ISSN 1974-4382

LA TRADUCTION MULTILINGUE : ANALYSE D'UNE PROUESSE TECHNOLOGIQUE

FRANÇOIS YVON
CNRS PARIS

francois.yvon@limsi.fr

Citation: Yvon, François (2023) "La traduction multilingue : analyse d'une prouesse technologique", in Maria Margherita Mattioda, Alessandra Molino, Lucia Cinato e Ilaria Cennamo (a cura di) *L'intelligenza artificiale per la traduzione: verso una nuova progettazione didattica?*, *mediAzioni* 39: A17-A34, https://doi.org/10.6092/issn.1974-4382/18785, ISSN 1974-4382.

Résumé : Les systèmes de traduction automatique (TA) neuronale ont fait ces dernières années des progrès tangibles, qui les ont rendus utilisables pour un nombre croissant de domaines et de couples de langues. Les systèmes neuronaux reposent sur des algorithmes d'apprentissage automatique et leur développement nécessite de grands corpus électroniques de textes parallèles, alignés au niveau des phrases, des ressources qui n'existent que pour un petit nombre de couples de langues et de domaines. Pour pallier ce manque, une proposition récente consiste à développer des systèmes de traduction dits « multilingues ». Ces développements ont été impulsés en particulier par les grands acteurs de l'Internet, pour qui le traitement d'un maximum de langues est un enjeu majeur. La principale caractéristique de ces architectures est de pouvoir traiter, avec un unique système de traduction, de multiples langues, aussi bien du côté source que du côté cible. Dans cette contribution, nous exposons les principes généraux qui sous-tendent ces systèmes et les innovations qui les ont rendus possibles, avant d'en discuter les principales forces et faiblesses.

Mots-clés : Traduction automatique neuronale ; Traduction automatique multilingue ; Transfert interlingue ; Interlingua neuronale.

Abstract: Neural machine translation (NMT) systems have made tangible progress in recent years, making them usable for an increasing number of domains and language pairs. The development of neural systems is based on machine learning algorithms and requires large electronic corpora of parallel texts, aligned at the sentence level. Such resources however only exist for a small number of language pairs and domains. To overcome this problem, a recent proposal is to develop so-called "multilingual" translation systems. These developments have been driven in particular by major Internet players, who need to develop automatic language processing tools for as many languages as possible. The main characteristic of these systems is to process multiple

languages, both on the source and target sides, with a single translation engine. In this paper, we present the general principles underlying these systems and the innovations that have made them possible, before discussing their main strengths and weaknesses.

Keywords: Neural Machine Translation; Multilingual Machine Translation; cross-lingual transfer; neural interlingua.

1. Introduction

Les méthodes de traduction automatique (TA) neuronale ont fait ces dernières années des progrès significatifs, qui les ont rendues utilisables pour un nombre croissant de domaines et de couples de langues, et ont conduit à leur déploiement rapide sur les plateformes de traduction en ligne ainsi que dans les outils professionnels de traduction automatique et de traduction assistée par ordinateur (TAO). Comme pour la génération précédente de systèmes statistiques, l'apprentissage et le développement de systèmes de TA neuronale[1] s'appuient sur de grands corpus électroniques de textes parallèles, alignés au niveau des phrases. De telles ressources n'existent pourtant que pour un nombre restreint de couples de langues et de domaines, ce qui réduit de fait l'applicabilité des méthodes neuronales.

Pour pallier ce manque, les systèmes de traduction neuronaux dits « multilingues »[2] (Dabre *et al.* 2020) se sont progressivement imposés, sous l'impulsion en particulier des grands acteurs de l'Internet pour qui le besoin de traiter le maximum de langues est un enjeu majeur – en particulier dans le but d'exploiter automatiquement les traductions automatiques (vers l'anglais) à des fins d'indexation, de filtrage ou d'analyse automatique des textes. La principale caractéristique de ces architectures est de pouvoir traiter, avec un *unique système de traduction*, de multiples langues, aussi bien du côté source que du côté cible.

Un premier bénéfice de cette approche est qu'elle permet d'étendre l'application des modèles neuronaux à des langues ou paires de langues dites « peu dotées », c'est-à-dire pour lesquelles les corpus parallèles existants sont insuffisants pour construire des systèmes bilingues performants (Haddow *et al.* 2022). En mutualisant l'entraînement du système de traduction sur de multiples langues, on observe en effet un *transfert interlingue* positif, qui fait que chaque direction de traduction bénéficie de la présence de phrases parallèles en d'autres langues.

Une autre motivation pour développer des systèmes multilingues est de nature plus opérationnelle et prend tout son sens lorsque l'on considère des systèmes massivement multilingues. Au lieu de devoir préparer et entraîner un système pour chaque direction de traduction, c'est un unique système qui va prendre en charge simultanément tous les couples de langues. On remplace ainsi des centaines, voire des milliers de systèmes, par un système unique, ce qui

[1] Les principales caractéristiques des systèmes neuronaux sont rappelées à la section 2.
[2] En dépit de son caractère tautologique, c'est la terminologie qui semble s'être imposée, opposant ainsi les systèmes « multilingues » aux systèmes « bilingues », qui ne traitent qu'une seule direction de traduction.

https://doi.org/10.6092/issn.1974-4382/18785

constitue alors une simplification majeure, aussi bien du point de vue de l'entraînement des systèmes que du point de vue de leur déploiement et de leur maintenance.

Dans cette contribution, nous exposons les principes généraux qui sous-tendent ces systèmes de TA un peu particuliers, en décrivant les innovations successives qui les ont rendus possibles, ainsi que les nouvelles applications de ces technologies multilingues, avant de discuter des principales questions qui se posent lorsque l'on cherche à comprendre et à améliorer ces systèmes.

2. La TA neuronale : principes et concepts

2.1. Traduire par apprentissage

Les systèmes de traduction neuronale ont été introduits en 2014 (Cho *et al.* 2014). Ces systèmes se sont rapidement imposés dans leurs différentes moutures (Bahdanau *et al.* 2015 ; Gehring *et al.* 2017 ; Vaswani et *al. 2017*) comme étant les plus performants pour une vaste gamme d'applications et d'usages. À l'instar de la génération précédente de systèmes statistiques (Koehn 2010), ils reposent principalement sur des méthodes d'apprentissage automatique exploitant de larges corpus de textes parallèles, alignés au niveau des phrases. Dans un système neuronal, la traduction est formalisée par un algorithme, A, à qui on présente une phrase source (notée S), et qui produit en réponse une phrase cible (notée C). Le calcul réalisé par l'algorithme A dépend de millions, voire de milliards de paramètres numériques, dénotés collectivement par θ. Pour marquer cette dépendance, nous notons $A_\theta(S)$ la traduction automatique de la phrase source S.

L'entraînement du système consiste à trouver les paramètres θ qui produisent les meilleures traductions possibles. À cet effet, on présente de manière répétée à A_θ des exemples (S, C) de traductions humaines extraites de très gros corpus parallèles. La traduction automatique $A_\theta(S)$ est ensuite comparée avec C, ce qui conduit à ajuster θ afin de corriger les différences observées entre la sortie calculée $A_\theta(S)$ et la sortie désirée (C). En répétant cette procédure un très grand nombre de fois, le comportement de A_θ se rapproche progressivement du comportement souhaité, qui est de reproduire fidèlement les traductions du corpus d'apprentissage.

2.2. Au cœur de l'algorithme : encodage et décodage

Si l'on détaille maintenant le fonctionnement de l'algorithme A_θ, on peut en première approche le décomposer en deux étapes principales. La première est l'**encodage** du texte source, qui consiste à transformer la séquence de symboles sources présentés à l'entrée du système en une représentation purement numérique (un tableau de nombres). Symétriquement, le **décodage** produit mot après mot une séquence en langue cible, chaque prédiction dépendant à la fois de toute la phrase source ainsi que des mots cibles qui ont déjà été produits. Dans cette approche, les vocabulaires source et cible sont fixés à l'avance, et contiennent des dizaines de milliers de mots sélectionnés en fonction de leur

https://doi.org/10.6092/issn.1974-4382/18785

fréquence. Cette architecture dite *encodeur-décodeur* peut être réalisée computationnellement de multiples manières, l'implémentation la plus répandue s'appuyant aujourd'hui sur le modèle Transformer de Vaswani *et al.* (2017). Il est important de noter que les paramètres qui déterminent les comportements respectifs de l'encodeur et du décodeur sont distincts : θ se compose donc de deux sous-ensembles θ_S pour l'encodeur, θ_C pour le décodeur. L'architecture encodeur-décodeur est schématisée Figure 1.

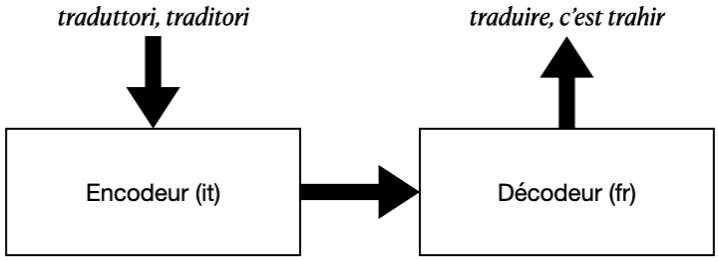

Figure 1. Une schématisation de l'architecture encodeur-décodeur d'un système bilingue. Encodeur et décodeur ne prennent en charge qu'une langue : l'italien en source, le français en cible.

3. Premiers pas vers la traduction multilingue

3.1. Traduire par pivot

La question du manque de données parallèles pour apprendre des modèles de TA s'est déjà posée pour les systèmes statistiques. Pour y répondre, de multiples variantes de l'approche **par pivot** ont été proposées. Lorsqu'il s'avère impossible de traduire directement depuis la langue L_A vers la langue L_B par manque de données, l'approche par pivot consiste à utiliser une langue tierce L_C, vers laquelle on sait traduire depuis L_A, et depuis laquelle on sait traduire vers L_B[3]. Pour des raisons de disponibilité des données, l'anglais est souvent le choix par défaut pour cette langue tierce. On note qu'alors la traduction $L_A \rightarrow L_B$ est **indirecte** et résulte d'un double processus automatique $L_A \rightarrow L_C \rightarrow L_B$. Ceci la rend d'autant plus risquée puisque la sortie cumulera les erreurs des deux étapes successives[4]. Pour autant, cette approche pragmatique a longtemps été la seule manière de répondre à la demande de traduire toujours plus de nouvelles paires de langues à l'aide de modèles statistiques (Utiyama et Isahara 2007). On notera qu'elle s'applique indifféremment aux systèmes statistiques et neuronaux.

L'approche par pivot peut également être détournée pour fabriquer des données d'apprentissage artificielles : partant d'un corpus parallèle pour le couple de langues (L_A, L_C), traduire automatiquement toutes les phrases de L_C vers L_B fournit un corpus parallèle (L_A, L_B), dans lequel le côté source contient des phrases correctes et le côté cible des phrases (imparfaitement) traduites. On

[3] Cette pratique de traduction indirecte ne se limite pas à la TA et a également été très utilisée pour produire des traductions par retraduction.
[4] La stratégie du pivot a longtemps été utilisée par Google Translate pour de nombreuses directions de traduction, voir par exemple Kaplan et Kianfar (2015).
https://doi.org/10.6092/issn.1974-4382/18785

obtient le résultat inverse en **rétrotraduisant** vers L_A le côté source d'un corpus (L_B, L_C). Ces approches font partie d'un ensemble de travaux relatifs à l'**augmentation de données** ; elles permettent de produire automatiquement des données artificielles, qui peuvent ensuite être utilisées pour construire de nouveaux systèmes (directs), statistiques ou neuronaux.

Avec le développement des techniques neuronales, l'approche par pivot a toutefois été supplantée par des méthodes plus efficaces pour pallier le manque de données parallèles.

3.2. Un encodeur, plusieurs décodeurs

Le principe fondamental des méthodes de traduction multilingues est d'éviter d'avoir à apprendre un ensemble de paramètres $\boldsymbol{\theta}$ pour chaque couple de langues. Il s'avère en effet bien plus rentable et computationnellement efficace de mutualiser ces apprentissages et de les rendre mutuellement bénéfiques.

Un premier pas dans cette direction est l'architecture de Dong *et al.* (2015), qui explorent différentes manières de traduire automatiquement depuis l'anglais vers plusieurs langues européennes (espagnol, français, néerlandais, portugais). L'approche par défaut consisterait à entraîner un jeu de paramètres distinct pour chaque direction selon les principes présentés à la section 2. Dans la mesure où la langue source est toujours la même, l'alternative étudiée consiste à n'apprendre qu'un seul encodeur (pour l'anglais) et quatre décodeurs (un par langue cible), donnant lieu à un système *multicible*. L'apprentissage se déroule comme expliqué ci-dessous, à ceci près que les données d'apprentissage rassemblent des phrases parallèles associant l'anglais avec quatre langues cibles différentes. Lorsque l'exemple à reproduire est une traduction vers le français, l'apprentissage ajuste les paramètres de l'encodeur (anglais) et du décodeur (français) ; si c'est une traduction vers le portugais on ajuste l'encodeur (anglais) et le décodeur (portugais), etc. Comme le système apprend simultanément plusieurs tâches de traduction, on parle alors d'apprentissage multitâches. Les auteurs montrent que cette approche est doublement bénéfique : d'une part, l'encodeur bénéficie de plus de données d'apprentissage ; d'autre part, la proximité entre langues cibles fait que chacune semble bénéficier des exemples fournis dans une autre langue, un effet qui est encore plus intéressant lorsque les corpus parallèles bilingues dont on dispose sont de petite taille. Cette architecture est représentée Figure 2.

https://doi.org/10.6092/issn.1974-4382/18785

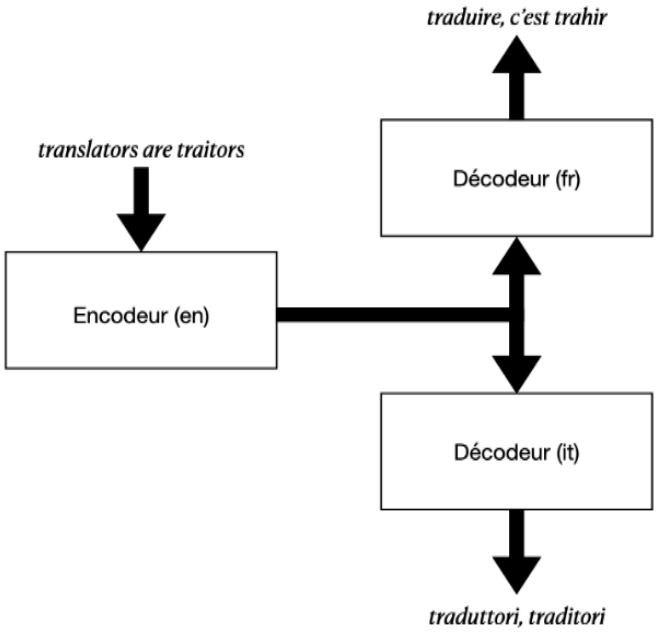

Figure 2. Une illustration de l'architecture encodeur-décodeur pour un système multicible, comprenant un unique encodeur et deux décodeurs. Chaque décodeur ne prend en charge qu'une langue.

3.3. Plusieurs encodeurs, plusieurs décodeurs

Cette stratégie s'applique à l'identique lorsque l'on inverse les directions de traduction, conduisant à entraîner plusieurs encodeurs et un seul décodeur. L'étape suivante consistera à appliquer cette stratégie à la fois pour entraîner plusieurs encodeurs (un par langue source) et plusieurs décodeurs (un par langue cible) (Firat *et al.* 2016a ; Ha *et al.* 2016), permettant de réaliser une économie considérable (voir la Figure 3). Considérons en effet un ensemble de N langues et supposons que l'on souhaite traduire automatiquement toutes les paires de cet ensemble. L'approche directe (bilingue) conduit alors à entraîner $N \times (N-1)$ encodeurs et autant de décodeurs. Dans la même configuration, l'approche multilingue ne demande que d'entraîner N encodeurs et N décodeurs, ce qui représente une économie considérable (pour $N = 100$, cela représente environ 100 fois moins de paramètres à entraîner). Pour la mettre en œuvre, on utilisera un corpus agrégeant des traductions pour toutes les langues sources et cibles possibles, en sélectionnant à chaque fois le bon couple encodeur-décodeur.

https://doi.org/10.6092/issn.1974-4382/18785

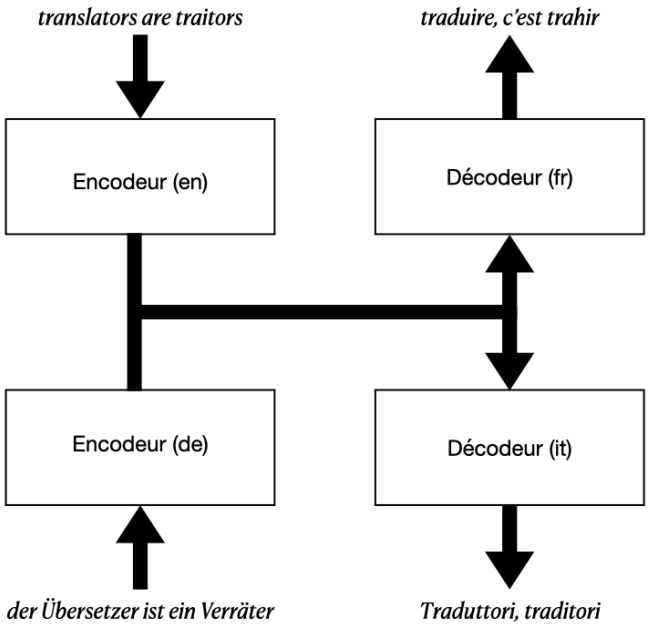

Figure 3. Une illustration de l'architecture encodeur-décodeur pour un système multisource-multicible, comprenant deux encodeurs et deux décodeurs. Ce système unique produit des traductions automatiques pour 4 couples de langues

3.4. Traduire sans données parallèles

Comme le réalisent très vite Ha *et al.* (2016), l'architecture multilingue présente un intérêt supplémentaire, à savoir pouvoir traduire automatiquement d'une langue L_A vers une langue L_B *sans avoir observé d'exemple pour ce couple de langues*, une configuration appelée **traduction zéro-ressource** (*zero-shot* translation)[5]. L'observation d'exemples associant L_A à d'autres langues que L_B suffit en effet pour apprendre à encoder L_A; de même, des exemples de phrases traduites dans la langue L_B suffiront pour apprendre à décoder cette langue. Une fois l'apprentissage terminé, coupler l'encodeur pour L_A et le décodeur pour L_B produira les traductions désirées. Les premières expériences décrites par ces auteurs ont permis de mettre en évidence le potentiel, mais aussi les limitations des architectures zéro-ressource, qui produisent majoritairement des traductions de très mauvaise qualité. Elles ont également confirmé l'intérêt des méthodes à base de pivot pour engendrer des données d'apprentissage artificielles.

3.5. Deux innovations supplémentaires

3.5.1. Contrôler la langue du décodeur

Plusieurs innovations supplémentaires vont permettre d'aboutir à des systèmes beaucoup plus performants. La première (Johnson *et al.* 2017) adapte à la

[5] Les recherches sur les systèmes de traduction non-supervisés (*unsupervised machine translation*) bilingues, capables d'apprendre à traduire sans données parallèles, utilisent des techniques similaires (pivot, augmentation de données) à celles qui sont présentées ci-dessus, et ont beaucoup influencé les travaux sur les systèmes multilingues (Lample *et al.* 2018).

https://doi.org/10.6092/issn.1974-4382/18785

traduction multilingue une méthode très simple pour conditionner les TA neuronales par des informations externes. Elle consiste à insérer avant le début de la phrase source à traduire des *pseudo-mots*[6] qui expriment des connaissances supplémentaires (relative à la phrase source) ou des contraintes (relatives à la phrase cible). Cette approche a été proposée pour injecter des informations sur le domaine (source) dans Kobus *et al.* (2017) ou des contraintes sur le degré de formalité de la cible (Sennrich *et al.* 2016a). Dans un scénario multilingue, en remplaçant l'exemple ('The grass is green', 'l'herbe est verte') par ('[2fr] The grass is green', 'l'herbe est verte'), on introduit, grâce au préfixe [2fr], une nouvelle information : la phrase cible est en français. En combinant alors des exemples pour le couple (anglais, français) et des exemples pour le couple (anglais, espagnol), par exemple '([2es] Car drivers are crazy', 'los conductores estàn locos'), dans laquelle [2es] identifie une traduction vers l'espagnol, on peut traiter **avec un seul couple (encodeur, décodeur)** les deux paires de langue. En effet, une fois l'apprentissage réalisé, l'insertion du préfixe '[2es]' ou '[2fr]' avant la phrase source conditionnera le choix de la langue à utiliser pour produire la phrase cible.

On notera que dans cette architecture, il est inutile de spécifier la langue source, l'encodeur apprenant implicitement à construire des représentations "universelles" qui ne dépendent plus la langue en entrée. D'une certaine manière, ce mécanisme s'apparente à celui d'une traduction par pivot, dans laquelle le rôle habituellement dévolu à l'anglais est tenu par la représentation multilingue issue de l'encodeur, qui tient alors lieu de représentation source utilisée pour décoder vers toutes les langues cibles connues du modèle. La nouveauté principale, qui est d'utiliser un unique décodeur pour produire au choix des énoncés en deux langues, ne semble toutefois possible que parce que les langues cibles ont des vocabulaires proches, ou du moins partagent un même alphabet. Cette approche est illustrée à la Figure 4.

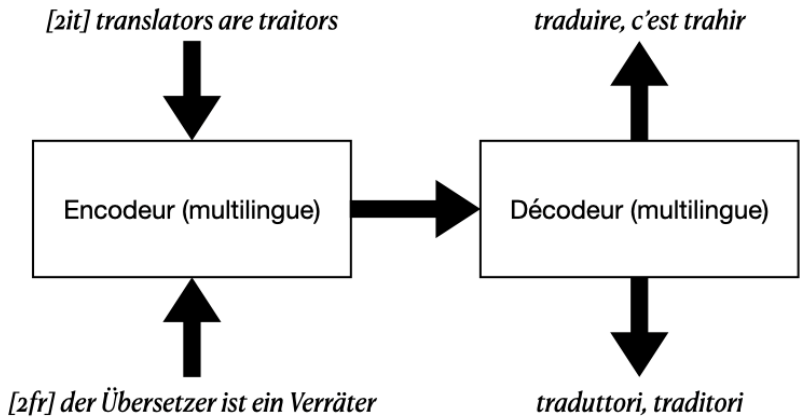

Figure 4. Une illustration de l'architecture encodeur-décodeur pour un système multilingue, comprenant un encodeur multilingue et un décodeur multilingue. Ce système traduit 4 couples de langues.

[6] Insérés avant le premier mot de la phrase source, ces préfixes particuliers seront notés entre [ ] pour les distinguer des mots du lexique courant.

https://doi.org/10.6092/issn.1974-4382/18785

3.5.2. Vers un vocabulaire universel

C'est ici qu'une seconde innovation intervient, qui consiste à se **défaire de la notion de mot** et à opérer la sélection des vocabulaires source et cible par un second mécanisme d'apprentissage automatique. Depuis ses débuts, la TA neuronale bute en effet sur un écueil, qui est la modélisation et la génération de mots imprévus, absents de l'ensemble d'apprentissage. Par définition, ces mots ne peuvent pas être traduits et causent des erreurs systématiques. La solution de *Wu et al.* (2016) consiste à pré-segmenter les phrases sources et cibles sur la base d'un ensemble prédéfini *d'unités sous-lexicales* correspondant soit à des mots fréquents, soit des fragments de mots, soit, au pire, à un découpage en lettres. L'important est que toute phrase puisse être ainsi décomposée de manière unique.

Tableau 1. Segmentation universelle appliquée à quatre langues[7].

fr	trad _uire c' est tra _hir
en	translator _s are trait _ors
it	trad _utto _ri trad _itori
de	der Über _set _zeristein Ver _rät _er

Pour garantir que les mots les plus fréquents ne seront pas segmentés, l'inventaire d'unités sous-lexicales se fonde sur une analyse statistique des corpus d'apprentissage. Cette idée s'étend sans difficulté au cadre multilingue et conduit à modéliser un ensemble universel d'unités, susceptible d'être utilisé pour encoder et décoder toutes les langues, indépendamment même de leur système d'écriture. Le tableau 1 illustre cette segmentation pour quatre versions de la même phrase. On note que certaines unités apparaissent (par exemple 'trad') dans plusieurs langues, ce qui peut faciliter le transfert entre langues. Avec cette seconde avancée, les systèmes multilingues seront alors entraînés avec des exemples représentant de multiples couples de langues, toutes partageant un même vocabulaire, toutes contribuant à alimenter un unique encodeur universel et un unique décodeur universel (cf. Figure 4).

3.6. Le passage à l'échelle

Tous les ingrédients sont alors réunis pour passer à l'échelle. Alors que les premiers systèmes multilingues ne traitent que quelques couples de langues (12 pour Johnson *et al.* 2017), Neubig et Hu (2018) étudient un système qui traduit près de 60 langues, et Aharoni *et al.* (2019), puis Fan *et al.* (2021), proposent des systèmes de TA capables de réaliser toutes les traductions possibles entre 102 langues. Cette augmentation du nombre de langues s'est accompagnée d'une amélioration des performances, les systèmes multilingues surpassant même

[7] Les unités précédées d'un '_' (par ex. '_uire') correspondent à des fragments internes et se distinguent des unités apparaissant en début de mot (par ex. 'trad', que l'on trouve en français et en italien).

https://doi.org/10.6092/issn.1974-4382/18785

parfois les systèmes monolingues. Ce mouvement n'est pas achevé puisque l'état de l'art le plus récent revendique un doublement du nombre de langues couvertes (Bapna *et al.* 2022 ; Costa-jussà *et al.* 2022) – soit plus de 40 000 couples de traductions possibles réalisés par un unique système.

Pour l'essentiel, cette extension du domaine de la TA multilingue repose sur les principes exposés précédemment, ainsi que sur un énorme effort d'ingénierie, qui vise d'une part à collecter des corpus d'apprentissage réellement multilingues, d'autre part à entraîner des modèles toujours plus volumineux. Le premier effort est nécessaire pour dépasser les limites de modèles entraînés sur des corpus parallèles dans lesquels une langue (l'anglais) est dominante, et présente (en source ou en cible) dans la majorité des exemples. Ceci demande en particulier de fouiller à très large échelle les sources, parallèles ou quasi parallèles, présentes sur Internet, ainsi qu'à développer de nouvelles méthodes d'augmentation de données par pivot ou rétrotraduction (voir la section 3.1). Pour gérer ces immenses corpus d'apprentissage, qui, devenus multilingues, contiennent des milliards de phrases parallèles, ainsi que l'accroissement corrélatif de la taille des ensembles d'unités "universels" et des ensembles de paramètres associés, un autre axe de développement majeur a été l'amélioration des performances computationnelles des algorithmes d'apprentissage. Il n'est pas surprenant que l'entraînement et le déploiement de ces méga-modèles soient devenus l'apanage quasi-exclusif des géants de l'Internet, qui ont les motivations et les ressources pour s'engager dans cette course au gigantisme.

4. Extensions, défis et questionnements

L'avènement de méthodes de TA multilingue a d'évidentes implications pour le développement des technologies, démontrant la possibilité à large échelle de construire des systèmes de traitement automatiques capables de prendre en charge de multiples langues. Accompagnant ces développements, de nombreuses études se sont penchées sur la généralisation de ces méthodes à d'autres tâches complexes : résumé multilingue, dialogue multilingue etc. Nous commençons cette section en présentant diverses applications fructueuses de ces méthodes, avant d'en pointer également certaines limites.

4.1. Des traitements toujours plus multilingues

La traduction multilingue accompagne un mouvement plus large en traitement des langues pour apprendre des systèmes capables de traiter simultanément plusieurs langues, un des enjeux déjà mentionné dans la cadre de la TA étant celui de transférer des connaissances ou systèmes de traitement à des langues peu dotées et pour lesquelles il serait difficile d'apprendre un système neuronal depuis zéro. Cette thématique fait l'objet d'une large littérature, voir par exemple Ruder *et al.* (2019), qui excède de loin les applications à la traduction. Dans ce contexte, l'effort d'acquisition de corpus parallèles hautement multilingues est immédiatement valorisable, de nombreuses études ayant montré que ces corpus

https://doi.org/10.6092/issn.1974-4382/18785

facilitent l'apprentissage de représentations multilingues au niveau des mots et des phrases (Conneau *et al.* 2020).

Ces représentations multilingues sont, en retour, exploitées dans des nouveaux outils d'identification de phrases parallèles. Ces outils utilisent la capacité de l'encodeur d'un système de TA multilingue à construire des représentations universelles qui capturent, d'une certaine manière, la sémantique des phrases en entrée. Soient alors deux phases s_A et s_B, respectivement écrites dans les langues L_A et L_B et représentées numériquement par l'encodeur multilingue comme e_A et e_B: plus ces deux représentations seront proches, plus il sera plausible que ces deux phrases soient des traductions mutuelles. Des mesures de similarité peuvent alors servir à calculer des alignements de phrases dans des corpus parallèles, ou encore à détecter de possibles phrases parallèles dans des corpus multilingues (Artetxe et Schwenk 2019).

Une autre application des modèles de traduction multilingues est proposée par Thompson et Post (2020) : considérant que ces modèles permettent de sélectionner librement la langue source à associer à une cible donnée, les auteurs de ce travail considèrent alors des **traductions monolingues** et appliquent l'algorithme de traduction à des phrases sources telles que '[2fr] le vélo c'est la santé!' dans lesquelles on produit une cible en français depuis une source également en français – alors même que de tels couples n'ont jamais été observés à l'apprentissage. Le système de TA sert alors à produire des paraphrases ou bien encore à mesurer la proximité sémantique de deux phrases écrites dans la même langue.

4.2. Nouvelles questions linguistiques

L'architecture des systèmes de traduction multilingue, dont l'encodeur est capable de traiter des phrases écrites dans des langues variées et à en dériver des représentations "universelles", ravive l'espoir de faire émerger des **représentations interlingues du sens**. Ces représentations, qui encodent des significations indépendamment de la langue source, ont joué un rôle majeur dans le développement des systèmes de TA à base de règles (Dorr *et al.* 2004). Cette analogie entre représentations numériques et interlingues soulève de nombreuses questions relatives à l'encodage des significations qui est réalisé par ces modèles, à leur caractère réellement universel, à la possibilité de les analyser ou d'en extraire des connaissances sous une forme symbolique.

Une autre promesse des systèmes de traduction multilingues, régulièrement mise en exergue, est l'existence d'un **transfert positif**[8] entre langues et couples de langues, par le truchement duquel des exemples de traductions entre L_A et L_B permettraient d'améliorer l'apprentissage de la traduction entre L_A et L_C, voire entre des paires n'impliquant ni L_A ni L_B. L'analyse des effets de transfert, ainsi que des mécanismes par lesquels il se manifeste, est également un sujet très ouvert étudié par exemple dans Arivazhagan *et al.* (2019). Mieux le comprendre permettrait d'orienter la sélection des langues à prendre en compte à

[8] Il existe également un transfert négatif, tout aussi mal compris, qui se manifeste par une dégradation des performances quand on passe d'une traduction bilingue à une traduction multilingue. Ceci n'arrive toutefois que pour des paires de langues bien dotées.

https://doi.org/10.6092/issn.1974-4382/18785

l'apprentissage, ainsi qu'à mieux anticiper l'effet que produiraient de nouvelles données en L_A sur la traduction en L_B. Il est tentant de chercher à s'appuyer pour répondre à ces questions aux relations historiques ou typologiques entre langues et familles de langues, en faisant l'hypothèse que le transfert est plus fort entre langues d'une même famille (Kudugunta *et al.* 2019). Il semble toutefois que d'autres facteurs, en particulier le système d'écriture, jouent un rôle important pour faciliter ces transferts.

4.3. Limites et défis de la traduction automatique multilingue

Les systèmes actuels multilingues répondent à un besoin réel d'étendre l'applicabilité des technologies de traduction automatique, tout en facilitant la maintenance et la mise à jour de systèmes qui doivent gérer en production des milliers de couples de langues. Ils s'appuient sur des heuristiques robustes, qui laissent toutefois ouvertes plusieurs questions importantes.

La première porte sur le caractère égalitaire de l'apprentissage des systèmes de TA en régime multilingue. Il comprend plusieurs volets, également difficiles à aborder. Commençons par la question de l'inventaire des unités universelles, qui, appris à partir de corpus surreprésentant l'anglais, fait la part belle aux mots et morphèmes de cette langue (et par ricochet, des langues voisines), au détriment de langues utilisant d'autres systèmes d'écriture, dont les données sont plus éparses (Rust *et al.* 2021). L'acquisition d'un ensemble d'unités qui représenterait de manière équitable toutes les langues à partir de corpus de textes déséquilibrés reste une question très ouverte. Une alternative radicale récemment, explorée par exemple dans (Clark *et al.* 2022 ; Xue *et al.* 2022) consiste à faire reposer les processus d'encodage et de décodage sur une segmentation en caractères.

Un second aspect de cette question porte sur l'équilibrage des exemples présentés à la machine : présenter des phrases sources (respectivement cibles) majoritairement rédigées en anglais (ou dans une langue bien représentée) biaise l'apprentissage de l'encodeur (respectivement du décodeur) et produit des systèmes inégalement performants, qui traduiront mieux depuis et vers les langues bien dotées que vers des langues minoritaires[9]. La question posée alors porte sur la représentation quantitative des différents couples de langues dans les données d'apprentissage, ainsi que du mélange entre exemples réels et exemples artificiels ou bruités (Wang et Neubig 2019 ; Zhou *et al.* 2021).

Un troisième aspect porte enfin sur l'architecture et le dimensionnement du modèle. La force des systèmes multilingues repose sur le partage intégral des paramètres des différents encodeurs et décodeurs, qui induit selon les couples de langues des transferts positifs ou négatifs. Ceci suggère que cet aspect du modèle pourrait être optimisé par des stratégies de partage partiel des paramètres. On conçoit également qu'augmenter le nombre de langues, et la taille des corpus, s'accompagne d'un accroissement de la taille du modèle. Pour autant, il est difficile de quantifier précisément cet accroissement et de savoir jusqu'à quel

[9] Le tableau 33, p. 104 de Costa-jussà *et al.* (2022) montre ainsi que les scores de performance moyens pour traduire vers l'anglais, qui est la condition la plus favorable, sont plus de deux fois supérieurs au score de performance moyenné sur toutes les paires de langues.

https://doi.org/10.6092/issn.1974-4382/18785

point il est nécessaire. Pour progresser sur tous ces sujets, un préalable est de mieux formaliser l'objectif de l'apprentissage, à l'instar du travail de (Pham *et al.* 2021) pour la traduction multidomaine : désire-t-on un modèle également performant pour toutes les langues, ou bien est-il souhaitable qu'il soit plus performant pour les couples de langues les plus demandées par les internautes ou les mieux dotées, ou encore qu'il privilégie les traductions vers les langues les moins bien dotées ?

Rappelons enfin que l'évaluation de la qualité de la traduction automatique est une question difficile, encore mal résolue pour des couples de langues bien étudiées. On conçoit alors qu'évaluer la performance d'un unique système traduisant 40 000 paires de langue représente un défi colossal. Et ce d'autant plus que : (a) les métriques automatiques usuelles (Papineni *et al.* 2002 ; Banerjee et Lavie 2005) s'appuient sur la notion de mot pour effectuer leurs décomptes ; (b) les rares données de test aujourd'hui disponible[10] sont presque exclusivement des traductions depuis l'anglais, ce qui introduit un grand nombre de biais dans les mesures[11] ; (c) les mesures de performance rapportés dans les études publiées montrent des disparités énormes, et suggèrent que dans de nombreux cas, les traductions automatiques sont en fait inutilisables ; et (d) pour un grand nombre de couples de langues, il serait même difficile de trouver des experts bilingues capables d'évaluer les traductions produites. Deux questions découlent de cette observation : est-il possible d'améliorer significativement la traduction pour les couples peu dotés avec des techniques massivement multilingues ? Sera-t-il possible de les étendre pour capturer davantage encore de diversité[12]?

5. Conclusion

Les systèmes de traduction multilingues, capables de traduire dans un même système de multiples paires de langues, constituent une évolution majeure des systèmes de traduction neuronale. Ils apportent une réponse conceptuellement simple et opérationnellement efficace au besoin de traduire massivement depuis et vers de nombreuses langues, un besoin qui est partagé à la fois par les grands acteurs de l'Internet, les entreprises et les institutions multinationales. L'utilisation de systèmes multilingues permet également de proposer des solutions de traduction automatique pour des couples de langues peu dotés en ressources, pour lesquels la traduction automatique bilingue est virtuellement impossible.

Compte-tenu de leur impact, il est attendu que ces systèmes continuent à se développer et à se diffuser, en particulier pour (a) intégrer plus de sources de données et de langues, (b) mieux contrôler les phénomènes de transfert interlangue, ainsi que (c) exploiter des architectures qui permettront d'arbitrer

[10] Principalement résultant des efforts des chercheurs de Meta/Facebook et décrites en détail dans (Goyal *et al.* 2022).
[11] Par exemple, ces données surestiment la qualité de traduction vers l'anglais (Toral *et al.* 2018) ; et sous-estiment la qualité des traductions entre langues très proches, puisque les tests correspondants sont produits indépendamment depuis l'anglais.
[12] Le projet de Bapna *et al.* (2022) se donne pour objectif '1000 paires de langues', soit un million de directions de traduction.

https://doi.org/10.6092/issn.1974-4382/18785

plus finement entre les paramètres qui doivent être partagés entre plusieurs langues ou familles de langues, et ceux qui doivent au contraire être spécialisés.

Comme nous l'avons montré, le tour de force technologique que ces systèmes réalisent implique de mettre en place des mécanismes qui gomment les différences entre langues, aussi bien du côté de l'encodeur que du côté du décodeur. L'exemple le plus évident est l'abandon de la notion de mot linguistique, remplacé par des unités universelles encodant toutes les langues. Ce choix radical est une réponse pragmatique au besoin de traduire y compris des mots inconnus, mais implique un renoncement à toute forme de connaissance linguistique, lexicale ou terminologique, qui pourrait venir informer le processus de traduction. En ce sens, il semble que les systèmes multilingues évoluent dans une direction opposée à celle qui a conduit à des systèmes de traduction automatique de haute qualité pour des domaines restreints, pour lesquels ces ressources sont essentielles.

BIBLIOGRAPHIE

Aharoni, R., M. Johnson and O. Firat (2019) "Massively Multilingual Neural Machine Translation", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers)* (Minneapolis, Minnesota: Association for Computational Linguistics), 3874-3884. https://aclanthology.org/N19-1388/.

Arivazhagan, N., A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun *et al.* (2019) "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges", https://arxiv.org/abs/1907.05019.

Artetxe, M. and H. Schwenk (2019) "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond", *Transactions of the Association for Computational Linguistics* 7: 597-610, https://doi.org/10.1162/tacl_a_00288.

Bahdanau, D., K. Cho and Y. Bengio (2015) "Neural Machine Translation by Jointly Learning to Align and Translate", in *Proceedings of the First International Conference on Learning Representations* (San Diego, CA).

Banerjee, S. and A. Lavie (2005) "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* (Ann Arbor, Michigan), 65-72, http://www.aclweb.org/anthology/W/W05/W05-0909.

Bapna, A., I. Caswell, J. Kreutzer, O. Firat, D. Esch, A. van Siddhant *et al.* (2022) "Building Machine Translation Systems for the Next Thousand Languages", https://arxiv.org/abs/2205.03983.

Cho, K., B. Merrienboer, D. van Bahdanau and Y. Bengio (2014) "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* (Doha, Qatar), 103-111, http://www.aclweb.org/anthology/W14-4012.

https://doi.org/10.6092/issn.1974-4382/18785

Clark, J. H., D. Garrette, I. Turc and J. Wieting (2022) “Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation”, *Transactions of the Association for Computational Linguistics* 10: 73-91, https://doi.org/10.1162/tacl_a_00448.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán *et al.* (2020) “Unsupervised Cross-Lingual Representation Learning at Scale”, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 8440-8451, https://www.aclweb.org/anthology/2020.acl-main.747.

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan *et al.* (2022) “No Language Left Behind: Scaling Human-centered Machine Translation”, https://arxiv.org/abs/2207.04672.

Dabre, R., C. Chu and A. Kunchukuttan (2020) “A Survey of Multilingual Neural Machine Translation”, *ACM Computing Surveys* 53(5), https://doi.org/10.1145/3406095.

Dong, D., H. Wu, W. He, D. Yu and H. Wang (2015) “Multi-Task Learning for Multiple Language Translation”, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th international Joint Conference on Natural Language Processing* (volume 1: Long papers) (Beijing, China: Association for Computational Linguistics), 1723-1732, https://aclanthology.org/P15-1166.

Dorr, B. J., E. H. Hovy and L.S. Levin (2004) “Machine Translation: Interlingual Methods”, in K. Brown (ed) *Encyclopedia of Language and Linguistics*.

Fan, A., S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal *et al.* (2021) “Beyond English-Centric Multilingual Machine Translation”, *Journal of Machine Learning Research* 22: 1-48, http://jmlr.org/papers/v22/20-1307.html.

Firat, O., K. Cho and Y. Bengio (2016a) “Multi-way, Multilingual Neural Machine Translation with a Shared Attention Mechanism”, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California: Association for Computational Linguistics), 866-875, https://aclanthology.org/N16-1101/.

------, B. Sankaran, Y. Al-onaizan, F.T. YarmanVural and K. Cho (2016b) “Zero-Resource Translation with Multi-Lingual Neural Machine Translation”, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas: Association for Computational Linguistics), 268-277, https://aclanthology.org/D16-1026.

Gehring, J., M. Auli, D. Grangier, D. Yarats and Y.N. Dauphin (2017) “Convolutional Sequence to Sequence Learning”, in D. Precup and Y. W. Teh (eds) *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70 (International Convention Centre, Sydney, Australia), 1243-1252, http://proceedings.mlr.press/v70/gehring17a.html.

Goyal, N., C. Gao, V. Chaudhary, P-J. Chen, G. Wenzek, D. Ju *et al.* (2022) “The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation”, *Transactions of the Association for Computational Linguistics* 10: 522–538, https://doi.org/10.1162/tacl_a_00474.

Ha, T.-H., J. Niehues and A. Waibel (2016) “Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder”, in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)* (Vancouver, Canada: IWSLT).

Haddow, B., R. Bawden, A.V.M Barone, J. Helcl and A. Birch (2022) “Survey of Low-Resource Machine Translation”, *Computational Linguistics* 48: 673-732, https://doi.org/10.1162/coli_a_00446.

Johnson, M., M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen *et al.* (2017) “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation”, *Transactions of the Association for Computational Linguistics* 5: 339-351, https://doi.org/10.1162/tacl_a_00065.

Kaplan, F. et D. Kianfar (2015) « Il pleut des chats et des chiens : Google et l’impérialisme linguistique », *Le Monde diplomatique* 28, http://infoscience.epfl.ch/record/205081.

Kobus, C., J. Crego and J. Senellart (2017) “Domain Control for Neural Machine Translation”, in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* (Varna, Bulgaria: INCOMA Ltd.), 372-378, https://doi.org/10.26615/978-954-452-049-6_049.

Koehn, P. (2010) *Statistical Machine Translation*, Cambridge: Cambridge University Press.

Kudugunta, S., A. Bapna, I. Caswell and O. Firat (2019) “Investigating Multilingual NMT Representations at Scale”, 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)* (Hong Kong, China: Association for Computational Linguistics), 1565-1575, https://aclanthology.org/D19-1167.

Lample, G., A. Conneau, L. Denoyer and M. Ranzato (2018) “Unsupervised Machine Translation Using Monolingual Corpora Only”, *International Conference on Learning Representations* (Vancouver, Canada), https://openreview.net/forum?id=rkYTTf-AZ.

Luong, M.-T., Q.V. Le, I. Sutskever, O. Vinyals and L. Kaiser (2016) “Multi-Task Sequence to Sequence Learning”, in Y. Bengio and Y. LeCun (eds) *4th International Conference on Learning Representations, ICLR* (San Juan, Puerto Rico), http://arxiv.org/abs/1511.06114.

Neubig, G. and J. Hu (2018) “Rapid Adaptation of Neural Machine Translation to New Languages”, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium: Association for Computational Linguistics), 875-880, https://aclanthology.org/D18-1103.

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu (2002) “BLEU: a Method for Automatic Evaluation of Machine Translation”, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* ACL ’02 (Stroudsburg, PA, USA), 311-318, https://doi.org/10.3115/1073083.1073135.

Pham, M. Q., J. Crego and F. Yvon (2021) “Revisiting Multi-Domain Machine Translation”, *Transactions of the Association for Computational Linguistics* 9: 17-35, https://transacl.org/index.php/tacl/article/view/2327.

Ruder, S., I. Vulić and A. Søgaard (2019) “A Survey of Cross-Lingual Word Embedding Models”, *Journal of Artificial Intelligence Research* 65: 569-631.

Rust, P., J. Pfeiffer, I. Vulić, S. Ruder and I. Gurevych (2021) "How Good Is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 1: Long papers)* (Online: Association for Computational Linguistics), 3118-3135, https://aclanthology.org/2021.acl-long.243.

Sennrich, R., B. Haddow and A. Birch (2016a) "Controlling Politeness in Neural Machine Translation via Side Constraints", in *Proceedings of the 2016 Conference of the north American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California: Association for Computational Linguistics), 35-40, https://aclanthology.org/N16-1005.

------ (2016b) "Neural Machine Translation of Rare Words with Subword Units", in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers)* (Berlin, Germany): 1715-1725, https://aclanthology.org/P16-1162.

Thompson, B. and M. Post (2020) "Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing", in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (Online: Association for Computational Linguistics), 90-121, https://aclanthology.org/2020.emnlp-main.8.

Toral, A., S. Castilho, K. Hu and A. Way (2018) "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation", in *Proceedings of the Third Conference on Machine Translation: Research Papers* (Brussels, Belgium: Association for Computational Linguistics), 113-123, https://aclanthology.org/W18-6312.

Utiyama, M. and H. Isahara (2007) "A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation", *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of The Main Conference* (Rochester, New York: Association for Computational Linguistics), 484-491, https://aclanthology.org/N07-1061.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez *et al.* (2017) "Attention Is All You Need", in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan *et al.* (eds) *Advances in Neural Information Processing Systems* 30, 5998-6008, http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wang, X. and G. Neubig (2019) "Target Conditioned Sampling: Optimizing Data Selection for Multilingual Neural Machine Translation", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy: Association for Computational Linguistics), 5823-5828, https://aclanthology.org/P19-1583.

Wu, H. and H. Wang (2007) "Pivot Language Approach for Phrase-Based Statistical Machine Translation", *Machine Translation* 21: 165-181, https://doi.org/10.1007/s10590-008-9041-6.

Wu, Y., M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey *et al.* (2016) "Google's Neural Machine Translation System: Bridging the Gap between

Human and Machine Translation", https://doi.org/10.48550/arXiv.1609.08144.

Xue, L., A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale *et al.* (2022) "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models", *Transactions of the Association for Computational Linguistics* 10: 291-306, https://arxiv.org/abs/1609.08144.

Zhou, C., D. Levy, X. Li, M. Ghazvininejad and G. Neubig (2021) "Distributionally Robust Multilingual Machine Translation", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Online; Punta Cana, Dominican Republic: Association for Computational Linguistics): 5664-5674, https://aclanthology.org/2021.emnlp-main.458/.