الجمهورية الجزائرية الديمقراطية الشعبية

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHESCIENTIFIQUE

UNIVERSITE YAHIA FARES - MEDEA

FACULTE DES SCIENCES

Département Mathématiques et Informatique



Projet de fin d'études

Présenté pour l'obtention du Diplôme de

Master en Informatique

Option : Ingénierie des Systèmes et Technologies du Web

THEME :

Amélioration de la traduction automatique neuronale de la langue arabe à l'aide de l'augmentation des données

---

Proposé et dirigé par :

Mr. BABAALI Baligh

Préparé et présenté par :

Mr. BRAHIMI Slimane

Mr. OULDKHAOUA Abdelrezak

Année Universitaire 2023-2024

Dédicaces

Je dédie ce modeste travail

À mes chers parents, à qui je dois tout, et à qui j'exprime toute ma gratitude et ma reconnaissance, en espérant que Dieu me permette de les rendre heureux et qu'ils soient fiers de moi. À tous les membres de ma famille, à mon superviseur, à tous ceux qui m'ont enseigné ou qui m'ont apporté leur aide, à tous mes amis et mes proches, à tous ceux qui m'aiment de près ou de loin. Je souhaite également dédier ce travail à tous mes frères musulmans, en particulier aux opprimés partout dans le monde, en priant Dieu de les soutenir et de les soulager.

FREE GAZA 

Slimane

Dédicaces

Je dédie ce modeste travail

À mes chers parents, sans qui je n'aurais pas pu arriver jusqu'ici. Leur soutien indéfectible et leur amour ont été les fondations de mes réussites.je souhaite leur exprimer toute ma gratitude et mon amour éternel. Je souhaite également exprimer ma gratitude à mon encadrant, mes amis et à tous ceux qui ont contribué à m'aider.

Abdelrezak

Remerciements

En préambule à ce mémoire nous remercions ALLAH qui nous aide et nous donne la patience et le courage tout au long de notre vie et durant notre année d'étude.

Nous souhaitons adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire.

Nous tenons à exprimer notre profonde reconnaissance et la preuve de notre grande gratitude pour Mr. Baligh BABAALI, pour nous avoir proposé ce thème, pour son aide, ses conseils, sa disponibilité et le suivi qu'il nous a apporté tout au long de notre travail.

Nous tenons à remercier, également, les membres du jury pour avoir bien voulu examiner et avoir accepter de juger ce travail.

Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenu et encouragé au cours de la réalisation de ce mémoire.

Merci à toutes et à tous.

ملخص:

شهد مجال برمجة الكمبيوتر تقدماً كبيراً في السنوات الأخيرة، مع تطور التطبيقات في عدة تخصصات. شهدت تكنولوجيا الترجمة الآلية (MT)، وهي مجال فرعي، تطورات كبيرة من خلال منهجيات وتقنيات مختلفة. على الرغم من وجود العديد من أنظمة الترجمة الآلية (MT) والأدوات الداعمة لها، الترجمة العربية موجودة بالفعل؛ ومع ذلك، فإن جودتها ليست في المستوى المطلوب وتحتاج إلى تحسين.

بالإضافة إلى ذلك، كان باحثو الترجمة الآلية العربية (AMT) مدفوعين بالحاجة الملحة إلى تقنيات فعالة لمعالجة وترجمة المعلومات العربية. ونتيجة لذلك، اقترحوا تقنيات وحلول جديدة، مع التركيز بشكل خاص على طريقة الترجمة الآلية العصبية (NMT) المستخدمة على نطاق واسع. تقدم هذه الدراسة مراجعة ومقارنة متعمقة لمختلف طرق الترجمة الآلية العصبية (NMT)، مع التركيز في المقام الأول على الجهود البحثية في الترجمة العربية-الدارجة (والدارجة العربية) باستخدام الترجمة الآلية. تتناول المنهجيات التي تمت مناقشتها العديد من العقبات والقضايا اللغوية والتقنية، في حين تظهر نجاحًا كبيرًا على الطرق التقليدية. نتائج هذه الدراسة يمكن أن تفيد الأكاديميين والمهنيين من خلال إبقائهم على اطلاع وتزويدهم بالأدوات الأساسية لنموذج وتحسين AMT. تشمل الموارد المجموعات ومجموعات الأدوات والمنهجيات والنماذج الجديدة. وتسلط النتائج التي تم الحصول عليها الضوء على سلسلة من الاستنتاجات والانتقادات والأسئلة التي لم يتم حلها في هذا المجال.

كلمات مفتاحية: تعزيز البيانات، الترجمة الآلية، الشبكة العصبية، اللغة العربية الفصحى، اللهجة الجزائرية.

Résumé :

Le domaine de la programmation informatique a connu des progrès importants ces dernières années, avec le développement d'applications pour plusieurs disciplines. La technologie de traduction automatique (TA), un sous-domaine, a connu des progrès significatifs grâce à diverses méthodologies et techniques. Malgré l'existence de nombreux systèmes de traduction automatique (MT) et outils de support, La traduction arabe existe déjà ; cependant, sa qualité est médiocre et doit être améliorée.

De plus, les chercheurs en traduction automatique arabe (AMT) étaient motivés par le besoin pressant de technologies efficaces pour traiter et traduire l'information arabe. En conséquence, ils ont proposé de nouvelles techniques et solutions, en se concentrant particulièrement sur la méthode de traduction automatique neuronale (NMT) largement utilisée. Cette étude fournit un examen approfondi et une comparaison de diverses méthodes de traduction automatique neuronale (NMT), se concentrant principalement sur les efforts de recherche dans la traduction de l'arabe-darija (et du darija arabe) à l'aide de la traduction automatique. Les méthodologies discutées abordent divers obstacles et problèmes linguistiques et techniques, tout en présentant un succès significatif par rapport aux méthodes conventionnelles. Les résultats de cette étude peuvent profiter aux universitaires et aux professionnels en les tenant informés et en les dotant d'outils essentiels pour modéliser et améliorer l'AMT. Les ressources comprennent des corpus, des boîtes à outils, des méthodologies et de nouveaux modèles. Les résultats obtenus mettent en évidence une série de conclusions, de critiques et de questions non résolues dans ce domaine.

Mots clés : Augmentation des données, Traduction automatique, Réseau neuronal, Arabe MSA, Dialecte algérien.

Abstract:

The field of computer programming has seen significant progress in recent years, with the development of applications for several disciplines. Machine translation (MT) technology, a subfield, has seen significant advancements through various methodologies and techniques. Despite the existence of numerous machine translation (MT) systems and supporting tools, The Arabic translation already exists; however, its quality is poor and needs improvement.

Additionally, Arabic machine translation (AMT) researchers were motivated by the pressing need for efficient technologies to process and translate Arabic information. As a result, they proposed new techniques and solutions, particularly focusing on the widely used neural machine translation (NMT) method. This study provides an in-depth review and comparison of various neural machine translation (NMT) methods, primarily focusing on research efforts in translating Arabic MSA-Darija (and Darija-Arabic MSA) using machine translation. The methodologies discussed address various linguistic and technical obstacles and issues, while exhibiting significant success over conventional methods. The results of this study can benefit academics and professionals by keeping them informed and equipping them with essential tools to model and improve AMT. Resources include corpora, toolkits, methodologies and new models. The results obtained highlight a series of conclusions, criticisms and unresolved questions in this area.

Keywords : Data augmentation, Machine translation, Neural network, Arabic MSA, Algerian dialectal.

TABLE DES MATIERES

LISTE DES FIGURES I

LISTE DES TABLEAUX II

LISTE DES ACRONYMES III

INTRODUCTION GENERALE 1

CHAPITRE I : LA LANGUE ARABE ET LA TRADUCTION AUTOMATIQUE 3

I.1. Introduction 4

I.2. La langue arabe 4

I.3. Les réseaux sociaux 5

I.4. Traduction automatique 5

I.4.1. Les approches linguistiques 6
I.4.2. Les approches basées sur le corpus 8
I.4.3. Evaluation 11

I.5. Conclusion 14

CHAPITRE II :L'AUGMENTATION DES DONNEES 16

II.1. Introduction 17

II.2. Techniques et méthodes 18

II.2.1. Back-translation (Rétro-Traduction) 18
II.2.2. Right-Rotation Augmentation RRA (Augmentation par rotation à droite) 19

II.3. Conclusion 20

CHAPITRE III : LES RESEAUX DE NEURONES ET LA TRADUCTION AUTOMATIQUE 22

III.1. Introduction 23

III.2. Les réseaux de neurones 24

III.2.1. Les réseaux Feed-Forward Networks (FFN) 24
III.2.2. Les réseaux de neurones récurrents (RNN) 25
III.2.3. Long Short-Term Memory (LSTM) 26

III.3. La traduction automatique neuronale (NMT) 27

III.3.1. Apprentissage des modèles de traduction automatique neuronale 28
III.3.2. Décodage 29

III.4. Conclusion 30

CHAPITRE IV : CONCEPTION DE LA METHODE PROPOSEE 31

IV.1. Introduction 32

IV.2. Prétraitement **32**
IV.2.1. Nettoyage des données 32
IV.2.2. Préparation des données 32
IV.3. Description du modèle utilisé **33**
IV.3.1. Préparer des données à l'aide de BPE et créer un dictionnaire de mots 34
IV.3.2. Configurer les paramètres du model 35
IV.3.3. Entraînement du modèle 36
IV.3.4. Évaluation du modèle 37
IV.4. Étapes de la technique de Rétro-Traduction (Back-Translation) **38**
IV.5. Étapes de la technique de Rotation à Droite (Right Rotation) : **39**
IV.6. Conclusion **40**
CHAPITRE V : TESTS ET RESULTATS **41**
V.1. Introduction **42**
V.2. Matériel et logiciels utilisés **42**
V.3. Description de la base de données utilisée **43**
V.4. Apprentissage du modèle **44**
V.4.1. Paramètre 44
V.4.2. Résultats 45
V.4.3. Analyse des résultats 45
V.5. Conclusion **47**
CONCLUSION GÉNÉRALE **48**
REFERENCES BIBLIOGRAPHIQUES **51**

LISTE DES FIGURES

Figure 1. Le triangle de Vauquois, illustrant les fondements de la traduction automatique..... 6
Figure 2. Approche de traduction automatique statistique..... 9
Figure 3. Modèle NMT basé sur le paradigme Encode-Attend-Decode 10
Figure 4. Implémentation BART pour la traduction automatique basée sur le modèle BERT 11
Figure 5. L'approche d'augmentation de rotation a droite 20
Figure 6. Nombre d'articles mentionnant la « traduction automatique neuronale » par an selon Google Scholar. 24
Figure 7. L'architecture de NMT avec formation à la stabilité contradictoire. 28
Figure 8. Schéma du modèle Seq2Seq basé sur GRU avec Couche d'attention de Bahdanau 34

LISTE DES TABLEAUX

Tableau 1. Ordre des mots flexible en arabe 19
Tableau 2. Tableau des bibliothèques python utilisées 42
Tableau 3. Statistiques des ensembles de données MSA-DZDA utilisés 43
Tableau 4. Performances de traduction MSA-DZDA en utilisant les scores BLEU et ChrF .. 45

LISTE DES ACRONYMES

AMT: Arabic Machine Translation

BLEU: Bilingual Evaluation Understudy

BT: Back-Translation (Rétro-traduction)

ChrF: Character n-gram F-score

COMET: Crosslingual Optimized Metric for Evaluation of Translation

CPU: Central Processor Unit

DL: Deep Learning

DMT : Direct Machine Translation

DZDA : Dialecte Algérien

EBMT: Example-Based Machine Translation

GHz: Giga Hertz

GPU: Probability Density Function

GRU: Gated Recurrent Unit

LSTM: Long Short-Term Memory

ML: Machine Learning

MSA: Moderne Standard Arabe

NLP : Natural Language Processing

NMT : Neural Machine Translation (Traduction automatique neuronale)

PC: Personal Computer

RAM: Random Access Memory

RBMT: Rule-Based Machine Translation

RNN: Recurrent Neural Network

RR: Right-Rotation (Rotation à droite)

SMT: Statistical Machine Translation

TBMT: Transfer-Based Machine Translation

TPU: Tensor Processing Unit

INTRODUCTION GENERALE

Introduction Générale

La traduction automatique (TA) est devenue plus cruciale dans diverses applications et a réalisé des progrès significatifs pour presque toutes les langues [1]. En conséquence, les récents progrès de la traduction automatique (MT) ont grandement amélioré la qualité des traductions [2], [3] et même pour les paires de langues disposant de ressources abondantes, comme l'anglais-tchèque ou le chinois-anglais, la qualité de la traduction humaine n'a pas été égalée.

De plus, divers sous-domaines de recherche au sein des systèmes de traitement du langage naturel (NLP) ont réalisé des progrès significatifs dans le but de répondre aux demandes croissantes de traduction automatique (MT) dans divers domaines et applications, y compris les chatbots multilingues. Par conséquent, il existe un besoin croissant de traductions automatiques rapides et précises. Cependant, parvenir à une traduction appropriée et idéale est une entreprise difficile dans n'importe quel contexte linguistique [6], [7], [8]. Néanmoins, il est nécessaire d'améliorer davantage la qualité des traductions arabes en raison de nombreux problèmes soulignés dans diverses études de recherche, notamment les inexactitudes linguistiques [6], [9] – [13]. La traduction automatique présente de nombreux obstacles, qui peuvent être classés en deux groupes principaux : les défis linguistiques et les défis techniques. Un obstacle linguistique important est l'ambiguïté inhérente, la complexité linguistique et la diversité de la langue arabe par rapport aux autres langues. Des caractéristiques supplémentaires de l'arabe, telles que la possibilité d'avoir une liberté dans l'ordre des mots, divers schémas de diacritisation et un large éventail de variations dialectales basées sur des facteurs socio-économiques et géographiques, présentent des obstacles linguistiques substantiels pour la traduction automatique [14].

Toutefois, un des principaux défis demeure la rareté des données, ce qui limite l'efficacité des systèmes de traduction neuronale (NMT). Par conséquent, ce travail vise à augmenter les données disponibles pour améliorer la qualité de la traduction automatique neuronale.

CHAPITRE I : LA LANGUE ARABE ET LA TRADUCTION AUTOMATIQUE

I.1. Introduction

La traduction automatique (MT) est un domaine spécialisé du traitement du langage naturel (NLP) qui se concentre sur la création et l'amélioration de systèmes informatisés pour la traduction des langues. L'objectif d'un système de traduction automatique (MT) est de convertir mécaniquement un texte donné d'une langue à une autre tout en conservant son sens et son style, et en s'efforçant de générer une traduction linguistiquement fluide [15] [19].

L'arabe fait partie des cinq langues les plus parlées au monde, avec plus de 300 millions de locuteurs natifs. C'est également l'une des six langues officielles reconnues par les Nations Unies (ONU).6 C'est une langue sémitique réputée pour sa morphologie complexe et étendue, qui diffère considérablement de celle des langues indo-européennes comme l'anglais et le français. L'incorporation de la morphologie arabe, ainsi que d'autres facteurs linguistiques, a considérablement accru la complexité de la traduction automatique vers et depuis l'arabe. Malgré les progrès significatifs réalisés dans les paradigmes de traduction basés sur les données tels que les techniques statistiques et neuronales, les problèmes linguistiques continuent de poser de nombreux défis [1] [20].

I.2. La langue arabe

Il y a eu une croissance significative de la recherche consacrée au traitement du langage naturel arabe ces dernières années. La demande d'outils avancés de NLP arabe a augmenté en raison de la morphologie complexe de la langue. Ces outils sont nécessaires pour répondre aux exigences croissantes dans de nombreux domaines et applications.

La traduction automatique arabe (AMT) se heurte à des obstacles majeurs en raison de la nature complexe de la langue et de la morphologie arabes. Les difficultés se développent en relation avec le langage, la structure des phrases et le sens lorsque traduire le sens de la terminologie arabe en Darija et le contraire.

Le principal problème du la traduction de la langue arabe vers les autres langues est le manque de signes diacritiques dans les mots arabes, conduisant au choix de termes inappropriés dans la langue cible. D'autres problèmes incluent le manque de terminologie, des interprétations peu claires, des implications et des combinaisons de mots [1].

I.3. Les réseaux sociaux

La structure complexe de la langue arabe, ainsi que son manque de spécificité, entraînent un niveau d'ambiguïté important. Le niveau d'ambiguïté de la langue arabe utilisée dans les textes des réseaux sociaux continue d'augmenter.

Ces derniers temps, de nombreux problèmes sont apparus en raison de la prolifération des sites de médias sociaux. Le contenu trouvé dans les microblogs, comme

Les tweets sur Twitter se caractérisent par leur brièveté, leur volume élevé et leur orthographe et grammaire non conventionnelles. Il ne fait aucun doute que les utilisateurs des réseaux sociaux font souvent des fautes d'orthographe. Par exemple, ils ont tendance à translittérer les mots arabes en utilisant l'alphabet et les chiffres latins. De plus, chaque utilisateur transcrit les mots arabes de manière unique.

A titre d'illustration, la lettre arabe « ح » est communément représentée par le chiffre « 7 », tandis que la lettre « ق » est représentée par le chiffre « 9 », et ainsi de suite. Sur les réseaux sociaux, les utilisateurs translittèrent souvent le nom propre « احمد » (Ahmd) de différentes manières, telles que « ahmed », « ahmad », « ahmd », « a7mad », « a7med », « a7mmd », « a7md » ou "ahmmd".

En outre, l'utilisation de l'arabe dialectal (DA) sur les réseaux sociaux pose un défi important pour les applications de traitement du langage naturel (NLP) arabe. Premièrement, les méthodes existantes de traitement du langage naturel (NLP) sont incapables de traiter les dialectes couramment utilisés dans les textes des médias sociaux en raison de l'absence de conventions d'écriture rigides. Les tweets sont concis, cacophoniques et possèdent une structure complexe. Le contenu comprend plusieurs éléments distincts tels que le nom d'utilisateur, les hashtags, le retweet, etc. [2].

I.4. Traduction automatique

La traduction automatique est un processus informatique qui utilise un logiciel pour convertir du texte d'une langue naturelle (langue source, SL) vers une autre langue naturelle (langue cible, TL) [25]. Lors de tout processus de traduction, qu'il soit effectué par des humains ou des systèmes automatiques,

Les phrases sources doivent être fidèlement reproduites dans les phrases traduites cibles, même si cela peut sembler une tâche simple [2].

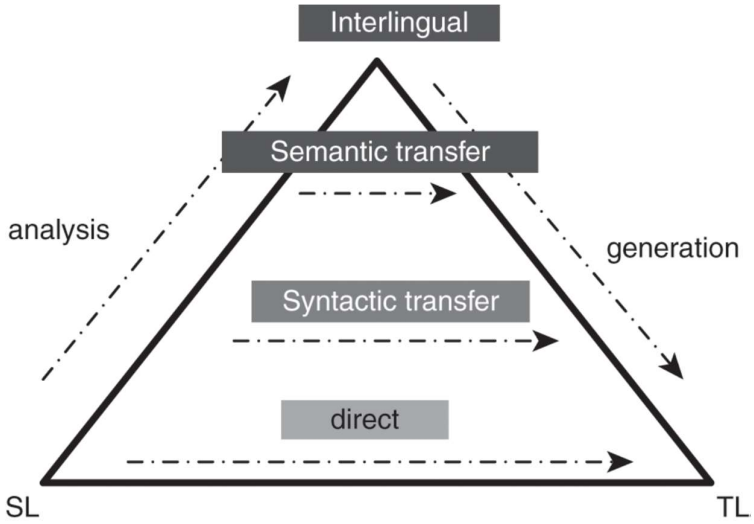


Figure 1. Le triangle de Vauquois, illustrant les fondements de la traduction automatique.

Les différentes méthodes de traduction automatique peuvent être classées en trois catégories distinctes :

1. S'appuyer sur la réglementation et l'expertise (à orientation linguistique). Les approches empiriques, qui s'appuient sur des données et des preuves, constituent une stratégie.
2. D'autres méthodes sont basées sur le corpus qui consistent à analyser de grandes collections de textes.
3. Enfin, des méthodes hybrides combinant différentes approches sont également utilisées.

I.4.1. Les approches linguistiques

Ces systèmes de traduction automatique visent à codifier toutes les connaissances essentielles nécessaires à la traduction, en employant des techniques expertes. La figure 1 représente le « triangle De Vauquois », qui sert de représentation générale de diverses stratégies. [2]

I.4.1.1. Approche directe (DMT)

La traduction automatique directe (DMT) est la méthode de traduction automatique la plus basique et la plus simple. Le système fonctionne au niveau des mots individuels, ce qui

signifie que la traduction des mots est effectuée sur une base individuelle, de la même manière que fonctionne un dictionnaire. Cependant, le contexte global ou la relation entre les mots est généralement peu pris en compte.

Le terme « signification » est défini comme indiqué dans la référence.

I.4.1.2. Traduction basée sur les règles (RBMT)

Le système RBMT repose sur la connaissance linguistique des langues source et cible, qui est principalement obtenue via des dictionnaires et grammaires bilingues. Ces ressources couvrent les principales règles morphologiques, syntaxiques et sémantiques de chaque langue. L'approche RBMT présente le défi de ne pas être en mesure d'écrire toutes les règles pour chaque langue en raison de l'expertise linguistique étendue et cruciale requise pour ce travail [11].

I.4.1.3. Traduction Interlingua (Interlingual MT)

L'interlingua est un terme utilisé pour décrire une langue qui sert d'intermédiaire entre deux langues différentes. Dans cette approche, la langue source (SL) est transformée en une langue intermédiaire (représentation) qui n'est pas influencée par les langues impliquées dans le processus de traduction. Par la suite, cette forme auxiliaire est utilisée pour indiquer précisément le verset traduit de la langue cible (TL). Cette méthode met l'accent sur l'utilisation d'une représentation unifiée pour plusieurs langues.

I.4.1.4. Traduction basée sur les transferts (TBMT)

Il est comparable à Interlingual-MT car il produit une traduction à partir d'une structure intermédiaire qui imite le sens du texte source. Le texte original est traduit dans une langue intermédiaire plus générale.

Ensuite, le formulaire est converti en une structure de langue cible possédant une structure similaire et le texte est produit dans la langue cible.

Le processus de transfert utilise les informations morphologiques, syntaxiques et/ou sémantiques des langues source et cible. Par conséquent, TBMT peut utiliser les connaissances des langues source et de destination.

I.4.2. Les approches basées sur le corpus

Les techniques basées sur le corpus emploient des méthodes empiriques pour garantir que toutes les connaissances linguistiques sont acquises par l'observation et l'analyse de corpus. Les corpus sont des collections d'ensembles de données parallèles comprenant des expressions source et cible qui sont traduites les unes dans les autres. [2]

I.4.2.1. Traduction basée sur les exemples (EBMT)

Le concept principal sous-jacent de (EBMT) est l'analogie. L'idée principale est de construire de nouvelles traductions basées sur des exemples existants. Les systèmes EBMT sont formés à l'aide de corpus parallèles bilingues constitués de paires de phrases.

Cet outil est conçu pour traduire des phrases qui semblent similaires en recherchant l'exemple source le plus comparable au mot ou à la phrase source dans des corpus parallèles. Nagao a catégorisé avec précision cette méthode en trois parties :

Chaque fragment est comparé à une base de données d'échantillons authentiques.

Identifier les éléments de traduction qui ont une forte corrélation (Alignement)

Combiner ces éléments pour former le texte souhaité.

I.4 .2.2. Traduction basée sur les statistiques (SMT)

Le système de traduction automatique statistique (SMT) produit des traductions potentielles dans une langue cible sur la base d'une phrase dans une langue source. Il sélectionne la traduction avec la probabilité conditionnelle la plus élevée *P(t|s)*. Le sens de traduction sera inversé en traduction.

La règle de Bayes sera appliquée pour incorporer le modèle (TM) *P(s|t)* et le modèle de langage (LM) *P(t)*. L'équation fournie est utilisée pour maximiser la probabilité d'obtenir la traduction la plus optimale.

$$\boldsymbol{t_{best} = argtmax\big(P(t|s)\big) = argtmax\big(P(s|t) \times P(t)\big)}\ \mathbf{1}$$

La valeur de *tbest* est l'argument qui maximise la probabilité de *P(t|s)*, ce qui équivaut à maximiser le produit de *P(s|t)* et *P(t)* comme le montre l'équation (1).

P(s|t) représente la probabilité de transition de l'état s à l'état t dans la machine de Turing (TM), tandis que P(t) représente la probabilité de l'état t dans le modèle de langage (LM).

Généralement, SMT (Statistical Machine Translation) nécessite l'utilisation d'un modèle de langage, d'un modèle de traduction et d'un mécanisme de décodage. La MT garantit que l'hypothèse cible générée correspond à la phrase source. En revanche, le modèle de langage (LM) garantit que le résultat est grammaticalement précis (Fig. 2).

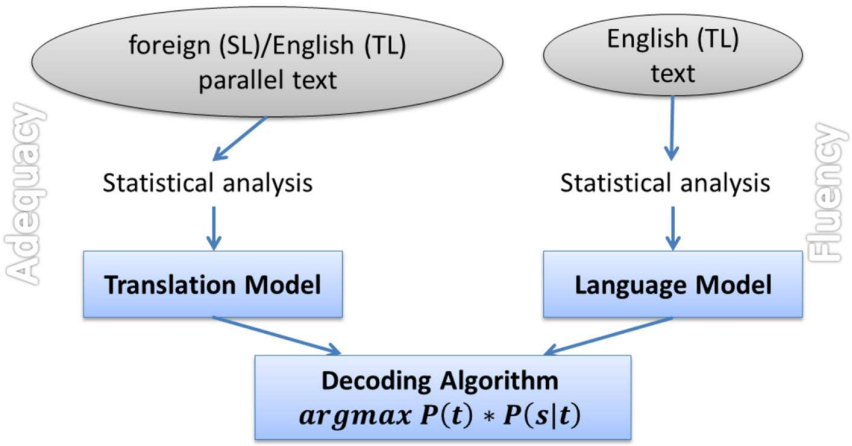


Figure 2. Approche de traduction automatique statistique

I.4.2.3. Traduction basée sur les réseaux neurones (NMT)

Traduction automatique neuronale Ce système utilise l'apprentissage profond et les réseaux neuronaux avec l'architecture.

Les modèles SMT conventionnels nécessitent une quantité importante de mémoire, mais les modèles NMT n'en nécessitent qu'une petite partie. De plus, contrairement aux systèmes de traduction conventionnels [21].

Tous les composants du modèle de traduction automatique neuronale (NMT) sont formés simultanément, ce qui signifie qu'ils sont appris du début à la fin afin d'optimiser les performances de traduction. Le paradigme Embed-Encode-Attend-Decode est l'approche la plus couramment utilisée dans la traduction automatique neuronale (NMT), comme le montre la figure 3. Les algorithmes les plus couramment utilisés basés sur ce modèle sont le réseau neuronal récurrent (RNN), conceptions de réseaux neuronaux convolutifs (CNN) et de réseaux d'auto-attention/feed-forward (SA/FFN). NMT a démontré des résultats de pointe pour de nombreuses paires de langues, ce qui est tout à fait remarquable [17] [18].

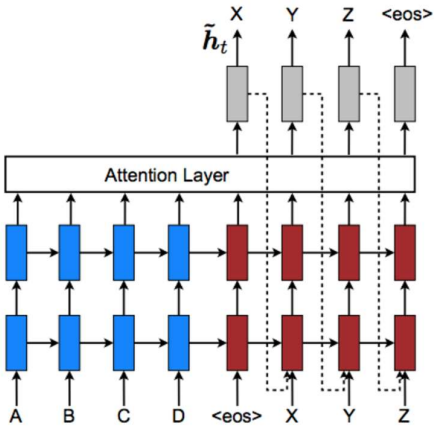


Figure 3. Modèle NMT basé sur le paradigme Encode-Attend-Decode

Récemment, l'approche dominante appelée modèles de langage basés sur des transformateurs (LM), telle que la représentation d'encodeur bidirectionnel à partir de transformateurs (BERT) [41], a incorporé des mécanismes d'attention pour gérer efficacement la tâche principale.

Au lieu d'utiliser un RNN comme base pour les représentations de séquences, il utilise l'auto-attention, qui est appliquée à deux états à l'intérieur de la même séquence [38]. Cependant, les modèles pré-entraînés ont démontré des performances impressionnantes dans un large éventail de tâches de PNL, en particulier celles liées à la compréhension du langage.

La maîtrise des modèles BERT et un effort manuel sont nécessaires pour accomplir avec succès la tâche MT. Une autre méthode d'intégration et d'utilisation de BERT dans la traduction automatique neuronale (NMT) consiste à extraire des représentations pour une séquence d'entrée. Ces représentations sont ensuite combinées avec chaque couche de codeur et de décodeur en utilisant des mécanismes d'attention. BART [42] utilise une structure seq2seq conventionnelle dotée d'un encodeur bidirectionnel. Il prédit les jetons masqués en exploitant l'intégralité de la séquence pour recueillir suffisamment d'informations pour des prédictions précises, à l'instar de BERT. De plus, BART utilise un décodeur de gauche à droite, similaire à GPT, pour prédire le mot suivant à l'aide d'un masque causal, comme le montre la figure 4.

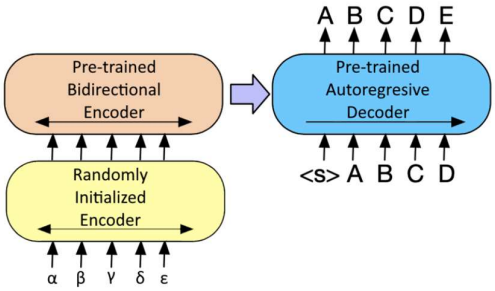


Figure 4. Implémentation BART pour la traduction automatique basée sur le modèle BERT

I.4.3. Evaluation

L'évaluation de la traduction automatique englobe diverses méthodes permettant d'évaluer l'efficacité d'un système de traduction automatique. Le score de qualité MT est une méthode utilisée pour évaluer les performances des systèmes de traduction automatique, fournissant une base fiable pour comparer l'efficacité de différents systèmes. Pour y parvenir, l'évaluation de la traduction automatique utilise des mesures mesurables.

L'évaluation des performances d'un système MT est nécessaire pour deux raisons principales. La première étape consiste à évaluer son adéquation à une mise en œuvre dans le monde réel. Le deuxième objectif est de servir de référence dans le processus de recherche et de développement.

Il existe deux méthodes distinctes pour évaluer les performances d'un système MT. La révision humaine implique une évaluation manuelle menée par des professionnels humains, tandis que l'évaluation automatique utilise des mesures basées sur l'IA spécialement conçues pour évaluer la qualité de la traduction sans avoir recours à une interaction humaine. Chaque option possède son propre ensemble d'avantages et d'inconvénients. Nous fournirons des informations plus complètes sur les deux types d'évaluation de la traduction automatique. [3][4]

I.4.3.1. Evaluation Humaine

L'évaluation humaine est la méthode la plus efficace pour évaluer la qualité de la traduction automatique, mais elle est plus coûteuse et plus longue [24].

La révision humaine est souvent considérée comme l'approche prédominante pour évaluer et quantifier la qualité des traductions [26]. Les spécialistes en linguistique et en traduction peuvent évaluer la qualité du résultat d'un système de traduction sous deux angles :

Aisance/Fluidité

Le degré de fluidité et de cohésion du texte traduit par rapport aux normes linguistiques de la langue cible, y compris l'exactitude grammaticale, la compréhensibilité et la lucidité. Au cours du processus d'évaluation de la maîtrise, les évaluateurs, qui maîtrisent uniquement la langue cible, ne reçoivent que la traduction évaluée. Ils n'ont pas accès au texte source, car celui-ci n'est pas important pour l'évaluation de la maîtrise.

Adéquation/Précision

La précision est un autre terme utilisé pour désigner ce concept. L'adéquation de la traduction fait référence à la mesure dans laquelle le texte cible correspond exactement au texte source, englobant l'utilisation de techniques d'expression et la représentation fidèle du contenu informatif.

Dans ce scénario, les évaluateurs qui maîtrisent à la fois la langue source et la langue cible ont la possibilité de réviser à la fois le texte original et les traductions évaluées. Ils considèrent systématiquement le contexte environnant de la phrase. [3][4]

I.4.3.2. Evaluation Automatique

L'évaluation automatique, en revanche, utilise des programmes spécialement conçus pour évaluer la qualité de la traduction automatique, mais elle est moins fiable au niveau des phrases mais constitue une option évolutive pour évaluer la qualité globale de la traduction sur plusieurs documents.

Le principal avantage de l'évaluation automatisée par rapport à l'évaluation humaine est sa capacité à évoluer efficacement. L'exécution de centaines d'instances d'évaluation informatisée est nettement plus rapide que la réalisation d'une seule série d'évaluations humaines. Cette fonctionnalité en fait une alternative parfaite pour effectuer des ajustements ou améliorer le système de traduction automatique, ce qui nécessite des résultats rapides.

Actuellement, il existe une variété de mesures d'évaluation automatique distinctes. Voici quelques illustrations de ceux actuellement utilisés :

BLEU

BLEU est largement considéré comme la mesure la plus populaire. C'est un système couramment utilisé pour évaluer la traduction automatique (TA) qui a démontré une forte association avec l'évaluation humaine. Le processus consiste à évaluer chaque phrase en la comparant à des traductions de référence précises, puis à calculer la note moyenne pour produire la note BLEU finale pour le système de traduction automatique. Le score quantifie le degré de similarité entre le résultat du système de TA et la traduction de référence effectuée par des humains, servant d'indicateur de qualité [3][4] [16]. Les scores BLEU sont calculés sur la base de n-grammes, qui sont des morceaux de texte consécutifs. La méthode évalue le nombre de chevauchements entre n-grammes, en attribuant un score allant de 0 à 1. Un score de 1 indique une correspondance parfaite entre la référence et la phrase de sortie. Les scores BLEU n'atteignent une signification statistique que lorsqu'ils sont considérés dans le cadre d'un échantillon de texte ou d'un corpus substantiel [32]. Des critères supplémentaires, tels que des sanctions en cas d'utilisation excessive de mots ou de phrases extrêmement brèves, ont également un impact sur le score BLEU. Bien que BLEU présente certaines limites, elle reste très appréciée et reste aujourd'hui la technique la plus couramment utilisée pour évaluer la traduction automatique [33].

COMET

La métrique multilingue optimisée pour l'évaluation de la traduction (COMET) est une méthode utilisée pour évaluer la traduction automatique multilingue à l'aide de techniques de classement et de régression. Il consiste en un estimateur et un modèle de classement des traductions qui reposent sur la détermination humaine. L'encodeur utilise une double modélisation multilingue, prenant en compte à la fois la source d'entrée et la phrase cible. COMET utilise des objectifs de formation pour simuler une évaluation humaine, prédire les scores de qualité de la traduction automatique et minimiser les écarts entre les résultats de qualité élevée et faible.

Les progrès de l'apprentissage profond dans la traduction automatique génèrent des résultats basés sur une compréhension plus approfondie de la langue, et pas seulement sur la substitution de mots. Pour évaluer la qualité de la traduction, COMET peut utiliser un encodeur multilingue pour comprendre plusieurs langues, montrant une forte corrélation avec l'évaluation humaine, abordant le problème des mots qui se chevauchent dans les résultats de traduction. [3][4] [16]

CHRF

Des enquêtes récentes ont démontré l'importance des n-grammes au niveau des caractères dans l'évaluation automatique, en particulier lorsqu'ils sont incorporés à des métriques plus complexes telles que MTERATER et BEER. Cependant, aucune recherche n'a été menée jusqu'à présent pour les examiner en tant que mesure distincte. À l'inverse, les F-scores dérivés des n-grammes, en particulier ceux qui sont linguistiquement motivés,

La recherche a démontré une forte corrélation entre les étiquettes et les morphèmes des parties du discours et les évaluations humaines. Il a été constaté que ces mesures surpassent les mesures largement utilisées comme BLEU. Nous considérons que ce score a le potentiel d'être utilisé comme mesure autonome. En effet, il a été démontré qu'il s'agit d'un élément important des mesures complexes évoquées plus haut. De plus, à l'instar du F-score basé sur les morphèmes, il prend en compte certains phénomènes morpho-syntaxiques. De plus, contrairement à d'autres mesures similaires, cette mesure est simple, car elle ne nécessite aucun outil ni source d'information supplémentaire. De plus, il est totalement indépendant du langage et de la tokenisation [34].

I.5. Conclusion

Dans ce chapitre, nous avons présenté une vue d'ensemble de la traduction automatique (MT), un domaine clé du traitement du langage naturel (NLP). Nous avons exploré les défis particuliers posés par la langue arabe, notamment sa morphologie complexe et son utilisation sur les réseaux sociaux. Malgré les avancées significatives dans les techniques de traduction, comme les méthodes statistiques et neuronales, ces défis persistent, rendant la traduction vers et depuis l'arabe particulièrement difficile.

Les diverses approches de traduction automatique, qu'elles soient basées sur des règles linguistiques, des corpus de données ou des réseaux neuronaux, ont chacune leurs avantages et limitations. Les méthodes linguistiques comme la traduction basée sur les règles (RBMT) et la traduction interlingua nécessitent des connaissances expertes étendues, tandis que les approches basées sur le corpus, telles que la traduction basée sur les exemples (EBMT) et la traduction statistique (SMT), s'appuient sur des données empiriques pour améliorer la précision. La traduction automatique neuronale (NMT) a démontré des résultats prometteurs grâce à l'utilisation de réseaux neuronaux et de modèles de langage avancés comme les transformateurs.

Nous avons également abordé l'importance de l'évaluation des systèmes de traduction automatique pour déterminer leur efficacité. Bien que les méthodes d'évaluation humaine soient précises, elles sont coûteuses et lentes. En revanche, les évaluations automatiques comme BLEU, COMET et CHRF offrent des alternatives évolutives mais moins précises.

En résumé, bien que la traduction automatique ait parcouru un long chemin, des défis significatifs demeurent, en particulier pour des langues comme l'arabe. Les progrès futurs devront continuer à affiner les méthodes existantes et à développer de nouvelles techniques pour surmonter ces obstacles et améliorer la fluidité et l'exactitude des traductions automatiques.

CHAPITRE II :L'AUGMENTATION DES DONNEES

II.1. Introduction

Le domaine du traitement du langage naturel (NLP) a récemment montré un intérêt croissant pour l'augmentation des données. Cela est principalement dû à l'attention accrue portée aux domaines à faibles ressources, à l'émergence de nouvelles tâches et à l'utilisation généralisée de réseaux neuronaux à grande échelle qui nécessitent des quantités substantielles de données de formation. Bien qu'il y ait eu une augmentation récente de l'activité, ce sujet particulier est encore relativement inexploré, peut-être en raison des difficultés présentées par les caractéristiques discrètes des données linguistiques.

L'augmentation des données (DA) englobe des techniques visant à améliorer la variété des instances de formation sans qu'il soit nécessaire de collecter activement des données supplémentaires. Récemment, la recherche sur l'apprentissage automatique (ML) a montré un intérêt significatif pour certaines techniques telles que l'UDA [38], qui utilisait la rétro-traduction [39], l'Auto Augment [40], Rand Augment [41] et MIXUP [42]. La vision par ordinateur (CV) sert souvent de champ d'exploration initial pour ces concepts, tandis que l'adaptation de domaine (DA) au traitement du langage naturel (NLP) est considérée comme secondaire et relativement inexplorée. Cela peut être attribué aux défis posés par la nature discrète du langage, qui empêche l'utilisation d'un bruit continu et rend plus difficile le maintien de l'invariance.

Qu'est-ce que l'augmentation des données ?

L'augmentation des données fait référence à la technique consistant à augmenter artificiellement la taille d'un ensemble de données en appliquant diverses transformations ou modifications aux données existantes. Ce processus contribue à améliorer la diversité et la variabilité de l'ensemble de données, ce qui améliore à son tour les performances et la robustesse des modèles d'apprentissage automatique. L'augmentation des données fait référence aux techniques utilisées pour améliorer la diversité des données de formation sans qu'il soit nécessaire de collecter directement des données supplémentaires.

II.2. Techniques et méthodes

II.2.1. Back-translation (Rétro-Traduction)

La traduction porte sur de vastes corpus parallèles constitués de textes alignés dans les langues source et cible. Cependant, la quantité de données monolingues disponibles est limitée, ce qui en fait une option plus appropriée pour former des modèles linguistiques. Dans le domaine de la traduction automatique neuronale (NMT), des efforts considérables ont été entrepris pour améliorer les modèles en utilisant des données monolingues. Ces efforts impliquent la fusion de modèles linguistiques, le processus de rétro-traduction et la pratique du double apprentissage. Ces méthodes offrent des avantages uniques et peuvent être utilisées pour atteindre une précision exceptionnelle, améliorant ainsi l'efficacité et l'efficience de la traduction automatique [22] [23] [27].

La rétro-traduction (BT) est une technique partiellement supervisée qui exploite des données bilingues et monolingues dans la langue souhaitée. Le processus implique la formation d'un système intermédiaire utilisant des données parallèles pour traduire les données monolingues cibles dans la langue source. Cela produit un corpus parallèle, dans lequel le côté source comprend un résultat de traduction automatique synthétique, tandis que le côté cible est constitué de matériel réel rédigé par des humains. Ensuite, le corpus parallèle artificiel est combiné avec le bitexte authentique pour créer le système de traduction ultime. Cette approche a démontré son utilité dans la traduction au niveau des phrases, la traduction automatique neuronale (NMT) et la traduction automatique non supervisée.

La recherche par faisceau ou la recherche gourmande Beam search or greedysearch sont couramment utilisées en rétro-traduction pour produire des phrases sources artificielles. Ces algorithmes sont des algorithmes d'approximation qui déterminent la sortie maximale a posteriori (MAP), qui fait référence à la phrase avec la probabilité estimée la plus élevée en fonction d'une entrée donnée. Néanmoins, l'utilisation de la prédiction Maximum A Posteriori (MAP) peut donner lieu à des traductions moins nuancées, en particulier dans les tâches qui impliquent un niveau d'incertitude important, comme la production de dialogues et d'histoires. Cela constitue un défi pour les méthodes d'augmentation des données telles que la rétro-traduction, car elles se concentrent principalement sur les résultats les plus probables du modèle, conduisant à la génération de phrases sources artificielles qui ne représentent pas de manière adéquate la distribution réelle des données.

Afin de résoudre ce problème, différentes approches sont suggérées, telles que le prélèvement d'échantillons à partir de la distribution du modèle et l'introduction de bruit dans les sorties de l'algorithme de recherche de faisceaux. L'échantillonnage sans contrainte produit une large gamme de résultats, même si certains d'entre eux peuvent être extrêmement improbables. D'un autre côté, nous examinons également un échantillonnage limité aux termes les plus probables. Ce texte examine la position intermédiaire entre MAP (maximum a posteriori) et échantillonnage illimité.

Une autre option consiste à introduire du bruit dans les sorties de la recherche de faisceaux, ce qui s'est révélé avantageux pour les configurations d'auto-encodeur. Les phrases sources subissent trois types de bruit : suppression de mots avec une probabilité de 0,1, remplacement de mots par un jeton de remplissage avec une probabilité de 0,1 et échange de mots sous forme de permutation aléatoire entre les jetons.

II.2.2. Right-Rotation Augmentation RRA (Augmentation par rotation à droite)

La langue arabe fait preuve d'une polyvalence exceptionnelle dans l'ordre des mots, ce qui la rend parfaitement adaptée à l'augmentation des données dans les applications de traduction automatique (TA) (voir Tableau I).

Order	**Sentence**	**English translation**
VSO	شرب عمر الماء	Drank Omar the water
OVS	الماء شرب عمر	The water drank Omar
SVO	عمر شرب الماء	Omar Drank the water
VOS	شرب الماء عمر	Drank the water Omar

Tableau 1. Ordre des mots flexible en arabe

En exploitant cette adaptabilité, une méthode efficace pour améliorer un ensemble de données de traduction automatique (TA) limitée consiste à tourner chaque phrase du corpus original vers la droite, le premier mot étant placé à la fin (car l'arabe s'écrit de droite à gauche)**.**

Cette transformation maintient le cadre fondamental et la construction de l'arabe, tout en produisant de nouvelles phrases avec une disposition des mots modifiée. En implémentant

cette méthode sur l'ensemble de données d'origine, nous pouvons l'améliorer avec de nombreuses phrases supplémentaires, augmentant ainsi la variété des instances de formation (voir l'algorithme 1). Ce corpus élargi expose le modèle de traduction automatique neuronale (NMT) à une plus grande variété d'arrangements de mots et de constructions de phrases, lui permettant d'extraire des informations plus subtiles et d'améliorer ses capacités de traduction. L'approche suggérée utilise la flexibilité naturelle de l'ordre des mots arabes pour améliorer l'augmentation des données et augmenter les performances des systèmes NMT [35].

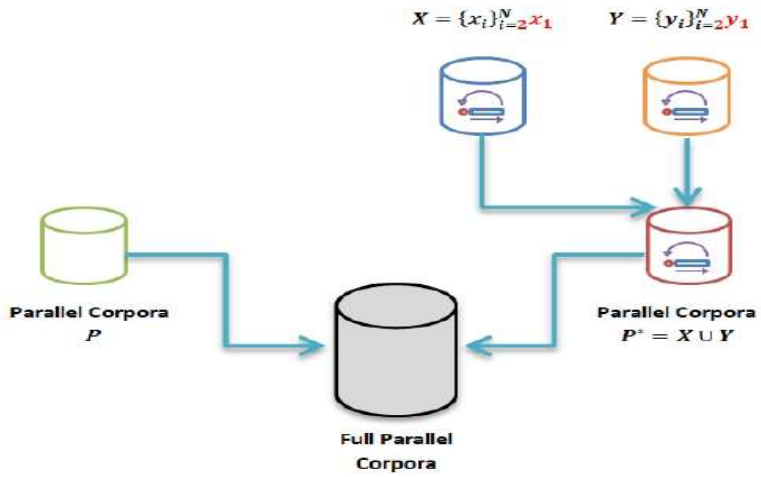


Figure 5. L'approche d'augmentation de rotation a droite

II.3. Conclusion

Dans ce chapitre, nous avons présenté un aperçu détaillé de l'augmentation des données dans le domaine du traitement du langage naturel (NLP). Nous avons examiné l'importance croissante de cette technique, notamment dans les domaines à faibles ressources et pour les modèles de réseaux neuronaux à grande échelle nécessitant des quantités substantielles de données de formation. L'augmentation des données se distingue par sa capacité à améliorer la diversité des instances de formation sans nécessiter la collecte active de nouvelles données.

Nous avons exploré diverses techniques et méthodes d'augmentation des données, en mettant particulièrement l'accent sur la rétro-traduction (BT). Cette technique exploite à la fois des données bilingues et monolingues pour créer des corpus parallèles synthétiques qui peuvent être utilisés pour améliorer la performance des modèles de traduction automatique. La recherche par faisceau et l'introduction de bruit dans les sorties ont été discutées comme

des moyens d'améliorer la diversité des phrases générées et de mieux représenter la distribution réelle des données.

En outre, nous avons présenté l'augmentation par rotation à droite (RRA), une méthode innovante qui introduit des variations structurées dans les données textuelles pour améliorer la robustesse et la performance des modèles d'apprentissage automatique.

En résumé, bien que l'augmentation des données dans le NLP soit un domaine relativement nouveau et encore en exploration, ses techniques montrent un potentiel significatif pour améliorer les modèles actuels. Les défis posés par la nature discrète du langage nécessitent des approches créatives et continues pour surmonter les obstacles et maximiser les bénéfices de ces techniques dans les applications linguistiques.

CHAPITRE III : LES RESEAUX DE NEURONES ET LA TRADUCTION AUTOMATIQUE

III.1. Introduction

La redécouverte des réseaux de neurones a considérablement amélioré divers domaines du traitement du langage naturel (NLP) (voir Goldberg, 2016 pour un examen complet). Néanmoins, l'incorporation de réseaux neuronaux dans les systèmes de traduction automatique (TA) était au début superficiel. Dans les premières étapes, les chercheurs ont utilisé des modèles de langage neuronal à action directe pour réévaluer les réseaux de traduction pour la langue souhaitée. Les modèles neuronaux initiaux, qui prenaient en compte la langue source, ont développé ce concept en utilisant le même modèle avec des tuples bilingues au lieu des mots de la langue cible. Soit-il eu évalué directement les paires de phrases à l'aide d'un réseau de rétroaction, soit ils ont incorporé une fenêtre de contexte source dans le modèle de langage neuronal Kalchbrenner et Blunsom (2013) et Cho et coll. (2014b) ont été les premiers à utiliser des réseaux récurrents pour la modélisation de la traduction. Ces systèmes utilisaient des réseaux de neurones comme éléments d'un système de traduction automatique statistique conventionnel. En conséquence, ils ont conservé la combinaison de modèles log-linéaires et ont simplement remplacé certains composants de l'architecture conventionnelle. La traduction automatique neuronale (NMT) a résolu cette division en employant un réseau neuronal étendu et solitaire qui convertit directement la langue source en phrase cible. L'émergence de la traduction automatique neuronale (NMT) constitue indéniablement une percée significative dans le domaine de la traduction automatique (MT), entraînant un changement radical et brutal dans la recherche dominante, s'éloignant des axes de recherche antérieurs. Cela ressort clairement de l'augmentation significative du nombre d'articles scientifiques sur la traduction automatique neuronale (NMT) ces dernières années (Figure 6), La traduction automatique neuronale (NMT) a été largement adoptée par plusieurs secteurs et est actuellement utilisée dans les systèmes opérationnels par de grands des entreprises telles que Google, Microsoft, Facebook, Amazon, SDL, Yandex et autres. Ce chapitre fournira une introduction aux principes fondamentaux de la traduction automatique neuronale (NMT) et offrira un aperçu complet des dernières avancées dans ce domaine de recherche. [5]

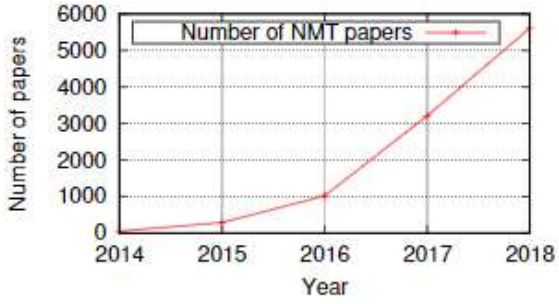


Figure 6. Nombre d'articles mentionnant la « traduction automatique neuronale » par an selon Google Scholar.

III.2. Les réseaux de neurones

Construction d'une machine neuronale Le modèle de traduction comporte trois étapes fondamentales : la tokenisation du texte, la formation du modèle NMT à l'aide d'appariements de phrases source et de destination tokenisées, et le décodage ou la traduction. Au départ, chaque [6]

Une phrase est convertie en une série de jetons, qui peuvent être constitués de mots, de sous-mots ou de caractères individuels. La séquence de jetons sera convertie en une séquence entière. Au cours de la deuxième étape, une fois qu'une architecture neuronale est sélectionnée, les paramètres du modèle NMT sont ajustés en fonction d'un objectif d'entraînement. Pendant l'entraînement, le modèle NMT reçoit une paire de séquences d'entiers qui représentent une phrase source et une phrase cible. Au cours de la dernière étape, une fois entraîné, le modèle NMT est capable de produire une traduction de n'importe quelle phrase dans la langue source en utilisant une technique de décodage telle que la recherche par faisceau (Koen, 2004). Lors de l'étape d'inférence, le modèle NMT ne prend en compte que la séquence de nombres qui représente la phrase source. [6]

III.2.1. Les réseaux Feed-Forward Networks (FFN)

Les réseaux de neurones à flux direct sont un type de réseaux de neurones artificiels dans lesquels les nœuds sont disposés de manière à ce qu'il n'y ait pas de boucles. Ce réseau neuronal est communément appelé réseau neuronal multicouche car les informations ne sont transmises que dans le sens aller.

Au cours du processus de flux de données, les nœuds d'entrée reçoivent des données, qui traversent ensuite des couches cachées avant de finalement sortir par les nœuds de sortie. Il n'existe aucun lien disponible dans le réseau pouvant être utilisé pour transmettre des informations à partir du nœud de sortie.

Un réseau neuronal à rétroaction se rapproche des fonctions en faisant passer les données d'entrée à travers une série de couches interconnectées, sans aucune connexion de rétroaction.

Un algorithme dérive des classificateurs en utilisant la formule $y = f^*(x)$.

L'entrée x est classée comme y.

Le modèle feed-forward indique que la variable de sortie y est déterminée par la variable d'entrée x et les paramètres θ, via la fonction f. Cette valeur définit l'approximation la plus précise de la fonction.

Lorsque le réseau neuronal à rétroaction est simplifié, il peut être représenté comme un perceptron à une seule couche.

Ce modèle effectue une opération de multiplication entre les entrées et les poids au fur et à mesure de leur réception par la couche. Par la suite, les valeurs d'entrée auxquelles sont attribuées les pondérations sont combinées pour obtenir la somme totale. Si le total des valeurs dépasse un seuil spécifique, fixé à zéro, la valeur résultante est généralement 1. À l'inverse, si le total tombe en dessous du seuil, la valeur résultante est généralement -1.

Le perceptron monocouche, qui est un type de modèle de réseau neuronal à rétroaction, est couramment utilisé pour les tâches de classification. Les perceptrons monocouches peuvent être couplés à l'apprentissage automatique. Les réseaux de neurones peuvent modifier leurs poids à l'aide de la règle delta pendant l'entraînement, leur permettant ainsi de comparer leurs sorties avec les valeurs souhaitées.

Le déclin progressif est une conséquence de la formation et de l'apprentissage. De même, les perceptrons multicouches modifient leur poids. Cependant, cette procédure est communément appelée rétro-propagation. Dans un tel scénario, les couches cachées du réseau seront modifiées en fonction des valeurs de sortie générées par la couche finale. [7]

III.2.2. Les réseaux de neurones récurrents (RNN)

Cette section donne un aperçu de la première architecture utilisée dans la traduction automatique neuronale (NMT), connue sous le nom d'architecture de traduction automatique

neuronale récurrente (RNMT). Le RNMT se compose d'un encodeur récurrent, d'un décodeur récurrent et de tables d'incorporation de mots. L'encodeur récurrent et le décodeur récurrent [8]

En règle générale, les réseaux de neurones récurrents (RNN) utilisent des types de couches similaires, telles que l'unité récurrente fermée (GRU) et la mémoire à long terme (LSTM), qui seront développées plus en détail dans la section suivante. Le modèle RNMT est intrinsèquement auto-régressif, ce qui signifie que chaque état caché dans le codeur et le décodeur doit passer par chaque état intermédiaire pour accéder aux informations à partir de n'importe quel pas de temps précédent. Les états cachés de la traduction automatique neuronale récurrente (RNMT) préservent les informations de commande, ce qui lui confère un avantage par rapport à la traduction automatique neuronale convolutive (CNMT) et à la traduction automatique neuronale basée sur l'attention (ANMT). Néanmoins, l'absence de corrélations directes entre les lieux de la séquence d'entrée présente de nombreux défis dans la formation à la traduction automatique neuronale récurrente (RNMT). L'un de ces défis est le problème de gradient de disparition rencontré lors de la rétropropagation dans le temps. [8]

III.2.3. Long Short-Term Memory (LSTM)

Mémoire à long terme et à court terme Hochreiter et Schmidhuber (1997) ont initialement conçu la méthode de déclenchement pour les réseaux neuronaux récurrents (RNN) afin de faciliter la transmission transparente des informations et des gradients depuis des positions distantes vers la position actuelle. LSTM implémente formellement l'équation 2.3. [9]

Comme indiqué ci-dessous :

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \mathbf{c}^{\sim}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} \left[w_{lstm} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b_{lstm} \right],$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \mathbf{c}^{\sim}_t,$$

$$h_t = o_t \odot tanh(c_t),$$

Soit W_{lstm} une matrice de taille R4d×2d et b_{lstm} un vecteur de taille R4d. Le symbole représente la multiplication par élément. Les fonctions $\sigma(\cdot)$ et $tanh(\cdot)$ représentent respectivement les fonctions d'activation tangente sigmoïde et hyperbolique. En plus de l'état

caché h_t, LSTM fournit un état récurrent distinct appelé cellule mémoire ct, spécialement conçu pour conserver et stocker les informations précédentes. La mise à jour des cellules mémoire, telle que décrite par l'équation 2.5, suit une structure linéaire. Cette structure linéaire permet à ct de conserver des informations anciennes, améliorant ainsi la capacité de modélisation à longue portée du LSTM.

Le modèle LSTM régule son comportement interne à l'aide de trois portes : la porte d'entrée (it) contrôle le flux d'informations de l'entrée actuelle vers la cellule mémoire, la porte d'oubli (ft) détermine la quantité d'informations à conserver de l'état précédent de la cellule, et la porte de sortie (ot) surveille les informations qui sont transmises à l'état caché de sortie. La coopération de ces portes permet la transmission efficace des gradients et simplifie l'optimisation du modèle. L'optimisation joue un rôle en guidant le processus d'apprentissage de ces portes et influence leur capacité à fournir des représentations de modèles significatives. Tous les facteurs contribuent aux performances exceptionnelles du LSTM sur un large éventail de tâches, y compris les fonctions d'activation sigmoïde et tangente hyperbolique (tanh(·)).[9]

III.3. La traduction automatique neuronale (NMT)

Les modèles de traduction automatique neuronale (NMT) ont amélioré le niveau technologique actuel en construisant un réseau neuronal solitaire plus efficace pour acquérir des connaissances et une compréhension [30].

Le réseau neuronal comprend deux composants : Un réseau d'encodeurs est utilisé pour transformer la phrase d'entrée en une série de représentations distribuées. Ces représentations sont ensuite utilisées par un réseau de décodeurs, qui intègre un modèle d'attention, pour générer la traduction. Plusieurs modèles de traduction automatique neuronale (NMT) ont été développés sur la base de ce paradigme codeur-décodeur, ce qui a permis d'améliorer les performances des systèmes de traduction automatique. NMT a la capacité de mieux appliquer ses connaissances à un nouveau texte en utilisant les similitudes entre les mots dans les intégrations et en considérant des contextes plus larges pour capturer la réorganisation à longue distance de manière continue. [10]

III.3.1. Apprentissage des modèles de traduction automatique neuronale

L'objectif de cette étude est de présenter une méthode complète pour améliorer la résilience des modèles NMT contre les perturbations d'entrée. Notre concept fondamental est d'assurer la cohérence des comportements dans le modèle NMT à la fois pour la phrase source originale x et sa version modifiée x 0. Le modèle NMT se compose de deux techniques pour transformer une phrase source x en sa phrase cible correspondante y. Le codeur code x dans une séquence de représentations Hx, tandis que le décodeur génère y en utilisant Hx comme entrée. Notre objectif est d'acquérir des connaissances sur le codeur et le décodeur invariants aux perturbations. [10]

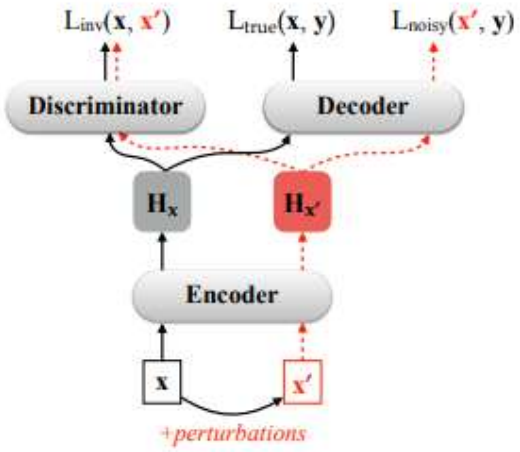


Figure 7. L'architecture de NMT avec formation à la stabilité contradictoire.

Les lignes sombres et pleines de flèches représentent le flux d'informations de passe avant pour la phrase d'entrée x, tandis que la flèche pointillée rouge lignes pour la phrase d'entrée bruyante x0, lequel est trans formé à partir de x en ajoutant de petites perturbations [31].

La figure 7 décrit la structure de notre méthodologie. Étant donné une phrase d'entrée x, nous générons un ensemble de phrases modifiées N(x), où chaque phrase x0 est créée en introduisant des modifications mineures à x. Pour que x soit considéré comme une variation subtile de x0, il faut qu'ils possèdent une sémantique identique.

Lorsque nous recevons la paire d'entrées (x, x0), nous avons deux attentes : (1) La représentation codée Hx0 doit être similaire à Hx. (2) Si on nous donne Hx0, le décodeur devrait être capable de produire la sortie fiable y.

III.3.2. Décodage

Décodeur pour transformateurs Le décodeur de transformateur a une structure similaire à celle du codeur, avec l'ajout d'une sous-couche d'attention codeur-décodeur [9]. Cette sous-couche est utilisée pour aligner et extraire les informations pertinentes pour la traduction de la source en fonction de toutes les sorties du codeur XL [9] :

$$\bar{y}^{l} = transformerdecoderlayer^{(Y^{l-1},X^{L})} \Rightarrow$$

$$\bar{y}^{l} = \underbrace{LN\left(Rc\left(Y^{l-1}, MHATT(Y^{l-1}, X^{l-1})\right)\right)}_{decoder\ (masked) self-attention}$$

$$\ddot{y}^{l} = \underbrace{LN\left(Rc\left(\bar{y}^{l}, MHATT(\bar{y}^{l}, X^{l})\right)\right)}_{encoder-decoder\ or\ cross\ attention}$$

$$Y^{l} = LN\left(Rc\left(\ddot{y}^{l}, FFN(\ddot{y}^{l})\right)\right)$$

Les sorties des sous-couches d'auto-attention et d'attention du codeur-décodeur sont notées Y¯ l ∈ R |Y|×d et Y¨ l ∈ R |Y|×d, respectivement. Le modèle Transformer utilise la génération auto-régressive comme paramètre par défaut, ce qui signifie que la sous-couche d'auto-attention du décodeur ne permet pas aux mots cibles de s'occuper de leurs futurs homologues. Pour ce faire, l'algorithme Transformer applique un masque causal à la matrice de masquage de l'équation 2.11, où les indices i et j vont de 1 à la longueur de Y [28] [29].

$$m_{i,j} = \begin{cases} 0, & if\ i \geq j \\ -\infty, & \text{otherwise} \end{cases}$$

Par conséquent, nous appelons le décodeur l'auto-attention cachée.

Le décodeur auto-régressif est largement accepté comme approche standard pour le modèle Transformer. Il modélise directement la dépendance entre les jetons du côté cible et atteint des performances de pointe sur la plupart des tâches de génération. Cependant, il présente l'inconvénient d'une inférence lente lors du décodage.

Bien que le processus de formation puisse être entièrement parallélisé, l'étape d'inférence est grandement entravée par la prédiction séquentielle. L'architecture sous-jacente du décodeur, la dépendance de l'auto-attention du décodeur à l'égard des états antérieurs et l'attention croisée ont tous un impact néfaste sur l'efficacité du processus d'inférence. [9]

III.4. Conclusion

Ce chapitre nous a permis de présenter une exploration détaillée des réseaux neuronaux appliqués à la traduction automatique (TA), en mettant en évidence leur évolution et leurs contributions significatives au domaine du traitement du langage naturel (NLP). Nous avons examiné les premières approches qui incorporaient superficiellement les réseaux neuronaux dans les systèmes de TA, en passant par les modèles de langage neuronal à action directe et les réseaux récurrents utilisés pour la modélisation de la traduction.

Nous avons ensuite discuté de la traduction automatique neuronale (NMT), qui a marqué un tournant décisif dans la recherche en TA en proposant un réseau neuronal solitaire étendu capable de convertir directement une langue source en une phrase cible. L'architecture NMT, avec ses composants encodeur et décodeur, a amélioré de manière significative les performances des systèmes de traduction en capturant les similitudes entre les mots et en prenant en compte des contextes plus larges.

Les différents types de réseaux de neurones, tels que les réseaux de neurones à flux direct (FFN), les réseaux de neurones récurrents (RNN) et les modèles de mémoire à long terme et à court terme (LSTM), ont été analysés pour comprendre leurs rôles et leurs contributions respectives dans l'amélioration des modèles de NMT. Chaque type de réseau a apporté des solutions uniques aux défis posés par la modélisation de la traduction, en particulier en termes de gestion de la séquence et de la transmission d'informations sur de longues distances.

Enfin, nous avons abordé les techniques d'apprentissage et de décodage utilisées dans les modèles NMT, mettant en avant l'importance de la formation à la stabilité contradictoire et des techniques de décodage auto-régressif. Ces méthodes ont permis d'améliorer la robustesse des modèles NMT face aux perturbations d'entrée et d'optimiser le processus de traduction pour obtenir des résultats plus précis et fiables.

En somme, ce chapitre a fourni une vue d'ensemble des avancées technologiques et des innovations dans le domaine de la traduction automatique neuronale, soulignant l'impact majeur des réseaux neuronaux sur la qualité et l'efficacité des systèmes de traduction. Les défis restants et les directions futures de recherche ouvrent la voie à de nouvelles explorations pour améliorer encore davantage les capacités des modèles de NMT.

CHAPITRE IV : CONCEPTION DE LA METHODE PROPOSEE

IV.1. Introduction

La traduction automatique est l'un des domaines clés du traitement du langage naturel. Notre projet vise à comparer les performances d'un modèle de traduction automatique lors de l'utilisation de différentes techniques d'augmentation de données. Dans cette section, nous expliquerons les étapes que nous avons suivies pour préparer les données, les nettoyer, entraîner le modèle et l'évaluer à l'aide de la bibliothèque OpenNMT.

IV.2. Prétraitement

Le prétraitement est une étape essentielle et vitale dans les projets de traitement linguistique et les modèles de traduction automatique. Cela implique de nettoyer et préparer les données de manière à les rendre adaptées à une utilisation dans les modèles de formation. Cette étape nécessite une compréhension approfondie des données et des technologies utilisées pour améliorer leur qualité et leur précision.

IV.2.1. Nettoyage des données

Les données brutes contiennent souvent du bruit et des éléments indésirables tels que des signes de ponctuation et des caractères spéciaux. Nous avons nettoyé les données pour garantir la qualité des données saisies dans le modèle. Le nettoyage des données comprend les étapes suivantes :

- Convertir le texte en caractères ASCII : pour standardiser l'encodage et éviter les problèmes avec les caractères spéciaux.
- Supprimez les signes de ponctuation : pour éviter leur impact négatif sur le processus d'apprentissage.
- Supprimer les espaces de fin : pour standardiser le formatage du texte.

IV.2.2. Préparation des données

Après avoir nettoyé les données, nous les avons divisées en ensembles de formation, de validation et de test. Cela aide à évaluer objectivement les performances du modèle.

- **Ensemble d'entraînement :** utilisé pour entraîner le modèle.
- **Ensemble de vérification :** utilisé pour ajuster les paramètres et garantir qu'il n'y a pas de "overfitting".
- **Ensemble de test :** utilisé pour évaluer les performances finales du modèle.

Après avoir préparé les données en les nettoyant et en les divisant en ensembles de formation, de validation et de test. Nous écrivons ces données dans des fichiers texte afin qu'OpenNMT puisse les lire facilement.

IV.3. Description du modèle utilisé

OpenNMT est un écosystème open source pour la traduction automatique neuronale et l'apprentissage de séquences neuronales, Lancé en décembre 2016 par le groupe Harvard NLP et SYSTRAN, le projet a depuis été utilisé dans plusieurs applications de recherche et industrielles. Il est actuellement entretenu par SYSTRAN et Ubiqus.

La structure séquence à séquence, présentée pour la première fois par [43], a servi de base au modèle utilisé dans cette enquête. Le réseau codeur-décodeur qui utilise des cellules d'unités récurrentes fermées (GRU) constitue ce système. (Voir la figure 8). Tout au long de la phase de formation, le travail de l'encodeur consiste à reconnaître les intégrations de la phrase d'entrée et à générer un vecteur de contexte qui capture les idées principales de la phrase. Après cela, le décodeur utilise ce vecteur pour créer la phrase cible, en commençant par la saisie du symbole.

Pour produire la sortie actuelle (yt), le procédé consiste à alimenter le décodeur, pour chaque pas de temps (t), l'état caché (dt−1) et la sortie prédite précédente (yt−1). Avant la rétro-propagation, ceci est utilisé pour calculer la perte et appliquer des gradients. La méthode d'inférence prend en entrée la sortie de l'encodeur, l'état caché et les prédictions précédentes, tandis que le forçage de l'enseignant est utilisé pour la formation. Lorsque le modèle prédit le jeton qui indique la fin de la phrase, il s'arrête.

Attention [98], qui donne des poids à chaque mot de la phrase source lors de la génération du mot cible, améliore le modèle. Le vecteur de contexte est généré en ajoutant les états cachés de l'encodeur, et les poids d'attention, qui classent les états cachés en fonction de leur importance pour la production du mot cible au temps (t), sont obtenus en appliquant une fonction d'alignement et la fonction SoftMax.

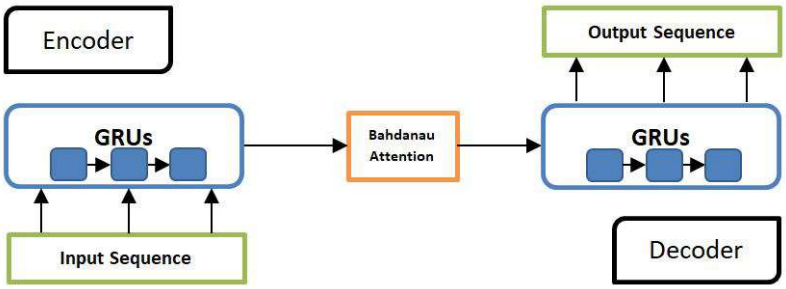


Figure 8. Schéma du modèle Seq2Seq basé sur GRU avec Couche d'attention de Bahdanau

IV.3.1. Préparer des données à l'aide de BPE et créer un dictionnaire de mots

L'utilisation de BPE dans la préparation des données constitue une étape importante dans le traitement du langage naturel, car elle permet de réduire la taille des données traitées et d'augmenter l'efficacité du modèle. La construction de dictionnaires joue également un rôle clé dans l'amélioration de la qualité de la traduction en fournissant un ensemble défini de mots possibles, ce qui réduit les erreurs de traduction et améliore la précision globale du modèle.

IV.3.1.1. Copie de fichiers

Les fichiers sont copiés sous de nouveaux noms pour être utilisés ultérieurement. Les textes source et cible sont conservés dans des fichiers séparés pour faciliter les processus d'apprentissage et de candidature.

IV.3.1.2. Apprentissage des codes BPE

Le modèle apprend les codes BPE à partir de textes de formation, une technique permettant de compresser les textes et de réduire la taille du dictionnaire tout en conservant une capacité de décodage efficace.

IV.3.1.3. Application BPE

Les codes appris sont appliqués aux scripts de formation, de validation et de test, aidant à transformer les scripts en une représentation compacte plus facile à traiter.

IV.3.1.4. Construction du dictionnaire

Un dictionnaire de mots est créé et utilisé dans le modèle pour identifier les mots possibles et les traductions possibles. La taille du dictionnaire est déterminée pour garantir des performances optimales du modèle.

IV.3.2. Configurer les paramètres du model

Un fichier YAML est créé contenant les paramètres du formulaire, notamment :

IV.3.2.1. Localisation des données : Nous définissons les chemins des dossiers de formation et de validation.

IV.3.2.2. Paramètres de l'optimiseur

- optimizer: type d'optimiseur utilisé dans la formation. Dans cet exemple, nous utilisons AdamOptimizer, connu pour son efficacité dans l'optimisation des modèles approfondis.
- optimizer_params: paramètres d'optimisation supplémentaires tels que beta1 et beta2, qui sont des paramètres utilisés pour ajuster la façon dont les pondérations du modèle sont mises à jour.
- learning_rate: taux d'apprentissage, un paramètre qui détermine la rapidité avec laquelle les poids du modèle peuvent être ajustés pendant l'entraînement.

IV.3.2.3. Paramètres d'entraînement

- batch_size: la taille du lot utilisée pour l'entraînement. Des lots plus importants signifient des mises à jour moins fréquentes mais plus stables.
- train_steps: le nombre d'étapes d'entraînement. Le modèle est mis à jour à chaque étape.
- save_checkpoints_steps: Le nombre d'étapes avant d'enregistrer le point de contrôle. Les points de contrôle permettent de restaurer le modèle à partir d'un certain stade en cas d'arrêt brutal.
- keep_checkpoint_max: le nombre de points de contrôle conservés. Il permet de conserver les anciennes versions du modèle.
- save_summary_steps: nombre d'étapes avant d'enregistrer un résumé. Cela aide à suivre les performances du modèle pendant l'entraînement.
- single_pass: indique si une seule passe doit être effectuée sur les données d'entraînement. Si c'est vrai, l'entraînement se termine après un passage dans les données.

- maximum_features_length et maximum_labels_length: longueur maximale de la séquence de fonctionnalités et d'étiquettes pendant l'entraînement. Aide à réduire la mémoire utilisée et à ajuster les performances.

IV.3.2.4. Paramètres d'évaluation

- batch_size: taille du lot utilisée pour la vérification. Des lots plus petits peuvent être plus rapides mais moins précis.
- num_threads: le nombre de threads utilisés pour traiter les données en parallèle. Cela peut accélérer le processus de vérification.
- prefetch_buffer_size: nombre de lots à pré-extraire de manière asynchrone. Aide à améliorer les performances.
- save_eval_predictions: s'il faut enregistrer les prédictions d'évaluation. Si c'est vrai, les résultats sont enregistrés pour une évaluation ultérieure.
- Exporters: exportateurs modèles utilisés après évaluation. Dans cet exemple, seul le modèle le plus récent est conservé.
- eval_delay: évaluation toutes les secondes données. Cela aide à suivre périodiquement les performances du modèle.
- external_evaluators: évaluateurs externes utilisés pour évaluer le modèle. Dans cet exemple, nous utilisons la métrique BLEU et CHRF.

IV.3.2.5. Paramètres d'inférence

- batch_size: la taille du lot utilisée dans la traduction lors de l'inférence. Des lots plus importants signifient une traduction plus rapide mais nécessitent plus de mémoire.

IV.3.3. Entraînement du modèle

La commande "onmt-main train_and_eval" est utilisée pour entraîner le modèle. Cette commande prend en compte les paramètres du modèle du fichier YAML et démarre le processus de formation à l'aide du GPU pour améliorer les performances.

IV.3.4. Évaluation du modèle

Nous avons utilisé deux mesures pour évaluer la qualité des traductions obtenues par rapport aux traductions de référence :

- BLEU (Étude d'évaluation bilingue).
- ChrF (caractère n-gramme F-score).

IV.3.3.1. La métrique BLEU

L'échelle BLEU est l'une des mesures les plus populaires pour évaluer la traduction automatique. Il compare les phrases traduites avec un ensemble de traductions de référence et est basé sur des n-grammes correspondants.

Étapes pour calculer BLEU :

- Analyser les phrases en n-grammes : nous segmentons les phrases traduites et les phrases de référence en n-grammes.
- Calcul de la précision par n-gramme : Nous calculons le pourcentage de n-grammes présents dans la phrase traduite qui sont également présents dans la phrase de référence.
- Calculer la pénalité de brièveté (BP) : pour éviter de récompenser des phrases trop courtes.
- Calcul du BLEU : nous prenons la précision moyenne sur tous les n-grammes (généralement 1 à 4) et la multiplions par la pénalité de brièveté.

Équation BLEU :

BLEU = $BP \cdot exp(\sum_{n=1}^{N} \mathrm{w}_n log p_n)$

Où :

BP est une pénalité de brièveté.

w_n C'est le poids de n-gramme (généralement égal).

p_n est la précision pour n-gramme [32].

IV.3.3.2. La métrique ChrF

La mesure CHRF est basée sur le calcul du score F pour faire correspondre les n-grammes au niveau des lettres.

Étapes pour calculer le ChrF :

- Analyser les phrases en n-grammes de caractères : nous segmentons à la fois les phrases traduites et les phrases de référence en n-grammes au niveau des caractères.
- Calcul de la précision et du rappel : Nous calculons le pourcentage de n-grammes dans la phrase traduite qui sont également présents dans la phrase de référence (précision) et vice versa (rappel).
- Calcul du F-score : basé sur la précision et le rappel.

Équation ChrF :

$$\text{ChrF} = (\beta^2 + 1)\cdot \frac{precision \cdot recall}{recall + (\beta^2 \cdot precision)}$$

Précision : le pourcentage de n-grammes présents dans la phrase traduite qui sont présents dans la phrase de référence.

Rappel : le pourcentage de n-grammes de caractères de la phrase de référence qui se trouvent dans la phrase traduite.

β : facteur de pondération entre précision et rappel (généralementβ = 2) [34].

IV.4. Étapes de la technique de Rétro-Traduction (Back-Translation)

Après avoir formé un modèle pour traduire des textes du dialecte vers l'arabe classique en utilisant l'ensemble de données disponibles.

- Nous utilisons ce modèle pour traduire un autre groupe de textes familiers DZA en arabe classique MSA.
- Nous utilisons ensuite un autre modèle entraîné pour traduire des textes de l'arabe classique vers le dialecte.
- Nous traduisons les textes obtenus lors de la première étape (textes classiques MSA) en textes familiers DZA.
- Nous combinons les textes originaux et générés en un seul ensemble de données pour obtenir un ensemble plus grand.

IV.5. Étapes de la technique de Rotation à Droite (Right Rotation) :

Nous prenons chaque phrase de l'ensemble de données d'origine et déplaçons le premier mot de la phrase vers la fin de la phrase.

- Après avoir déplacé l'emplacement du premier mot, il crée des copies supplémentaires des phrases dans des ordres différents.
- Phrase originale == traduction originale
- Phrase modifiée == traduction originale
- En utilisant la même technique pour les deux groupes, nous avons obtenu 4 fois plus pour chaque phrase comme suit :

MSA ==== DZA où : MSA est la phrase arabe originale

MSA` ==== DZA MSA` est la phrase arabe moderne

MSA ==== DZA` DZA est la phrase darija originale

MSA` ==== DZA` DZA` est la phrase darija moderne

Ainsi, nous avons quadruplé l'ensemble des données.

Algorithme : *Augmentation par Rotation à Droite*

Input: A sentence pair S_1-T_1 (where S_1 represents the source sentence and T_1 denotes the target sentence)

Output: Four augmented sentence pairs: S_1--T_1, S'_1--T_1, S_1--T'_1, and S'_1--T'_1

/ Apply Right Rotation to S_1;*

1. S'_1 ← Rotate the source sentence S_1 by moving the first word to the end;

/ Apply Right Rotation to T_1;*

2. T'_1 ← Rotate the target sentence T_1 by moving the first word to the end;

3. **return** Original sentence pair S_1--T_1, Augmented pair S'_1--T_1, Augmented pair S_1-T'_1, Augmented pair S'_1-T'_1

IV.6. Conclusion

Dans ce chapitre, nous avons présenté en détail toutes les étapes de la méthode proposée dans ce travail afin de réduire la dimension des différentes bases de tests, et pour pouvoir mieux classifier ou prédire les classes des nouvelles instances issues de la réduction de dimensionnalité.

CHAPITRE V : TESTS ET RESULTATS

V.1. Introduction

Dans ce chapitre, nous passerons en revue les ressources utilisées au cours du projet, du matériel aux environnements virtuels et aux outils qui ont facilité notre travail. De plus, nous présenterons et analyserons les résultats obtenus par notre application de différentes méthodes d'augmentation des données **.**Avec une analyse des raisons possibles de l'écart de performances entre ces méthodes. Nous chercherons également à formuler des recommandations qui, nous l'espérons, aideront à orienter les recherches futures dans ce domaine.

V.2. Matériel et logiciels utilisés

Pour mener à bien notre travail, nous avons choisi de travailler dans un environnement expérimental caractérisé par un PC doté de 8 Go de RAM, 1 CPU i5-8365U cadencé à 1.60GHz, tournant sous *Windows 10* 64 bits. En ce qui concerne le coté logiciel, notre développement est fait sur *Anaconda* qui est doté de *Python* 3.7 ainsi que l'éditeur *Jupiter* 3.3.4 qui a été choisi pour sa simplicité et efficacité. Plusieurs bibliothèques ont été utilisées : *OpenNMT*-tf, *Tensorflow, numpy, pandas, sklearn-pandas,nltk, scikit-learn*. Ainsi que la plateforme Google Colab.

Bibliothèque	**Version utilisée**
OpenNMT-tf	2.32.0
Tensorflow	2.15.0
Numpy	1.16.3
Pandas	0.24.2
sklearn-pandas	2.2.0
scikit-learn	1.2.2
nltk	3.8.1
System (python)	3.7.3 (default, Apr 24 2019, 15:29:51) [MSC v.1915 64 bit (AMD64)]

Tableau 2. Tableau des bibliothèques python utilisées

Parmi toutes ces bibliothèques utilisées dans ce travail, la plus importante est bien *Tensorflow*.Cette dernière est une bibliothèque de logiciels open source destinée au calcul numérique de haute performance.Son architecture flexible permet un déploiement facile du calcul sur diverses plates-formes (processeurs*CPU*, processeurs graphiques*GPU*, processeurs tenseurs *TPU*), et des ordinateurs de bureau aux clusters de serveurs, aux périphériques mobiles et périphériques.La bibliothèque *Tensorflow*[1] a été développée à l'origine par des chercheurs et des ingénieurs del'équipe*Google Brain* au sein de l'organisation *GoogleAI*, elle prend en charge l'apprentissage automatique et l'apprentissage profond et le cœur de calcul numérique flexible est utilisé dans de nombreux autres domaines scientifiques.

V.3. Description de la base de données utilisée

Vous trouverez ci-dessous une brève description de la base de données utilisée dans l'étude :

Dataset	Phrases	DZDA			MSA		
		Tokens	Vocab.	Taille moyenne	Tokens	Vocab.	Taille moyenne
Baseline	10,212	119,664	11,926	12.71	132,512	11,491	14.04
BT.Corpus	63,998	652,699	41,494	11.77	665,547	37,912	11.99
RR.Corpus	37,169	435,604	11,926	12.59	480,588	11,491	13.87

Tableau 3. Statistiques des ensembles de données MSA-DZDA utilisés

L'ensemble de données de Baseline contient 10 212 phrases. Avec un total de 119 664 tokens, une taille de vocabulaire de 11 926 et une taille moyenne de phrase de 11,77 caractères pour DZDA. Quant au MSA, il contient 132 512 tokens, une taille de vocabulaire de 11 491 et la taille moyenne des phrases est de 11,99 caractères.

L'ensemble de données élargi à l'aide de la rétro-traduction "BT. Corpus" contient 63 998 phrases, avec un total de 652 699 tokens et une taille de vocabulaire de 41 494. La taille moyenne des phrases est de 12,71 caractères pour DZDA.

Il contient 665 547 tokens et une taille de vocabulaire de 37 912. La taille moyenne des phrases est de 14,04 caractères pour MSA.

Enfin, l'ensemble de données élargi à l'aide d'une rotation à droite" RR. Corpus " contient 37 169 phrases, avec un total de 435 604 tokens et une taille de vocabulaire de 11 926. La taille moyenne des phrases est de 12,59 caractères.

Le MSA contient 480 588 tokens et un vocabulaire de 11 491. La taille moyenne des phrases est de 13,87 caractères.

V.4. Apprentissage du modèle

Dans cette partie, nous présenterons les valeurs des réglages utilisés et les résultats obtenus, tout en proposant une comparaison et une analyse de ces résultats.

V.4.1. Paramètre

Nous discuterons des paramètres du modèle utilisés dans le processus de traduction automatique et expliquerons les valeurs des paramètres utilisées pour garantir les meilleures performances possibles.

V.4.1.1. Paramètres de l'optimiseur

- optimizer:AdamOptimizer
- beta1: 0.8
- beta2: 0.998
- learning_rate: 1.0

V.4.1.2. Paramètres d'entraînement

- batch_size: 64
- train_steps: 10000
- save_checkpoints_steps: 1000
- keep_checkpoint_max: 5
- save_summary_steps: 100

- single_pass: true
- maximum_features_length: 70
- maximum_labels_length: 70

V.4.1.3. Paramètresd'évaluation

- batch_size: 32
- num_threads: 1
- prefetch_buffer_size: 1
- save_eval_predictions: false
- exporters: last
- eval_delay: 3600 “Toutes 1 heure”

V.4.1.4. Paramètres d'inférence

- batch_size: 32

V.4.2. Résultats

Le tableau suivant représente les résultats obtenus :

Dataset	DZDA →MSA		MSA →DZDA	
	BLEU	ChrF	BLEU	ChrF
Baseline	19.1	41.5	17.7	45.4
BT.Corpus	30.0	60.4	28.8	53.5
RR.Corpus	27.7	60.7	26.5	53.8

Tableau 4. Performances de traduction MSA-DZDA en utilisant les scores BLEU et ChrF

V.4.3. Analyse des résultats

Nos résultats expérimentaux montrent des améliorations significatives de la qualité de la traduction lors de l'utilisation de techniques d'augmentation des données pour les tâches de

traduction DZDA vers MSA et MSA vers DZDA. La comparaison des performances de base à celles de la suite améliorée révèle des améliorations significatives dans tous les domaines. Les résultats BLEU et ChrF pour le système NMT formé sur des ensembles distincts sont rapportés dans le tableau précédent.

Pour la tâche de traduction DZDA→-MSA, les scores bruts BLEU sont de 19,1 et les scores ChrF sont 41,5 Lorsque la technologie d'augmentation BT Corpus est introduite, il y a une amélioration significative, augmentant le score BLEU à 30,0 et le ChrF à 60,4 Pendant ce temps, le groupe RR présente une augmentation légèrement moindre mais néanmoins significative, atteignant un score BLEU de 27,1 et 60,7 un score ChrF.

Dans le contexte de la tâche de traduction MSA → DZDA, nous observons des modèles similaires. Les scores BLEU initiaux sont de 17,7, tandis que les scores ChrF sont 45,4 L'utilisation du la technologie d'augmentation BT Corpus entraîne une amélioration significative, atteignant un score BLEU de 28,8 et un score ChrF de 53,5 , Bien que légèrement réduite, l'augmentation du groupe RR a tout de même montré des progrès significatifs, avec un score BLEU de 26,5 et des scores ChrF de 53,8.

Tout bien considéré, ces résultats soulignent à quel point l'augmentation des données est cruciale pour réduire les difficultés causées par les langues à faibles ressources et améliorer l'efficacité des systèmes NMT lors de la traduction de textes combinant l'arabe standard moderne et l'arabe dialectal. Des études ultérieures pourraient examiner d'autres méthodes d'augmentation et leur impact sur la qualité des traductions, ainsi que leur efficacité avec diverses paires de langues et domaines.

L'approche « Right-Rotation Augmentation » (RRA) présente les avantages et les inconvénients suivants :

Avantages:

- Diversité des données accrue : RRA alterne les phrases préexistantes pour en produire de nouvelles, ajoutant ainsi une plus grande variété d'ordres de mots et de structures de phrases au corpus. La diversité accrue a le potentiel d'améliorer la flexibilité et les compétences de généralisation du modèle de traduction automatique neuronale (NMT).
- Préservation de la structure des phrases : RRA préserve la structure de base des phrases tout en modifiant l'ordre des mots, garantissant que les phrases améliorées préservent les

composants essentiels de la composition de la langue arabe. Cela maintient intactes l'intégrité grammaticale et la cohérence du corpus accru.

- Facilité de mise en œuvre : RRA est une technique d'augmentation des données simple, efficace et facile à utiliser. Il est accessible et rapide pour améliorer les corpus MT à petite échelle, car il ne nécessite pas d'étapes de prétraitement complexes ni de ressources externes supplémentaires.

Inconvénients :

- Variation de sens limitée : la RRA ajoute de la variété aux structures linguistiques, mais les phrases enrichies peuvent ne pas subir de changements de sens appréciables. La signification et le contenu des phrases peuvent rester pratiquement inchangés, ce qui pourrait limiter la capacité du modèle NMT à reconnaître et à gérer de subtiles variations sémantiques.
- Possibilité de phrases anormales : la RRA peut produire des phrases qui semblent moins répandues dans l'usage courant du langage ou anormales, en fonction du degré de rotation et de la complexité de la phrase. Cela pourrait avoir un impact sur la précision et la fluidité des traductions du modèle NMT.
- Disparité des données : RRA ne traite pas spécifiquement les problèmes liés à la rareté des données. Même après avoir produit de nouvelles phrases, la taille du corpus de départ limite toujours la taille du corpus amélioré. Si le corpus d'origine est petit, il se peut qu'il n'y ait pas d'augmentation significative des données d'entraînement provenant du corpus complété.

Il est impératif de reconnaître que l'efficacité de la méthodologie RRA peut différer selon les complexités particulières de mise en œuvre, le calibre du corpus original et les attributs linguistiques de la langue améliorée. Il est nécessaire de mener des évaluations empiriques et des analyses comparatives avec d'autres approches d'augmentation des données afin de déterminer l'efficacité de la RRA.

V.5. Conclusion

Ce dernier chapitre décrit, d'une manière détaillée, les outils utilisés pour réaliser le travail expérimental et fournit les résultats obtenus avec une analyse et une comparaison des résultats de prédiction des deux méthodes utilisées.

CONCLUSION GÉNÉRALE

Nous rappelons que l'objectif de notre travail est de comparer les performances d'un modèle de traduction automatique lors de l'utilisation de différentes techniques d'augmentation de données, en se concentrant sur la langue arabe et ses défis spécifiques en matière de traduction automatique. La langue arabe présente des défis uniques en raison de sa morphologie complexe, de ses différents dialectes et de la variabilité de son utilisation dans différents contextes.

Dans la première partie, nous avons présenté la conception de notre méthode, en détaillant les étapes de prétraitement des données, la description du modèle proposé basé sur OpenNMT, et la préparation des données à l'aide de BPE et la création d'un dictionnaire de mots.

Dans la deuxième partie, nous avons abordé les tests et les résultats obtenus. Nous avons commencé par décrire le matériel et les logiciels utilisés, puis nous avons présenté les statistiques des ensembles de données MSA-DZDA utilisés, ainsi que les paramètres du modèle et les résultats des tests, analysant les performances de traduction en utilisant les scores BLEU et ChrF.

Les résultats expérimentaux ont montré des améliorations significatives de la qualité de la traduction lors de l'utilisation de techniques d'augmentation des données pour les tâches de traduction DZDA vers MSA et MSA vers DZDA. La comparaison des performances de base à celles des ensembles améliorés a révélé des améliorations significatives dans tous les domaines, soulignant l'importance de l'augmentation des données pour améliorer les performances des systèmes NMT.

En particulier, la rétro-traduction (BT Corpus) a conduit à des améliorations notables dans les scores BLEU et ChrF, montrant son efficacité pour enrichir le corpus d'entraînement. De même, la rotation à droite (RR Corpus) a également démontré son potentiel à améliorer la qualité des traductions, bien que dans une moindre mesure que la rétro-traduction.

Ces résultats confirment l'importance de l'augmentation des données pour surmonter les défis posés par les langues à faibles ressources, en particulier pour les paires de langues comme l'arabe standard moderne et l'arabe dialectal. Ils soulignent également la nécessité de poursuivre la recherche sur les méthodes d'augmentation des données pour améliorer encore les performances des systèmes de traduction automatique, en particulier dans des contextes linguistiques complexes comme l'arabe.

En conclusion, notre étude a non seulement contribué à améliorer les performances de traduction entre l'arabe standard moderne et l'arabe dialectal, mais elle a également ouvert de nouvelles perspectives pour la recherche future dans le domaine de la traduction automatique, en particulier pour les langues qui ont été historiquement sous-représentées dans ce domaine , en mettant en évidence l'importance de l'augmentation des données et en proposant des pistes pour de nouvelles méthodes et approches.

REFERENCES BIBLIOGRAPHIQUES

[1] Hadj Ameur, M. S., Meziane, F., &Guessoum, A. (2020). Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Review, 38*, 100305.

[2] Mallek, F., Belainine, B., & Sadat, F. (2017). Arabic social media analysis and translation. 'Procedia Computer Science, 117', 298-303. https://doi.org/10.1016/j.procs.2017.10.121

[3] Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A survey on evaluation metrics for machine translation. 'Mathematics, 11'(4), 1006. https://doi.org/10.3390/math11041006

[4] Tomedes. (2024, May 10). Évaluation de la traduction automatique: Le guide ultime. MachineTranslation.com. https://www.machinetranslation.com/fr/blog/machine-translation-evaluation-ultimate-guide

[5] Stahlberg, F. (2020). Neural machine translation: A review. 'Journal of Artificial Intelligence Research, 69', 343-418. https://doi.org/10.1613/jair.1.12093

[6] Pham, M.-Q. (2021). 'Multi-domain neural machine translation' (Thèse de doctorat, Université Paris-Saclay). https://theses.hal.science/tel-03546910

[7] Turing. (2024). Mathematical formulation of feed-forward neural network. MachineTranslation.com. https://www.turing.com/kb/mathematical-formulation-of-feed-forward-neural-network

[8] Pham, M.-Q. (2021). *Multi-domain neural machine translation* (Thèse de doctorat, Université Paris-Saclay). Université Paris-Saclay. https://theses.hal.science/tel-03546910

[9] Zhang, B. (2022). Towards Efficient Universal Neural Machine Translation (Doctoral dissertation). University of Edinburgh.

[10] Dabre, R., Chenhuichu, &Kunchukuttan, A. (Date non spécifiée). A Survey of Multilingual Neural Machine Translation. [Titre de la revue non spécifié].

[11] BABAALI, B., & SALEM, M. (2022). Traduction automatique arabe : Un aperçu panoramique. Consulté le 22 mai 2024, sur SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4312742

[12] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., &Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[13] David Y., Timothy B., & Anna K. (2013). Analyse morphologique et désambiguïsation pour l'arabe dialectal. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1115-1126). https://aclanthology.org/volumes/D13-1/

[14] Sutskever, I., Vinyals, O., &Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[15] Farajian, M. A., Turchi, M., Negri, M., & Federico, M. (2017). Traduction automatique neuronale multi-domaines via l'adaptation non supervisée. *Proceedings of the Second Conference on Machine Translation* (pp. 127-137). https://doi.org/10.18653/v1/W17-4713

[16] Wray, S., Camargo de Souza, J., Specia, L., & Way, A. (2020). Evaluating Metrics for Machine Translation: BLEU, chrF, COMET, and Beyond. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7618-7629). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.612

[17] Merhav, Y., & Ash, K. (2018). A Simple and Effective Approach to Automatic Post-Editing for Statistical and Neural Machine Translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 570-576). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1091

[18] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 86-96). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1009

[19] Guzmán, F., Sajjad, H., Sennrich, R., & Vogel, S. (2016). Incorporating Linguistic Knowledge into Neural Machine Translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 44-54). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1005

[20] Bahdanau, D., Cho, K., &Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[21] Luong, M.-T., Pham, H., &Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1412-1421). https://arxiv.org/abs/1508.04025

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (NeurIPS 2017) (pp. 5998-6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[23] Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Proc. 34th Int. Conf. Machine Learning (Vol. 70, pp. 1243-1252). Available online: https://proceedings.mlr.press/v70/gehring17a.html

[24] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. Retrieved from https://arxiv.org/abs/1609.08144

[25] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, 4*, 339-351.

Retrieved from https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00065/43400/Google-s-Multilingual-Neural-Machine-Translation

[26] Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhang, H. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint* https://arxiv.org/abs/1803.05567.

[27] Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. *arXiv preprint* arXiv:1808.09381. Retrieved from https://arxiv.org/abs/1808.09381.

[28] Lample, G., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. *arXiv preprint* arXiv:1711.00043. https://arxiv.org/abs/1711.00043

[29] Artetxe, M., Labaka, G., & Agirre, E. (2018). Unsupervised neural machine translation. *arXiv preprint* arXiv:1710.11041. https://arxiv.org/abs/1710.11041

[30] He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T. Y., & Ma, W. Y. (2016). Dual learning for machine translation. In *Advances in neural information processing systems* (pp. 820-828). https://proceedings.neurips.cc/paper_files/paper/2016/hash/5b69b9cb83065d403869739ae7f0995e-Abstract.html

[31] Firat, O., Cho, K., Jean, S., Yarats, D., Deng, Y., Haddadi, A., ... & Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint* arXiv:1601.01073. https://arxiv.org/abs/1601.01073

[32] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 311-318). https://aclanthology.org/P02-1040

[33] Collison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 249-256). https://aclanthology.org/E06-1032

[34] Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392-395). https://aclanthology.org/W15-3049

[35] Babaali, B., Salem, M., & ElHarbe, N. R. (2024). Chasing the Recipe for Effective Low-Resource Neural Machine Translation. Unpublished manuscript.

[36] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6256-6268, 2020.

[37] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, pp. 86-96, 2016.

[38] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "AutoAugment: Learning Augmentation Policies from Data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113-123, 2018.

[39] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 702-703, 2020.

[40] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *International Conference on Learning Representations (ICLR)*, 2017.

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 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), 2019, pp. 4171–4186.

[42] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2014), 2014, pp. 3104–3112.