifier les questions sur lesquelles la TA continue de buter. Cette approche ancienne [King and Falkedal, 1990], revisitée récemment par Burchardt et al. [2017] et Isabelle et al. [2017] est peu adaptée à des évaluations

2. https://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html

automatiques. D'une part, elle demande une double intervention d'experts humains : en amont pour concevoir les exemples difficiles à soumettre au système, en aval pour vérifier si la machine a correctement surmonté les obstacles. D'autre part, les résultats produits sont très dépendants de la structure du jeu de test, qui n'est pas toujours représentatif des difficultés linguistiques rencontrées dans des applications réelles. Enfin, ces bancs d'essais doivent constamment être mis à jour, car les systèmes de TA progressent continuellement, et cherchent à corriger les erreurs détectés par ces jeux de tests.

Dans la littérature récente, deux voies sont explorées pour aller vers l'automatisation de ces approches : la conception (manuelle) de jeux de tests spécialisés, associés à des métriques automatisables ; la génération « automatique » de bancs d'essais, couplés à des mesures automatiques du succès.

4.1 Des bancs d'essais spécialisés

Une difficulté bien identifiée des systèmes statistiques et neuronaux, qui traitent chaque phrase séparément des phrases précédentes, concerne les références anaphoriques, comme dans l'exemple suivant : * Mary bought a brand new bike for her daughter. She will take it to ride to school. The funeral of the Queen Mother will take place on Friday. It will be broadcast live. * Les funérailles de la reine-mère auront lieu vendredi. Elles seront retransmises en direct. Pour traduire correctement le pronom 'it' vers le français, il faut connaître le genre grammatical de son référent, qui ne peut être calculé qu'en prenant en compte la phrase précédente. Diverses méthodes existent pour traiter ces problèmes de contexte étendus [Maruf et al., 2021]. Pour évaluer le succès de ces méthodes [Guillou and Hardmeier, 2016] propose un banc d'essai de 250 examples. Ces évaluations sont ensuite systématisées dans [Hardmeier et al., 2015, Guillou et al., 2016], qui proposent de réutiliser des *benchmarks* génériques, en remplaçant les mesures globales de la qualité telles que BLEU par une mesure spécifique, qui ne prend en compte que la traduction des pronoms.

Une approche identique est proposée par les auteurs de [Rios et al., 2018] pour évaluer la capacité de désambiguisation sémantique des systèmes de traduction depuis l'allemand vers l'anglais : étant donné un ensemble de phrases contenant des noms polysémiques, la mesure de qualité ne considérera que la correction des traductions proposées pour ces mots ambigus, qui peut être évaluée de manière automatique. Plus récemment, l'atelier *Terminology* de la « *Conference on Machine Translation* » de 2021 s'est intéressé à l'évaluation de la capacité des systèmes de TA neuronale à prendre en considération des ressources terminologiques [Alam et al., 2021].

4.2 Évaluation par des manipulations linguistiques

4.2.1 Dans la phrase source

La boîte noire neuronale produit une phrase cible et un score numérique en réponse à une entrée en langue source. Une première manière d'aller vers le diagnostic consiste à observer l'effet sur la sortie de changements '"contrôlés"' dans l'entrée. Cette approche est illustrée par les travaux sur la '"robustesse"' des systèmes de TA de Belinkov and Bisk [2018], où les auteurs analysent l'effet de petits changements superficiels de la phrase source, censés simuler les erreurs typographiques. Formellement, on compare donc $A_\theta(F)$ et $A_\theta(F')$, où F' est dérivée automatiquement de F en insérant, permutant, supprimant, ou substituant aléatoirement des symboles : le comportement idéal serait que ces variations n'affectent pas la traduction, ou du moins l'affectent le moins possible. Un système de TA sera alors jugé robuste si les scores (BLEU, TER, etc) obtenus avant et après changement de l'entrée sont proches.

Cette méthode est étendue dans [Burlot and Yvon, 2017, 2018] qui considèrent des contrastes linguistiquement motivés entre F et F', consistant principalement à manipuler un trait morphologique : par exemple, changer le temps ou la modalité du verbe principal, le nombre du sujet, remplacer un nom par un pronom, etc. Un exemple donné par ces auteurs pour analyser la direction anglais-français est le suivant : * F = That is what will keep you alive. * F' = That is what would keep you alive. On vérifie alors que le verbe principal dans la traduction de F' est bien au conditionnel. Cette approche demande (a) de dériver automatiquement F' depuis F ; (b) de vérifier automatiquement que la traduction automatique de F' présente bien la variation morphologique désirée. Selon les paires de langues considérées, les deux étapes (a) et (b) peuvent être plus ou moins faciles à mettre en oeuvre : ainsi, modifier le temps du verbe principal en anglais est relativement aisé ; le faire en allemand est nettement plus difficile du fait des restructurations syntaxiques qui sont nécessaires. Comme le notent les auteurs, cette méthode présente l'avantage de s'appuyer sur la génération automatique de tests, que l'on peut produire en grande quantité et utiliser pour obtenir des mesures statistiques fiables de la capacité du système de TA à prendre en compte des variations linguistiques. Cette technique a été récemment utilisée dans de plusieurs travaux sur les biais de genre de la TA, en faisant varier le genre du sujet entre F et F' [Saunders and Byrne, 2020, Wisniewski et al., 2021a].

4.2.2 Dans la phrase cible

Une alternative, qui préserve l'idée d'exploiter des contrastes contrôlés, tout en évitant les difficultés de l'étape (b) de la méthode précédente, est

de comparer des contrastes en langue cible, en s'appuyant sur les scores $S_\theta(F, E)$ calculés par $A_\theta()$. Partant d'une phrase parallèle correcte (F, E), on construit un contraste délibérément incorrect (F, E'), et l'on compare $S_\theta(F, E)$ et $S_\theta(F, E')$. Un système de TA sera évalué positivement chaque fois qu'il préfère la traduction correcte à la traduction incorrecte, qui doit donc avoir un moins bon score ; dans le cas contraire, il sera évalué négativement. [Sennrich, 2017] déploie cette technique, appelée '"évaluation contrastive"', à grande échelle pour la direction de traduction anglais-allemand. Les contrastes considérés sont variés et consistent par exemple à manipuler un trait morphologique pour tester l'accord ou encore à changer la polarité (positif / négatif) d'une proposition. À titre d'illustration, l'auteur présente la paire de traduction suivante pour la phrase anglaise $F =$"Prague Stock Market falls to minus by the end of the trading day", dont le second terme du contraste contient un accord sujet / verbe incorrect. * E : Die Prager Börse stürzt gegen Geschäftsschluss ins Minus. * E' : Die Prager Börse stürzen gegen Geschäftsschluss ins Minus. Si le système, confronté à ce type de contrastes, marque une claire préférence pour le premier terme, on pourra conclure qu'il a « appris »à réaliser correctement les accords entre sujets et verbes.

Plus simple à mettre en oeuvre que la méthode précédente, applicable à grande échelle, cette approche a été réutilisée dans de nombreux contextes. Ainsi [Bawden et al., 2018] s'en inspire pour étudier la traduction des références pronominales ainsi que la cohérence des choix lexicaux, une question également étudiée par [Voita et al., 2019] tandis que [Raganato et al., 2019] propose d'engendrer des contrastes pour des mots sémantiquement ambigus pour de multiples paires de langues. Une limite importante de cette méthode, toutefois, est que l'évaluation ainsi réalisée n'est pas complètement réaliste, car on y compare des traductions idéales (E) ou manipulées (E'), qui ne sont pas nécessairement celles que la machine calculerait pour traduire F.

5 Sous le capot, le moteur (de traduction)

5.1 Analyse des représentations (sondes linguistiques)

Les méthodes présentées dans les sections précédentes sont des méthodes dites en « boîte noire »(black box), qui fondent leur analyse sur la simple étude des entrées et sorties du système de traduction. Les développeurs de systèmes ont souvent les moyens d'en savoir plus, en observant les « représentations internes » construites et calculées lors de la traduction d'une phrase. Comme rappelé à la section 2, les systèmes neuronaux calculent leurs traductions en transformant les mots et phrases en entrée sous la forme de tableaux de chiffres qui, d'une certaine manière, encodent

l'information que le réseau de neurones a extrait de la phrase source et qu'il utilise pour calculer la sortie.

Il est alors possible de chercher à analyser ou à visualiser ces représentations, afin de savoir si des propriétés linguistiques que l'on pense essentielles pour la traduction ont correctement été extraites. Nous nous focalisons dans la suite sur les méthodes d'analyse, en renvoyant pour les méthodes de visualisation aux références de [Belinkov and Glass, 2019]. Un cas d'école est le nombre du groupe sujet, qui doit souvent être transféré depuis la langue source vers la langue cible, et qui devrait donc être extrait et représenté sous forme numérique. Pour tester cette hypothèse, une manière de procéder s'appuie sur des tâches auxiliaires (on parle également de « sondes linguistiques »), qui consistent à (a) isoler les représentations internes associées à chaque phrase et (b) les utiliser indépendamment du système de traduction pour prédire le nombre du sujet. Illustrons ce principe par l'exemple suivant : F= 'Depuis 1881, les meilleurs joueurs de tennis du monde se sont illustrées sur les terrains canadiens, Nombre=Pluriel'. Comme expliqué ci-dessus, le calcul d'une traduction encode F sous une forme numérique, notons cette représentation $R(F)$. En recueillant les représentations associées à un nombre suffisant de phrases et en leur associant la valeur N du trait de Nombre, on obtient une base de données qui permet d'utiliser à nouveau l'apprentissage automatique, pour apprendre la relation entre $R(F)$ et le trait de Nombre. S'il est possible de retrouver N à partir de $R(F)$, c'est que cette information a bien été extraite.

Cette approche est une des principales approches en « boîte blanche » pour analyser les modèles neuronaux en traitement des langues [Hupkes et al., 2018, Belinkov and Glass, 2019] – Les premiers parlent de classifieurs diagnostiques (*diagnostic classifieurs*), les secondes de sonde (*probes*), terme que nous reprenons ici – et en particulier en traduction automatique. Elle est exemplairement utilisée par [Shi et al., 2016], qui s'intéresse à l'extraction d'informations syntaxiques et par [Conneau et al., 2018], qui montre que les représentations internes des systèmes de traduction encodent effectivement divers traits de surface comme la longueur de la phrase source, le temps du verbe principal, ou encore le nombre du sujet et de l'objet direct.

5.2 Vers l'analyse causale

Comme le note toutefois [Vanmassenhove et al., 2017], qui étendent les travaux de [Shi et al., 2016] à l'étude de l'aspect du verbe principal, ce n'est pas parce qu' une information utile pour la traduction est correctement extraite qu'elle sera correctement utilisée. Cette observation est corroborrée par [Wisniewski et al., 2021b] qui étudie le transfert du genre du groupe sujet entre français et anglais sur un corpus contrôlé : alors que l'information de genre est correctement extraite et encodée par le système, celui-ci

s'avère pour autant extrêmement biaisé (en faveur du masculin) lorsqu'il s'agit de produire les traductions. Un autre point d'attention méthodologique est soulevé par [Voita and Titov, 2020, Hewitt and Liang, 2019], qui mettent en garde contre la capacité des méthodes à base de sondes à extraire y compris des associations artéfactuelles et arbitraires, et proposent des mesures pour pallier ce problème.

Les méthodes à base de sondes permettent de mettre en évidence des associations utiles entre les représentations extraites et les propriétés linguistiques de la phrase source. Du fait toutefois du caractère hautement multifactoriel des décisions probabilistes qui sont prises par un système neuronal, il est difficile de tirer des '"explications"' de ces associations. Pour progresser dans cette voie, plusieurs travaux récents proposent de mettre en place des '"interventions"' qui vont activement manipuler les représentations internes du réseau, de manière à rendre plus explicites des liens de causalité entre ces représentations et les décisions prises par le système de traduction. Cette approche est initialement proposée pour l'analyse des mécanismes d'accord par [Giulianelli et al., 2018], qui calcule ses interventions à partir des réponses de classifieurs auxiliaires. Lorsque les représentations internes associées au nom sujet, d'après la sonde, conduisent à une réponse erronnée sur le nombre du verbe, elles sont mises à jour selon une règle de gradient qui vise à renforcer l'encodage du nombre correct. Les auteurs montrent que cette méthode améliore effectivement la propagation du nombre dans la phrase, jusqu'à la sélection du nombre correct pour le verbe.

[Vig et al., 2020, Wisniewski et al., 2021a] proposent d'autres manipulations pour étudier les biais de genre respectivement en génération de textes et en traduction automatique. La question posée dans cette dernière étude est de savoir si le genre du groupe nominal (GN) sujet est correctement transféré du français vers l'anglais dans des phrases standardisées de la forme "DET N a fini son travail" où N est un nom d'occupation qui peut être soit féminin, soit masculin. Une première observation que font ces auteurs est que quel que soit le genre du N, la traduction anglaise contient majoritairement le pronom "his". Pour vérifier que le système de TA utilise pourtant l'information de genre présente en français, la manipulation proposée consiste à "neutraliser" les représentations calculées pour les mots du GN, c'est à dire à les rendre indépendantes du genre. L'hypothèse est que si le système de TA est sensible au genre du GN, cette manipulation devrait rendre "her" encore moins probable ; dans le cas inverse, on ne devrait pas observer de changement dans les sorties.

La caractéristique de cette méthode est donc qu'elle agit directement et de manière contrôlée sur tout ou partie des paramètres de θ pour mettre à jour des relations de causalité entre les valeurs de certains paramètres et

les décisions prises pour calculer les traductions, et par là, fournir des explications sur le comportement du système, et inspirer de nouvelles méthodes pour le réguler.

6 Conclusion

Les progrès récemment observés en matière de traduction automatique reposent sur l'utilisation de méthodes d'apprentissage automatique sophistiquées, qui mettent en jeu algorithmes calculant des représentations numériques des mots et phrases en langue cible et en langue source. Les systèmes de TA neuronales sont ainsi particulièrement opaques et il est difficile d'analyser finement leurs performances, de les metttre en rapport avec des types de difficultés de traduction, ou encore de développer des améliorations visant un problème particulier. En réponse à cette difficulté, les concepteurs de systèmes de TA ont développé diverses stratégies que nous avons passé en revue dans cette contribution : élaboration de bancs d'essais dédiés à l'étude d'un problème particulier, développement de méthodes d'analyse contrastives, conception de techniques pour visualiser ou analyser les représentations internes du système. Une question apparait de manière récurrente dans ces travaux : celle de la génération automatique de jeux de tests, et de la mesure automatique des performances, qui seules permettent de construire des métriques exploitables dans le cycle de développement de systèmes.

Toutes ces méthodes ont leurs limitations, et chacune ne donne qu'une vision partielle et imparfaite de la qualité d'un système de TA, suggérant qu'il faudra les combiner pour dessiner une vision plus complète des réelles capacités des systèmes de TA neuronaux. Il est toutefois possible que les approches actuelles, consistant à étudier ex-post les systèmes de TA à partir de l'observation de leurs entrées/sorties atteignent ces limites, et qu'il faille, à l'avenir, constuire des systèmes qui soient par construction plus transparents, et dotés dès leur conception, de la capacité d'expliquer en termes intelligibles par les utilisateurs les décisions qui sont prises au cours du calcul d'une traduction, comme le recommande [Rudin, 2019] pour les systèmes critiques à base d'intelligence artificielle.

Références

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, On-

line, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.69.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the first International Conference on Learning Representations*, San Diego, CA, 2015.

Antonio Balvet. Métriques d'évaluation en traduction automatique : le sens et le style se laissent- ils mettre en équation ? In T. Milliaressi, editor, *La Traduction épistémique : entre poésie et prose*, pages 315–356. Presses Universitaires du Septentrion, 2020. URL https://books.openedition.org/septentrion/93938.

Satanjeev Banerjee and Alon Lavie. METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, pages 65–72, Ann Arbor, Michigan, 2005. URL http://www.aclweb.org/anthology/W/W05/W05-0909.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 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)*, pages 1304–1313. Association for Computational Linguistics, 2018. doi : $10.18653/\mathrm{v1}/\mathrm{N18}\text{-}1118$. URL http://aclanthology.org/N18-1118.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing : A Survey. *Transactions of the Association for Computational Linguistics*, 7 :49–72, 04 2019. ISSN 2307-387X. doi : $10.1162/\mathrm{tacl\_a\_00254}$. URL https://doi.org/10.1162/tacl_a_00254.

Hervé Blanchon and Christian Boitet. Pour l'évaluation externe des systèmes de ta par des méthodes fondées sur la tâche. *Traitement Automatique des Langues*, 48 :33–65, 2007.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. A linguistic evaluation of rule-based, phrase-based, and neural mt engines. *The Prague Bulletin of Mathematical Linguistics*, 108 :159 – 170, 2017.

Franck Burlot and François Yvon. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference*

on Machine Translation, Volume 1 : Research Papers, pages 43–55, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{W17\text{-}4705}$. URL http://aclweb.org/anthology/W17-4705.

Franck Burlot and François Yvon. Evaluation morphologique pour la traduction automatique : adaptation au français. In *Conférence sur le Traitement Automatique des Langues Naturelles*, TALN, page 14 pages, Rennes, France, 2018. ATALA.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. Approaches to human and machine translation quality assessment. In *Translation quality assessment*, pages 9–38. Springer, 2018.

Eirini Chatzikoumi. How to evaluate machine translation : A review of automated and human metrics. *Natural Language Engineering*, 26(2) : 137–161, 2020.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation : Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. URL http://www.aclweb.org/anthology/W14-4012.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $& !#* vector : Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{P18\text{-}1198}$. URL https://aclanthology.org/P18-1198.

Mikel L. Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, pages 192–203, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{W18\text{-}6320}$. URL https://aclanthology.org/W18-6320.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context : A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9 :1460–1474, 2021. doi : $10.1162/\mathrm{tacl\_a\_00437}$. URL https://aclanthology.org/2021.tacl-1.87.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and

Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 2017. URL http://proceedings.mlr.press/v70/gehring17a.html.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood : Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{W18\text{-}5426}$. URL https://aclanthology.org/W18-5426.

Liane Guillou and Christian Hardmeier. PROTEST : A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology/L16-1100.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, pages 525–542, Berlin, Germany, August 2016. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{W16\text{-}2345}$. URL https://aclanthology.org/W16-2345/.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. Pronoun-focused MT and cross-lingual pronoun prediction : Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{W15\text{-}2501}$. URL https://aclanthology.org/W15-2501.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{D19\text{-}1275}$. URL https://www.aclweb.org/anthology/D19-1275.

Eduard Hovy, Margaret King, and Andrei Popescu-Belis. Principles of context-based machine translation evaluation. *Machine Translation*, 17 (1) :43–75, 2002.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61 :907–926, 2018.

Pierre Isabelle, Colin Cherry, and George Foster. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{D17\text{-}1263}$. URL https://www.aclweb.org/anthology/D17-1263.

Margaret King and Kirsten Falkedal. Using test suites in evaluation of machine translation systems. In *Papers presented to the 13th International Conference on Computational Linguistics*, COLING 1990, 1990. URL https://aclanthology.org/C90-2037.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. Just ask ! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.58.

Natalie Kübler. A comparable learner translator corpus : creation and use. In *Proc. of LREC 2008 Workshop on Building and Using Comparable Corpora*, BUCC, pages 73–78, Marrakech, Morocco, 2008.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Review*, 67 :653–672, 2020.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (MQM) : A framework for declaring and describing translation quality metrics. *Revista Tradumàtica : tecnologies de la traducció*, (12) :455–463, 2014. doi : $\mathrm{https://doi.org/10.5565/rev/tradumatica.77}$.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation : Methods and evaluation. *ACM Comput. Surv.*, 54(2), March 2021. ISSN 0360-0300. doi : $10.1145/3441691$. URL https://doi.org/10.1145/3441691.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*,

ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. doi : 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

John R. Pierce, John B. Carroll, Eric P. Hamp, David G. Hays, Charles F. Hockett, Anthony G. Oettinger, and Alan Perlis. Language and machines - computers in translation and linguistics. Technical report, ALPAC Report, National Academy of Sciences, Washington, DC, 1966. URL https://nap.nationalacademies.org/read/9547/chapter/1.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. The MuCoW test suite at WMT 2019 : Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy, August 2019. Association for Computational Linguistics. doi : 10.18653/v1/W19-5354. URL https://aclanthology.org/W19-5354.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET : A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.213.

Annette Rios, Mathias Müller, and Rico Sennrich. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation : Shared Task Papers*, pages 588–596, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi : 10.18653/v1/W18-6437. URL https://www.aclweb.org/anthology/W18-6437.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5) :206–215, 2019. doi : 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

Danielle Saunders and Bill Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. In *ACL*, pages 7724–7736, Online, July 2020. ACL. doi : 10.18653/v1/2020.acl-main.690. URL https://www.aclweb.org/anthology/2020.acl-main.690.

Carolina Scarton and Lucia Specia. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1579.

Rico Sennrich. How grammatical is character-level neural machine translation ? assessing MT quality with contrastive translation pairs. In *Procee-*

dings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers, pages 376–382, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2060.

Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural MT learn source syntax ? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas, November 2016. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/\mathrm{D16\text{-}1159}$. URL https://aclanthology.org/D16-1159.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA, 2006.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. *Quality estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2018.

Brian Thompson and Matt Post. 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)*, pages 90–121, Online, November 2020. Association for Computational Linguistics. doi : $10.18653/\mathrm{v1}/2020.\mathrm{emnlp\text{-}main}.8$. URL https://aclanthology.org/2020.emnlp-main.8.

Eva Vanmassenhove, Jinhua Du, and Andy Way. Investigating 'aspect' in NMT and SMT : Translating the english simple past and present perfect. *Computational Linguistics in the Netherlands Journal*, 7 :109–128, Dec. 2017. URL https://www.clinjournal.org/clinj/article/view/73.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum

description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November 2020. Association for Computational Linguistics. doi : 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14.

Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context : Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July 2019. Association for Computational Linguistics. doi : 10.18653/v1/P19-1116. URL https://www.aclweb.org/anthology/P19-1116.

Guillaume Wisniewski, Lichao Zhou, Nicolas Ballier, and François Yvon. Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire. In Pascal Denis, Natalia Grabar, Amel Fraisse, Rémi Cardon, Bernard Jacquemin, Eric Kergosien, and Antonio Balvet, editors, *Traitement Automatique des Langues Naturelles*, pages 11–25, Lille, France, 2021a. ATALA. URL https://hal.archives-ouvertes.fr/hal-03265895.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. Screening Gender Transfer in Neural Machine Translation. In *Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Punta Cana, Dominica, November 2021b. Association for computational linguistics. URL https://hal.archives-ouvertes.fr/hal-03424174.

Tianyi Zhang, Varsha Kishore*, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore : Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.