République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي

Université Mohamed El Bachir El Ibrahimi
Bordj Bou Arréridj
جامعة محمد البشير الابراهيمي **برج بوعريريج**
Faculté des Mathématiques et d'Informatique
Département d'informatique

Mémoire de Projet de Fin d'Études

Pour obtenir un diplôme de master en informatique

Option :

Technologies de l'Information et de la Communication (TIC)

Traduction des documents arabes par les transformes

Préparé par :
M. METAAI ilhem
M. YAHIAOUI hadjer

Encadré par :
M. BELAZZOUG mougoub

Devant le jury composé de :
M. ATTIA Abdelwahab
M. BAGHOURA Mohamed Amine

Promotion 2024/2025

Dédicace

Avant toute chose, Je remercions Dieu Tout-Puissant, pour la force, la patience et la lumière qu'Il a semée dans notre chemin.

Nous dédions ce travail à nos chers parents Metaai Ahmed,Beddar Nadia ces âmes nobles qui ont toujours cru en nous et qui ont offert sans compter, amour, prières et sacrifices silencieux. Ce travail est le fruit de leur foi en nous.

À mes familles, À mon encadrant, Monsieur Mouhoub Blaazoug, pour sa patience, son écoute et ses conseils éclairés.

À mon cher ami Ilyas Djabarni, surnommé "El Khattat", pour sa gentillesse, et sa présence fidèle jusqu'au bout.

À tous mon camarades de Master 2, avec qui nous avons partagé le savoir, le stress.

À tous ceux qui nous ont tendu la main, soutenu d'un mot, d'un sourire, d'une prière … Merci infiniment.

- ***ilhem***

Avant toute chose, Je remercions Dieu Tout-Puissant, pour la force, la patience et la lumière qu'Il a semée dans notre chemin.

Nous dédions ce travail à nos chers parents Yahiaoui ali, Slimani ghania ces âmes nobles qui ont toujours cru en nous et qui ont offert sans compter, amour, prières et sacrifices silencieux. Ce travail est le fruit de leur foi en nous.

À mes familles, À mon encadrant, Monsieur Mouhoub Blaazoug, pour sa patience, son écoute et ses conseils éclairés.

À mon cher ami Ilyas Djabarni, surnommé "El Khattat", pour sa gentillesse, et sa présence fidèle jusqu'au bout.

À tous mon camarades de Master 2, avec qui nous avons partagé le savoir, le stress.

À tous ceux qui nous ont tendu la main, soutenu d'un mot, d'un sourire, d'une prière … Merci infiniment.

- ***hadjer***

Remerciements

Tout d'abord on remercie « DIEU » pour nous avoir donné la force, la capacité, la volonté et le courage afin de mener à bien et à terme ce travail. Nous adressons également nos sincères remerciements à : Notre promoteur Mr BELAZOUG Mouhoub pour ses précieux conseils, son dévouement, pour son suivi et pour nous avoir aussi bien encadrées tout au long de la réalisation de ce projet. Nous tenons aussi à lui adresser notre gratitude pour tout le temps qu'il nous a consacrée, sa disponibilité ainsi que ses encouragements. Aux membres du jury pour avoir accepté de bien vouloir lire notre travail, l'examiner, l'évaluer et nous corriger.

Abstract

In recent years, the field of machine translation has seen a remarkable development thanks to the rapid advancement of artificial intelligence (AI) technologies, especially with the emergence of deep learning-based Transformers models. These models have contributed to improving translation quality, especially when it comes to languages that are not standardized or lack sufficient linguistic resources, such as Algerian Darija.

This study aims to build and evaluate a deep learning-based machine translation model for translating texts from Algerian Darija to English. To achieve this goal, a dataset containing sentences written in Algerian Darija and their corresponding English translations was collected and processed, and a model based on the Transformer architecture was trained using this data.

Keywords : Machine Translation, Artificial Intelligence, Deep Learning, Transformer Mo- dels, Human Translation, Traditional Translation, Adapters, Standard Arabic, Algerian Darija, Multilingual Dataset, Linguistic Features, Neural Translation

---

Résumé

Ces dernières années, le domaine de la traduction automatique a connu un développement remarquable grâce à l'avancée rapide des technologies d'intelligence artificielle (IA), notamment avec l'émergence de modèles Transformers basés sur l'apprentissage profond. Ces modèles ont contribué à l'amélioration de la qualité des traductions, en particulier dans le cas de langues non standardisées ou dépourvues de ressources linguistiques suffisantes, comme la darija algérienne.

Cette étude vise à construire et à évaluer un modèle de traduction automatique basé sur l'apprentissage profond pour la traduction de textes de la darija algérienne vers l'anglais. Pour atteindre cet objectif, un ensemble de données contenant des phrases écrites en darija algérienne et leurs traductions anglaises correspondantes a été collecté et traité, et un modèle basé sur l'architecture Transformer a été entraîné à l'aide de ces données.

Mots clés : Traduction automatique, Intelligence artificielle, Apprentissage profond, Modèles Transformers, Traduction humaine, Traduction traditionnelle, Adaptateurs, Arabe standard, Darija algérienne, Multilinguisme, Caractéristiques linguistiques, Traduction neuronale.

ملخص

في السنوات الأخيرة، شهد مجال الترجمة الآلية تطورًا ملحوظًا بفضل التقدّم السريع في تقنيات الذكاء الاصطناعي، ولا سيما مع ظهور نماذج المحولات (Transformers) المعتمدة على التعلم العميق. لقد ساهمت هذه النماذج في تحسين جودة الترجمة، خاصة عندما يتعلق الأمر باللغات غير الموحدة أو التي تفتقر إلى موارد لغوية كافية، مثل الدارجة الجزائرية.

تهدف هذه الدراسة إلى بناء وتقييم نموذج ترجمة آلية قائم على تقنيات التعلم العميق، مخصص لترجمة النصوص من الدارجة الجزائرية إلى اللغة الإنجليزية. ولتحقيق هذا الهدف، تم جمع ومعالجة مجموعة بيانات تحتوي على جمل مكتوبة بالدارجة الجزائرية مع ترجماتها المقابلة باللغة الإنجليزية، ثم تدريب نموذج قائم على بنية Transformer باستخدام هذه البيانات.

كلمات مفتاحية : الترجمة الآلية، الذكاء الاصطناعي، التعلم العميق، نماذج المحولات، الترجمة البشرية، الترجمة التقليدية، المحولات الفرعية ،(Adapters) اللغة العربية الفصحى، الدارجة الجزائرية، تعدد اللغات، الخصائص اللغوية، الترجمة العصبية.

Liste des abréviations

TA : Traduction Automatique

IA : Intelligence Artificielle

NMT : Neural Machine Translation (Traduction automatique neuronale)

SMT : Statistical Machine Translation (Traduction automatique statistique)

RBMT : Rule-Based Machine Translation (Traduction automatique à base de règles)

NLP : Natural Language Processing (Traitement du langage naturel)

PNL : Traitement du Langage Naturel

BERT : Bidirectional Encoder Representations from Transformers

mBART : Multilingual BART

GNMT : Google Neural Machine Translation

AraBERT : Arabic BERT

AraGPT2 : Arabic GPT-2

NLLB : No Language Left Behind

FFNN : Feed Forward Neural Network

ELU : Exponential Linear Unit

POS : Part Of Speech

BLEU : Bilingual Evaluation Understudy

Seq2Seq : Sequence-to-Sequence

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Le besoin de comprendre et de traduire des textes provenant d'autres langues n'a jamais été aussi pressant en raison de l'évolution rapide d'Internet et de la mondialisation de l'information. La traduction automatique est devenue un outil essentiel pour surmonter les obstacles linguistiques, que ce soit dans les domaines de l'éducation, du commerce international, de la diplomatie ou même des réseaux sociaux.

Cependant, malgré les progrès réalisés au cours des décennies, les systèmes de traduction traditionnels basés sur des règles linguistiques ou statistiques ont souvent montré leurs limites en termes de fluidité, de précision sémantique et de prise en compte du contexte global. Des modèles tels que BERT, T5, MarianMT ou mBART ont été développés pour traiter des textes multilingues et garantir une traduction plus précise et contextuelle. Ils s'appuient sur des systèmes sophistiqués capables d'apprendre de riches représentations linguistiques à partir de vastes corpus textuels.

Dans ce travail, nous nous concentrons sur le développement et le test d'un modèle de traduction automatique basé sur l'architecture Transformer, spécialement conçu pour traduire des textes arabes en anglais. Cette décision peut s'expliquer par la complexité linguistique de la darija, son manque de normalisation et son fort mélange avec d'autres langues comme le français. L'objectif est de déterminer dans quelle mesure un modèle de type Transformer peut répondre aux difficultés présentées par cette variété linguistique sous-représentée dans les ressources disponibles.

Contexte : Les progrès rapides de la technologie de l'intelligence artificielle ont fait de la traduction automatique un élément central des applications linguistiques modernes. Parmi les approches les plus performantes figurent les modèles Transformer, tels que BERT et GPT, qui ont montré une forte capacité à comprendre et générer du texte en plusieurs langues.

Dans ce contexte, notre travail vise à développer un modèle de traduction basé sur l'architecture Transformer, capable de traduire efficacement des textes de la darija algérienne vers l'anglais, tout en prenant en compte les particularités linguistiques et culturelles de cette langue.

Problématique :

Par ailleurs, face à la quantité croissante de textes produits quotidiennement, la traduction manuelle devient impraticable. D'où l'importance de développer des modèles automatisés capables de traiter efficacement cette variété linguistique spécifique. C'est dans cette optique que notre projet propose un modèle de traduction automatique basé sur l'architecture Transformer, ciblant la traduction de la darija algérienne vers l'anglais.

C'est pourquoi cette étude pose la question suivante :

Quelle est l'efficacité du modèle BERT personnalisé par rapport à la traduction manuelle et à la traduction générée par la machine lors de la traduction de phrases multilingues contenant des termes arabes, dialectaux et anglais ?

Contribution :

Ce travail vise à :

- Développer un modèle de traduction automatique basé sur l'architecture BERT, capable de traduire des phrases écrites en darija algérienne vers l'anglais.
- Construire un corpus annoté contenant des exemples représentatifs de la darija algérienne pour entraîner et évaluer le modèle.
- Étudier les défis linguistiques liés à la traduction de la darija, notamment la variabilité lexicale, l'absence de standardisation et l'influence d'autres langues.

Plan de travail :

Ce mémoire est organisé comme suit :

- **Chapitre 1** : Introduction générale à la traduction, ses types, ses approches, et ses enjeux dans le contexte des langues comme l'arabe et la darija algérienne.
- **Chapitre 2** : Présentation des travaux connexes ainsi qu'un aperçu des fondements du deep learning appliqué à la traduction automatique.
- **Chapitre 3** : Méthodologie de travail, incluant la préparation des données, la description du modèle utilisé, et les outils mis en œuvre pour la traduction.
- **Chapitre 4** : Présentation et analyse des résultats obtenus, discussion des performances du modèle, et conclusion avec pistes d'amélioration futures.

partie I

Notion d'étude

Chapitre 1

La traduction automatique

1.1 Introduction

Alors que le monde est de plus en plus connecté, les consommateurs internationaux recherchent de plus en plus des produits et des services de haute qualité, sensibles à la culture et adaptés à leurs besoins spécifiques. Dans le même temps, ils exigent des expériences transparentes, aussi conviviales et accessibles que possible. Dans le contexte de l'économie de l'information, l'adaptation d'un produit aux marchés locaux peut signifier qu'il faut fournir des dizaines de types de contenu dans plusieurs langues et à un large éventail de publics. Les technologies de traduction permettent aux entreprises non seulement de relever ces défis, mais aussi d'optimiser les dépenses de traduction en augmentant la rapidité et la qualité tout en réduisant les coûts. Comme c'est le cas pour la plupart des travaux assistés par la technologie, les outils technologiques de traduction peuvent accroître la productivité, l'efficacité et l'efficience globale de la gestion de contenus multilingues destinés à différents marchés ciblent. Avant l'apparition des technologies de traduction, la traduction était effectuée manuellement, les traducteurs consultant des dictionnaires papier et faisant preuve de discernement. L'impact négatif sur les entreprises était considérable, en raison des délais de mise sur lemarché, de la perte générale de cohérence du contenu, des coûts élevés d'opérations inefficaces et de la baisse de la qualité des résultats due à la nécessité de vérifier manuellement les erreurs.

1.2 Définition

La traduction est plus qu'une simple conversion de mots d'une langue à une autre, c'est un processus complexe qui nécessite une connaissance approfondie des langues concernées, ainsi qu'une compréhension culturelle et contextuelle. Le traducteur doit tenir compte de la structure linguistique, grammaticale et culturelle de la langue cible, ce qui fait de la traduction un art plus qu'un processus technique. Malgré les avancées significatives dans le domaine de la traduction automatique, la traduction humaine est toujours considérée comme essentielle dans de nombreux contextes où la précision et la compréhension du contexte culturel précis sont nécessaires (Team, 2023).

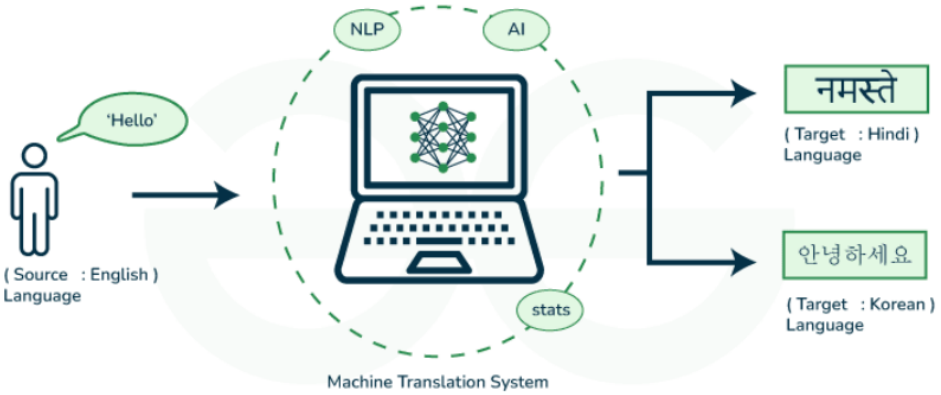

FIG. 1.1 : Translation machine system (SONG et al., 2024)

1.3 Les types de traduction

Dans un monde qui ne cesse de produire du contenu, comment peut-on traduire efficace- ment et dans un délai adapté aux besoins ? De nos jours, il est possible de choisir entre 3 types de traducteurs selon le cas et les besoins :

1.3.1 La traduction humaine

La traduction réalisée par des professionnels reste le principal atout pour les entreprises qui souhaitent obtenir un texte cible de qualité. Dans ce contexte, les compétences du professionnel, sa formation ainsi que son degré d'implication et de connaissance de la culture du pays du texte source sont indispensables. Comme partout, l'un des principaux handicaps auxquels sont confrontés les traducteurs est le temps disponible, l'habituelle urgence des délais de livraison. La technologie contribue à atténuer ce problème à travers des systèmes qui permettent à plusieurs traducteurs de travailler simultanément sur un même projet, avec des textes numériques qui se traduisent sur une même plateforme. Ces outils accélèrent le processus de traduction de manière significative et efficace (SANSCRIT, 2023).

Défis de la traduction : analyse et solutions

Lorsqu'un traducteur humain reçoit un texte, il doit interpréter, à l'aide d'une analyse préa- lable, toutes les significations éventuelles de chacune des phrases qui composent le texte de manière minutieuse et exhaustive. Pour réaliser ce travail, le professionnel doit être formé dans la langue source et la langue cible afin de maîtriser la sémantique et la grammaire du texte et de pouvoir l'adapter dans un cadre culturel adéquat. En règle générale, lorsqu'un traducteur est confronté à un texte, il doit être capable de résoudre certains problèmes récurrents lors de la traduction :

- Grammaticaux : la grammaire des différentes langues dans lesquelles on travaille est le moyen principal de construire correctement chacune des phrases qui composent

un texte

- Emantiques : le sens ou l'interprétation des signes linguistiques comme les symboles, les mots ou les expressions (Sanscrit, 2023).
- Culturels : les expressions et le vocabulaire propres au pays de la langue source et de la langue cible ;
- Syntaxiques : les relations d'accords et de hiérarchies entre les mots lorsqu'ils sont groupés pour former des phrases ;
- Intentionnels : l'intention du texte qui est traduit afin de l'interpréter. Par exemple, lorsque l'on est face à une phrase ironique ou humoristique .
- Idiomatiques : toutes les langues ont des mots ou des expressions qui ne peuvent généralement pas être traduits littéralement. Le traducteur doit donc savoir quelle est l'intention de l'auteur et garder le sens de l'expression dans la langue cible (Sanscrit, 2023).

1.3.2 Assistée par ordinateur – CAT

Les logiciels de traduction assistée par ordinateur, communément appelés outils de TAO,offrent une série de fonctionnalités permettant aux traducteurs de convertir le sens d'un texte d'une langue à l'autre de manière cohérente et rapide. Les outils de TAO modernes adoptent une approche hybride qui permet aux traducteurs et aux gestionnaires de projets de traduction de travailler en ligne ou hors ligne, à l'aide d'outils de bureau ou de navigateurs (Sanscrit, 2023).

Les avantages de l'utilisation d'un logiciel de traduction assistée par ordinateur

- Favoriser la cohérence : Grâce à l'utilisation de mémoires de traduction et de bases terminologiques, vous pouvez réutiliser les travaux de traduction antérieurs pour obtenir des résultats cohérents sur chaque projet.
- Accélérer les délais d'exécution : Plus la combinaison des mémoires de traduction, de la terminologie et de la traduction automatique vous permettent d'exploiter le contenu, plus votre processus de traduction sera rapide.
- Améliorer la qualité : Grâce à un ensemble de fonctions d'assurance qualité automatisées conçues pour compenser les erreurs humaines (telles que les correcteurs orthographiques, les algorithmes grammaticaux et les vérificateurs de chiffres), vous pouvez vous assurer que vos traductions sont précises et cohérentes.
- Augmenter les marges : En passant moins de temps à taper, éditer et formater manuellement le contenu, vous pourrez vous concentrer sur ce qui vous rapporte de l'argent : la traduction. De plus, vous pouvez faire profiter vos clients de ces avantages, ce qui en fait une situation gagnant-gagnant pour tout le monde.

1.3.3 La traduction automatique

La traduction automatique est réalisée par un ordinateur sans qu'un traducteur humain intervienne. Aujourd'hui, à l'ère du numérique, une vaste palette d'outils de traduction automatique est à disposition de tous pour traduire, en quelques secondes, des contenus dans des centaines de langues différentes. Google Translate est sans doute l'outil de traduction automatique grand public le plus connu. Il existe différentes méthodes de traduction automatique (la traduction neuronale, la traduction automatique statistique, et la traduction RBMT - à base de règles) (Bai & Yu, 2014).

La traduction automatique à base de règles (Rule-Based Machine Translation RBMT)

cyclopedia of Translation Technology. Routledge, 2014. La traduction automatique à base de règles repose sur des connaissances linguistiques for- melles, des dictionnaires bilingues et des règles grammaticales. Le système procède à une ana- lyse syntaxique et morphologique de la phrase source, puis applique des règles linguistiques pour produire la traduction. Ce type de traduction offre une certaine cohérence grammaticale, mais reste limité dans la gestion du contexte et nécessite un travail manuel important pour définir les règles (Koehn, 2009).

La traduction automatique statistique (Statistical Machine Translation - SMT)

La traduction automatique statistique repose sur l'exploitation de grands corpus bilingues. Elle apprend à traduire en calculant les probabilités de correspondance entre des segments de phrases dans différentes langues. Bien qu'elle ait représenté une avancée majeure, elle présente des lacunes dans la prise en compte du contexte global, ce qui peut engendrer des traductions imprécises ou peu naturelles.(Koehn, 2009)

La traduction automatique neuronale (Neural Machine Translation - NMT)

La traduction automatique neuronale est une méthode récente fondée sur les réseaux de neurones profonds, en particulier l'architecture Transformer. Elle permet de traiter une phrase dans son ensemble et de générer des traductions plus fluides, naturelles et contextualisées. Cette approche est actuellement la plus performante et est utilisée dans des systèmes comme Google Translate (version actuelle), DeepL, MarianMT, ou mBART (Koehn, 2009).

La traduction automatique hybride

Est une méthode de traduction automatique qui se caractérise par l'utilisation de plusieurs approches de traduction automatique au sein d'un seul système de traduction automatique. La motivation pour développer des systèmes de traduction automatique hybrides provient de l'incapacité d'une technique unique à atteindre un niveau de précision satisfaisant (M., 2025).

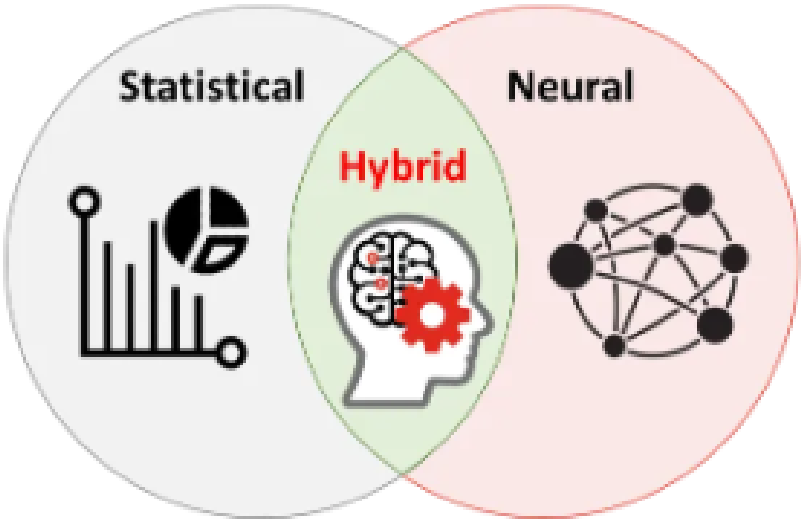

Fig. 1.2 : Translation automatique hybride
(RocketCDN, 2023)

Parallel Hybrid Cette approche de la traduction automatique hybride consiste à faire fonctionner plusieurs systèmes de traduction automatique en parallèle. Le résultat final est généré en combinant les résultats de tous les sous-systèmes. Le plus souvent, ces systèmes utilisent des sous-systèmes de traduction statistique et à base de règles, mais d'autres combinaisons ont été explorées. Par exemple, des chercheurs de l'université Carnegie Mellon ont réussi à combiner des sous-systèmes de traduction basés sur l'exemple, basés sur le transfert, basés sur la connaissance et statistiques en un seul système de traduction automatique.

Integrated Hybrid Cette approche consiste à utiliser des données statistiques pour générer des règles lexicales et syntaxiques. L'entrée est ensuite traitée avec ces règles comme s'il s'agissait d'un traducteur à base de règles. Cette approche tente d'éviter la tâche difficile et fastidieuse de création d'un ensemble de règles linguistiques complètes et fines en extrayant ces règles du corpus d'apprentissage. Cette approche souffre encore de nombreux problèmes.

liés à la traduction automatique statistique normale, à savoir que la précision de la traduction dépend fortement de la similarité entre le texte d'entrée et le texte du corpus d'apprentissage. Par conséquent, cette technique a connu le plus grand succès dans les applications spécifiques à un domaine et présente les mêmes difficultés d'adaptation au domaine que de nombreux systèmes de traduction automatique statistique.

Sequential Hybrid Cette approche implique le traitement en série des entrées à plusieurs reprises. La technique la plus courante utilisée dans les systèmes de traduction automatique multipasses consiste à prétraiter les entrées avec un système de traduction automatique basé sur des règles. Le résultat du préprocesseur basé sur des règles est transmis à un système de traduction automatique statistique, qui produit le résultat final. Cette technique permet de li- miter la quantité d'informations à prendre en compte

par un système statistique, réduisant ainsi considérablement la puissance de traitement requise. Elle évite également que le système basé sur des règles soit un système de traduction complet pour la langue, réduisant ainsi considéra- blement l'effort et le travail humains nécessaires à sa création.

Interactive Hybrid Cette approche diffère des autres approches hybrides par le fait que, dans la plupart des cas, une seule technologie de traduction est utilisée. Un indicateur de confiance est généré pour chaque phrase traduite, ce qui permet de décider s'il faut essayer une technologie de traduction secondaire ou poursuivre la traduction initiale. La TMS est égale- ment utilisée lorsque des schémas d'erreur courants, tels que la répétition de plusieurs mots, apparaissent successivement, comme c'est souvent le cas avec la TNM lorsque le mécanisme d'attention est perturbé.

1.4 fonctionnement de la technologie NLP

La technologie PNL permet à un programme informatique de comprendre le texte et la parole humaine. Étant donné que les ordinateurs ne comprennent que le langage binaire composé de 0 et de 1, nous avions besoin d'un système pour faire d'abord comprendre les mots à un ordinateur. Pour cela, la représentation des mots est utilisée, où les mots sont codés dans le langage informatique. Plusieurs techniques sont utilisées à cet effet, et l'one-hot est l'une de ces techniques. En plus de cela, une suite de techniques PNL est utilisée pour aider un ordinateur à comprendre le langage humain. Ceux-ci inclus : (Bai & Yu, 2014).

Fig. 1.3 : Fonctionnement de la technologie NLP (Koehn, 2009)

- Tige : Un processus dans lequel des mots similaires sont raccourcis jusqu'à leur mot d'origine, comme Finalize, depuis Final en éliminant les alphabets un par un.
- Lemmatisation : Il s'agit d'une technique par laquelle les mots sont érodés pour trouver leur structure de base significative.
- Tokenisation : Avec cette technique, les phrases sont décomposées en blocs plus petits pour en identifier les mots, les symboles et les chiffres.
- Analyse des sentiments : C'est là qu'un ordinateur tente d'identifier le ton et l'émotion derrière la phrase.

- Désambiguïsation du sens des mots : Cette technique est utilisée pour déterminer si le même mot a des significations différentes lorsqu'il est utilisé dans des contextes diffé- rents.

- Marquage d'une partie du discours (POS) : Le marquage POS est utilisé pour annoter chaque mot du texte. Cela inclut l'identification des verbes, des adverbes, des noms, des adjectifs et de toutes les autres parties du discours. En plus de ces techniques, un programme NLP utilise également des algorithmes pour com- prendre le texte et la parole générés par l'homme. Le système basé sur des règles est utilisé pour définir les règles permettant à la linguistique d'analyser les données.

1.4.1 Le rôle du traitement du langage naturel dans le développement des systèmes de traduction automatique modernes

Le traitement du langage naturel (TAL) joue un rôle central dans le développement des sys-tèmes de traduction automatique modernes, car il constitue la base sur laquelle sont construits les modèles linguistiques capables de comprendre et d'analyser les textes. Les systèmes de traduction modernes, en particulier ceux qui reposent sur des modèles d'apprentissage profond telsque Transformer, s'appuient sur des techniques de TAL pour analyser la structure syntaxiqueet sémantique des textes originaux, découvrir le contexte et les significations implicites, puisreproduire les textes traduits avec précision et fluidité. Grâce à des tâches telles que l'analyse syntaxique, la représentation sémantique, la reconnaissance des entités nommées et la tokenisation, les technologies NLP permettent aux systèmes de traduction d'aller au-delà de la traduction littérale et d'évoluer vers une traduction sensible au contexte culturel et linguistique. L'intégration de modèles de préentraînement (telsque BERT, GPT et mBART) a considérablement amélioré la précision de la traduction multilingue, en particulier dans les langues à ressources limitées. Ainsi, le traitement du langage naturel n'est pas seulement un outil d'appui, mais aussil'infrastructure sur laquelle les systèmes de traduction automatique s'appuient pour comprendre les langues humaines et générer des traductions de haute qualité,compréhensibles par l'homme (Jurafsky & Martin, 2025).

1.4.2 Les difficultés de la traduction automatique en arabe et dans les dialectes locaux à l'aide de NLP

L'une des utilisations les plus avancées et les plus importantes du traitement du langage naturel (TALN), en particulier dans un monde multilingue et globalisé, est la traduction automatique. Cependant, l'utilisation de cette technologie en arabe, compte tenu notamment de la variété de ses dialectes régionaux, présente un ensemble de difficultés particulières qui sont très différentes de celles auxquelles sont confrontées les langues occidentales. Ces difficultés découlent de caractéristiques linguistiques complexes, d'un manque de ressources linguistiques et d'un écart important entre l'arabe standard moderne et les dialectes régionaux.

La complexité morphologique et syntaxique de l'arabe

La langue arabe possède une structure morphologique riche et sophistiquée, les mots étant construits à l'aide de modèles spécifiques basés sur des races trilittéraires ou quadrilittéraires. Cette structure diffère de celle de la majorité des langues européennes, qui sont basées sur des séquences de mots directes. L'ordre des phrases arabes peut également changer en fonction de la situation, en commençant par un verbe ou un sujet. Cela présente des défis importants pour les modèles de traduction qui s'appuient sur la séquence grammaticale traditionnelle (Lakhfif & Laskri, 2017).

Plusieurs dialectes régionaux

Il existe plusieurs dialectes différents dans le monde arabe, notamment l'algérien, l'égyptien, le grec, le levantin et le marocain, qui diffèrent en termes de vocabulaire, de grammaire et de structures. Ces dialectes peuvent ne pas avoir d'informations ou d'informations sur les réseaux sociaux, ce qui pourrait rendre difficile leur détection ou leur analyse précise pour le système PNL. Ils peuvent être capables d'utiliser une variété de symboles sur le cadran et ne pas être reconnus dans d'autres situations.

La différence entre l'arabe standard et les dialectes

Il arrive que les textes officiels soient souvent rédigés en arabe contemporain, mais les interactions quotidiennes utilisent des dialectes. Cela représente un défi supplémentaire pour les systèmes de traduction qui doivent être capables de comprendre toutes les formes de la langue. Malheureusement, la majorité des systèmes de traduction actuellement utilisés sont basés uniquement sur la langue standard moderne, ce qui rend leurs performances médiocres lors de la traduction de textes familiers (Lakhfif & Laskri, 2017).

Gestion des données et des ressources linguistiques

Les modèles de traduction automatique modernes, en particulier la traduction automatique neuronale (NMT), s'appuient sur des ensembles de données parallèles extrêmement volumineux pour les entraîner. Cependant, ces ressources sont assez peu nombreuses pour l'arabe et encore plus rares pour les dialectes. Ce manque de données limite la capacité des systèmes à apprendre efficacement et a un effet négatif sur la qualité de la traduction.

Définitions des techniques de modélisation du langage

Même avec la disponibilité des données, un modèle efficace de l'arabe nécessite des outils sophistiqués pour analyser les races, l'inflexion, les signes diacritiques et l'intonation, qui sont tous des éléments essentiels pour comprendre le sens précis des mots. Cela pose un défi important au développement de modèles linguistiques efficaces, car toute erreur dans la première analyse affecte la traduction dans son ensemble (Lakhfif & Laskri, 2017).

Sémantiquement ambigu et multiforme

Un seul mot en arabe peut avoir plusieurs significations selon le contexte, ce qui en fait une langue riche en expressions et en significations. Pour répondre à cette multiplicité, il faut des systèmes intelligents capables de comprendre le contexte complet d'une phrase, ce qui constitue encore un défi pour de nombreux systèmes de traduction utilisés aujourd'hui.

1.4.3 Application des techniques NLP à la darija algérienne et à l'arabe avec le modèle Dziribert

Dziribert est le premier modèle linguistique basé sur *Transformer* qui a été créé spécialement pour le dialecte algérien. Le modèle est un outil puissant pour traiter une variété de textes arabes car il peut traiter des textes écrits à la fois en arabe et en latin (*Arabizi*). Dziribert a été construit à partir de plus d'un million de tweets algériens, fournissant une vaste source de données qui reflète les caractéristiques linguistiques et culturelles du dialecte algérien.

Contrairement à d'autres modèles tels que *MARBERT* et *CAMELBERT*, Dziribert a démontré un avantage significatif dans des tâches telles que l'étiquetage des sentiments et la reconnaissance des émotions, en particulier lorsqu'il s'agit de textes en langue latine. À titre d'exemple, Dziribert a surpassé MARBERT, qui a également atteint une précision de 80 % dans la classification des émotions, mais a été moins précis dans leur identification (Bouziane, Attia et al., 2021).

1.5 Application de modèles de transformateurs

Avec l'émergence des modèles Transformer, le domaine de la traduction automatique a subi un changement de paradigme qui a fondamentalement modifié la manière dont les systèmes d'IA sont conçus pour comprendre et gérer les langues. Parmi les modèles les plus significatifs figurent BERT et GPT, qui sont devenus des outils essentiels dans la création de systèmes de traduction basés sur l'IA en raison de leur capacité exceptionnelle à représenter le sens et à comprendre le contexte.

1.5.1 SYSTRAN "EC SYSTRAN : THE COMMISSION'S MACHINE TRANSLATION SYSTEM" - ACL Anthology

SYSTRAN, qui signifie « SYStem for Translation », est l'un des plus anciens systèmes de traduction automatique à base de règles. Il a été créé à la fin des années 1950 par le linguiste américain Peter Thoma, originaire de Hong Kong. Il se base sur des dictionnaires bilingues et une grammaire créés manuellement, c'est-à-dire qu'il analyse les textes à l'aide d'un certain nombre d'opérations morphologiques et syntaxiques avant de reconstruire le sens dans la langue ciblent à l'aide de règles prédéterminées. De nombreuses organisations gouvernementales et privées ont utilisé SYSTRAN, notamment le gouvernement américain pendant la guerre froide pour traduire des documents russes et la Commission européenne

pour traduire des documents entre les langues officielles de l'UE (Tiedemann, 2017).

1.5.2 Moses : Open Source Statistical Machine Translation System

Un cadre open source appelé Moses est utilisé pour créer des systèmes de traduction statistique automatique (SMT). Il a été créé à l'université de Cambridge et s'appuie sur des techniques d'apprentissage automatique pour extraire des modèles à partir de grandes quantités de textes parallèles dans plusieurs langues. Sur la base de ces données parallèles, le système crée un modèle statistique qu'il utilise ensuite pour traduire les textes dans les différentes langues.(Koehn et al., 2007) Moses fonctionne principalement de la manière décrite ci-dessous :

- L'analyse de textes parallèles consiste à utiliser une paire de textes prétraduits pour acquérir des connaissances.
- Développement d'un modèle statistique : Sur la base de la fréquence des modèles, des modèles probabilistes sont construits pour identifier la traduction la plus appropriée.
- Production de la traduction : Après avoir créé les modèles, Moses les utilise pour créer des traductions pour de nouveaux textes.
- Même si des modèles d'apprentissage profond plus récents comme le NMT ont remplacé Moses dans certaines applications, il est depuis longtemps une référence importante dans le domaine de la traduction statistique.

1.5.3 Le modèle BERT et son rôle dans la traduction

BERT, ou Bidirectional Encoder Representations de Transformers, est un modèle de conférence binaire, ce qui signifie qu'il peut comprendre le contexte d'un mot à la fois sur les côtés gauche et droit, ce qui lui donne une grande capacité à représenter des significations de mots subtiles. Bien que BERT n'ait pas été initialement destiné à la traduction, il a été utilisé dans le contexte de la bande d'analyse (encodeur) dans les systèmes de traduction neuronale automatique (Devlin et al., 2018).

1.6 Le processus de la traduction

La mise en place d'une traduction automatique efficace va au-delà du simple remplacementd'un mot par un autre. Elle implique un enchaînement complexe d'opérations linguistiques et techniques, mobilisant diverses ressources informatiques et modèles d'apprentissage. Chaqueétape contribue à garantir l'exactitude du message, la fluidité du texte et son adéquation au contexte linguistique et culturel de la langue cible. Dans ce qui suit, nous décrivons les principales étapes d'un système de traduction automatique, quel que soit le type de modèle (basé sur des règles statistiques ou neuronales)

1.6.1 Prétraitement du texte source (Text Preprocessing)

Le prétraitement est une phase essentielle qui prépare le texte brut pour une traduction automatique efficace. Il comprend plusieurs sous-étapes :

Nettoyage et filtrage

Cette étape consiste à éliminer les éléments indésirables du texte, tels que :

- Les caractères non imprimables ou les symboles spéciaux.
- Les segments vides ou contenant des erreurs d'encodage (par exemple, UTF-8 invalides).
- Les doublons ou les segments excessivement longs ou courts. Ces opérations permettent de réduire la taille des données et d'améliorer la qualité de l'entraî- nement des modèles de traduction.

Normalisation

La normalisation vise à uniformiser le texte en : ResearchGate+1Towards Data Science+1

- Convertissant toutes les lettres en minuscules (cas normalisation).
- Standardisant la ponctuation (par exemple, remplacer différents types de guillemets par un seul type).
- Éliminant les espaces superflus et les caractères spéciaux. Cette étape est cruciale pour réduire la variabilité du texte et faciliter son traitement ultérieur.

Tokenisation

La tokenisation consiste à segmenter le texte en unités linguistiques appelées "tokens" (mots, ponctuations, etc.). Cette segmentation est essentielle pour l'analyse syntaxique et sémantique du texte. Des outils comme NLTK ou SpaCy sont couramment utilisés pour cette tâche.

Suppression des mots vides (Stopword Removal)

Les mots vides sont des mots fréquents qui portent peu d'information sémantique (parexemple, "le", "et", "de"). Les supprimer permet de se concentrer sur les mots porteurs de sens et de réduire la dimensionnalité des données.

Stemming et lemmatisation

Ces techniques visent à réduire les mots à leur forme de base :

- Stemming : réduction des mots à leur racine en supprimant les suffixes (par exemple, "manger", "mangeons", "mangé" deviennent "mang").
- Lemmatisation : réduction des mots à leur lemme en tenant compte du contexte grammatical (par exemple, "mangé" devient "manger").

La lemmatisation est généralement plus précise que le stemming.

Cas particulier de la langue arabe

Pour les langues morphologiquement riches comme l'arabe, le prétraitement peut inclure :

- La séparation des clitiques (préfixes et suffixes).
- L'analyse morphologique approfondie.
- La désambiguïsation contextuelle.

Ces techniques améliorent significativement la qualité de la traduction automatique pour l'arabe.

1.6.2 Analyse linguistique

L'analyse linguistique vise à comprendre la structure grammaticale et le sens du texte source. Elle comprend plusieurs sous-étapes :

- Analyse morphologique : identification des racines des mots et de leurs affixes.
- Analyse syntaxique : détermination de la structure des phrases (sujets, verbes, objets).
- Analyse sémantique : compréhension du sens des mots et des phrases dans leur contexte.

1.6.3 Transfert linguistique

Le transfert linguistique consiste à convertir la représentation intermédiaire obtenue lors de l'analyse en une structure adaptée à la langue cible. Selon le type de système de traduction :

- RBMT (Rule-Based Machine Translation) : utilise des règles grammaticales et des dictionnaires bilingues pour effectuer le transfert.
- SMT (Statistical Machine Translation) : s'appuie sur des modèles statistiques dérivés de corpus parallèles pour estimer les probabilités de traduction.
- NMT (Neural Machine Translation) : utilise des réseaux neuronaux pour encoder le sens du texte source et le décoder dans la langue cible.

1.6.4 Génération du texte cible

Après le transfert, le système reconstruit la phrase dans la langue cible en respectant ses règles grammaticales et stylistiques. Cette étape peut inclure :

- Réorganisation syntaxique : adaptation de l'ordre des mots pour correspondre à la structure de la langue cible.
- Accord grammatical : ajustement des accords en genre, nombre et temps.
- Choix lexical : sélection des mots appropriés en fonction du contexte.

1.6.5 Post-traitement

Le post-traitement vise à améliorer la qualité finale de la traduction en :

- Corrigeant les erreurs résiduelles.
- Harmonisant la ponctuation.
- Supprimant les répétitions inutiles.
- Assurant la cohérence stylistique.

Dans les environnements professionnels, une post-édition humaine est souvent intégrée pour garantir une qualité optimale.

1.7 Les modèles de traduction pour un texte arabe

1.7.1 MarianMT

MarianMT est un modèle de traduction neuronale automatique construit par l'équipe Helsinki NLP de l'Université d'Helsinki, basé sur l'architecture Transformer. Le modèle a été mis en œuvre à l'aide du cadre MarianNMT, un cadre C++ qui permet un entraînement efficace des modèles de traduction neuronale. Les modèles MarianMT sont basés sur les ensembles de données OPUS, qui sont des ensembles de données parallèles à source ouverte couvrant plusieurs langues (Junczys-Dowmunt, 2018).

- Prise en charge multilingue : Prise en charge de plus de 1000 paires de langues, permettant la traduction entre un large éventail de langues.
- Grande efficacité : Grâce à l'architecture Transformer et à l'utilisation du cadre MarianNMT, le modèle offre des performances rapides et efficaces.
- Intégration avec Hugging Face : Le modèle peut être facilement utilisé grâce à la bibliothèque Transformers de Python, ce qui facilite son intégration dans différentes applications.
- Source ouverte : Le modèle est disponible sous la licence MIT, ce qui permet aux développeurs et aux chercheurs de l'utiliser et de le modifier librement.

1.7.2 Google Translate (NMT Proprietary Model)

GNMT (Google Neural Machine Translation), le système de traduction neuronale automatique de Google, est un modèle sophistiqué construit sur des techniques d'apprentissage profond pour améliorer la qualité des traductions dans Google Translate. Ce système a été présenté dans un texte intitulé : « Le système de traduction neuronale de Google : combler le fosséentre la traduction humaine et la traduction automatique »(Wu et al., 2016).

- Une architecture profonde : Le modèle consiste en un réseau profond LSTM à 8 sièges pour le codeur et le décodeur, avec des mécanismes d'attention et de communication supplémentaires.
- Meilleure gestion des mots rares : Pour mieux gérer les mots rares et optimiser la traduction, le système utilise la fragmentation des mots en unités plus petites appelées « wordpieces ».
- Accélération du processus de traduction : Lors de l'inférence, des calculs de faible précision sont utilisés pour accélérer le processus de traduction sans sacrifier la qualité.
- Amélioration de la qualité de la traduction : Des évaluations humaines ont démontré que, par rapport aux systèmes de traduction basés sur des phrases, le GNMT réduit les erreurs de traduction jusqu'à 60

1.7.3 mBART

Le modèle séquence-séquence de mBART est basé sur un codeur-décodeur qui utilise l'arhitecture Transformer. Avec l'utilisation de la cible BART, il est pré-entraîné sur des ensembles de données monolingues à grande échelle dans plusieurs langues, ce qui lui permet de comprendre et de produire des textes dans plusieurs langues (Liu et al., 2020).

- Prise en charge du multilinguisme : Le modèle prend en charge la traduction multilingue, ce qui le rend adapté aux tâches multilingues.
- Amélioration des performances dans les langues à ressources limitées : Le modèle a démontré une amélioration significative des performances de traduction pour les langues ne disposant pas de ressources de formation adéquates.
- Adaptabilité à diverses tâches : mBART peut être adapté à de nombreuses tâches, telles que la traduction automatique, les CV et d'autres tâches impliquant le traitement du langage naturel.

1.7.4 AraBERT

Le modèle AraBERT + Seq2Seq Decoder est une architecture hybride qui utilise AraBERT comme codeur et un modèle génératif comme AraGPT2 comme décodeur dans le

cadre Encoder-Decoder pour la traduction automatique et les tâches de génération de texte arabe (Antoun et al., 2020).

- AraBERT : Basé sur l'architecture BERT, AraBERT est un modèle linguistique de pré-entraînement pour l'arabe créé par le laboratoire d'intelligence artificielle de l'Université des États-Unis à Beyrouth (AUB-MIND).
- AraGPT2 : Basé sur l'architecture GPT-2, AraGPT2 est un modèle de génération de langue arabe utilisé comme décodeur dans les architectures codeurdécodeur.
- Architecture Seq2Seq efficace : Pour des tâches telles que la traduction automatique etla génération de textes arabes, la combinaison d'AraBERT comme codeur et d'AraGPT2 comme décodeur crée une architecture Seq2Seq efficace.

1.7.5 No Language Left Behind (NLLB) by Meta AI

NLLB est un modèle de traduction automatique multilingue basé sur l'architecture Transformer qui utilise des méthodes de pointe, notamment le mélange d'experts et l'apprentissage auto-supervisé. Le modèle vise à améliorer la qualité de la traduction pour les langues qui nedisposent pas de ressources de formation suffisantes, en réduisant l'écart entre les langues à ressources élevées et les langues à ressources faibles (Firat et al., 2022).

- Prise en charge étendue des langues : Le modèle prend en charge la traduction entre plus de 200 langues, dont plusieurs langues africaines et asiatiques, avec peu de ressources.
- Amélioration de la qualité de la traduction : Par rapport aux modèles précédents, le modèle a démontré une augmentation de 44atteignant 70
- Architecture avancée : Le modèle utilise l'architecture Sparse Mixture-of-Experts, qui permet d'adapter certains de ses composants à des langues particulières sans pour autant tomber dans l'excès de personnalisation.
- Source ouverte : Afin d'encourager la coopération et l'avancement des approches de traduction, le modèle, les ensembles de données et le code d'encodage sont mis à la disposition de la communauté des chercheurs.

1.8 Conclusion

En conclusion, la traduction est l'un des domaines fondamentaux de la linguistique et de l'informatique, car elle permet de combler un fossé important entre de nombreuses langues et civilisations. Les techniques de traduction ont évolué au fil du temps, passant de la traduction manuelle traditionnelle à l'utilisation de modèles informatiques tels que la traduction automatique basée sur des règles, puis à des modèles plus récents basés sur l'apprentissage profond, comme Transformers, qui ont révolutionné le domaine de la traduction. Ces technologies ont considérablement accéléré et simplifié le processus de traduction, améliorant ainsi la communication entre les nations et les peuples. Cependant, la mise en œuvre de ces modèles nécessite des données importantes et précises pour former les modèles de manière efficace, et la traduction automatique est toujours confrontée à des défis importants dans des domaines tels que le maintien d'un sens précis et l'alignement culturel entre les langues. Même si des progrès notables ont été réalisés, ces modèles doivent encore être optimisés en permanence pour produire les meilleurs résultats. En examinant le domaine, nous constatons que de nombreuses études ont permis d'améliorer les processus de traduction automatique et d'apporter des réponses créatives aux problèmes auxquels ils sont confrontés. Dans le prochain chapitre, nous examinerons certaines de ces études antérieures qui mettent en évidence les avancées et les tendances actuelles dans le domaine de la traduction automatique, quelque chose qui aidera à la compréhension générale du domaine.

Chapitre 2

Deep learning

2.1 Les études précédentes (Related work)

À l'ère numérique, la traduction automatique est un outil essentiel pour promouvoir la compréhension interculturelle et joue un rôle important dans des domaines tels que l'éducation, les affaires, les médias et les affaires internationales. Le besoin d'une traduction précise en temps réel n'a jamais été aussi pressant, alors que le contenu numérique augmente et que les langues deviennent de plus en plus multidirectionnelles.

Les progrès rapides de l'intelligence artificielle (IA), en particulier des techniques d'apprentissage profond, ont rendu les systèmes de traduction automatique plus efficaces et capables de comprendre le contexte linguistique et de traiter des structures complexes. Cette section passe en revue les recherches antérieures sur la traduction automatique de différents types et méthodologies, en mettant l'accent sur les changements intervenus dans le domaine, des modèles traditionnels aux modèles neuronaux contemporains :

- **Marwa Hadj Salah**
  Dans ce travail, les chercheurs ont créé un système de traduction automatique statistique de l'anglais vers l'arabe basé sur l'outil open-source Moses. Le système a été mis en œuvre à l'aide d'une collection de modèles statistiques, y compris le modèle de langage construit par IRSTLM avec cinq règles de grammaire et le modèle de traduction basé sur les segments, en plus de l'utilisation de plusieurs blogs parallèles, y compris LDC-Ummah, LDC-News, News Commentary, et TED Talks. Le système est également basé sur l'outil MADAMIRA, qui traite l'arabe en termes de segmentation, de modélisation morphologique et d'analyse morphologique pour traiter les problèmes de tokenisation et de dé-tokenisation dans les textes arabes. La performance du système a été évaluée à l'aide de la méthode BLEU et a obtenu un score de 24,51, ce qui indique un niveau raisonnable de qualité de traduction entre les deux langues (Salah et al., 2018).

- **K. REZEG & M. T. LASKRI**
  Ce livre décrit une nouvelle méthode de traduction automatique de l'arabe vers le français en utilisant des réseaux neuronaux, en particulier le modèle Elman dans la classe des réseaux neuronaux récurrents (RNN). En développant un système d'auto-

apprentissage capable de traduire à partir d'exemples d'entraînement, le travail vise à dépasser les limites des modèles traditionnels basés sur des règles symboliques. Le système, qui est divisé en phases de traitement linguistique comprenant l'analyse, l'interprétation et la génération, est basé sur la représentation des phrases comme une chaîne de mots en utilisant la technique de l'association de mots tout en tenant compte du contexte. Malgré certaines contraintes liées à la taille des données limitées, le modèle a été testé sur de nouveaux mots et introduit dans une base de données bilingue (arabe-français), donnant des premiers résultats prometteurs. Pour améliorer la qualité de la traduction, les outils utilisés comprennent Python et des bibliothèques de réseaux neuronaux (comme Tano), en se concentrant sur la syntaxe et la sémantique des phrases (Rezeg & Laskri, 2014).

- **Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan†, Chris Callison-Burch**
  Cette étude est basée sur le développement d'un système de traduction automatique de l'arabe vers l'anglais en créant un ensemble de données parallèles dérivées de textes en ligne, en les classant par dialectes (tels que l'égyptien et le levantin), puis en les traduisant à l'aide de techniques de crowdsourcing. En comparant l'efficacité de la traduction directe de la darija vers l'anglais avec celle de l'ASM, les chercheurs ont utilisé des approches telles que l'analyse morphologique et la segmentation thématique du texte. D'après les résultats, la traduction directe est plus performante lorsqu'il y a suffisamment de données dans la langue cible (plus de 400 000 mots), mais l'utilisation des seules données ASM réduit l'efficacité et peut même nuire à la performance. Des données de test indépendantes provenant de messages Facebook envoyés par des Égyptiens ont également été utilisées pour valider la précision du modèle, démontrant la capacité du système à s'adapter aux textes précédemment saisis. L'étude conclut que des techniques linguistiques spécifiques au dialecte, telles que la segmentation morphologique et la classification dialectale, sont nécessaires pour améliorer la qualité de la traduction et souligne l'importance d'adapter le modèle au type de textes utilisés pour la formation et le test (Zbib et al., 2012).

- **SHAHAB AHMAD ALMAAYTAH & SOLEMAN AWAD ALZOBIDY**
  Ce document présente un examen systématique de toutes les difficultés rencontrées lors de la traduction de textes arabes en anglais à l'aide de systèmes de traduction automatique. Quatre problèmes principaux affectant la qualité de la traduction ont été identifiés : l'incapacité à distinguer les différents sens des mots, l'identification des entités nommées, les structures morphologiques riches et complexes, et le manque de ressources linguistiques, en particulier dans les dialectes arabes. D'autres problèmes ont également été mentionnés, tels que l'absence de signes diacritiques, les mots ambigus et incongrus, les erreurs orthographiques, la diversité dialectale, l'ordre libre des mots, le mauvais alignement des mots, la longueur inadéquate des phrases, les performances médiocres et le bruit des données. Certaines de ces questions sont interdépendantes, ce qui complique encore le processus de traduction. L'étude propose un certain nombre de solutions, telles que l'utilisation de modèles hybrides d'apprentissage profond, la segmentation morphologique et la densification

des données à l'aide de la traduction inverse ou de modèles multilingues. L'étude conclut que la traduction automatique de l'arabe vers l'anglais pose encore des problèmes importants qui nécessitent des recherches plus approfondies et suggère de créer un nouveau système de traduction qui aborde ces problèmes fondamentaux à l'avenir (Almaaytah & Alzobidy, 2023).

- **Mourad Broura, *, Abderrahim Benabbou**
  Cette étude porte sur la création d'un système de traduction de l'arabe vers la langue des signes arabe (ArSL) appelé ATLASLang MTS1, qui repose sur deux approches : l'analyse morphologique à l'aide du système Khalil sur la plate-forme SAFAR, la réorganisation des éléments grammaticaux de la phrase pour se conformer à la structure de la langue des signes (sujet général, verbe et objet), et l'application d'un ensemble de règles pour convertir les phrases en langue des signes. Suivi d'une animation assistée par GIF ou, si le mot n'est pas trouvé dans la base de données, d'une épellation au doigt, et enfin de l'utilisation d'un ensemble de règles pour la traduction des mots en langues des signes. Des expériences de traduction d'un certain nombre de phrases ont démontré l'efficacité du système malgré certaines limites, telles que l'ambiguïté de l'analyse morphologique. L'étude a conclu que la base de données devrait être élargie, que la grammaire utilisée devrait être élargie et que, dans les prochaines itérations, les graphiques animés devraient être remplacés par des avatars en 3D. Cet article décrit un système de traduction automatique de l'arabe vers la langue des signes arabe (ArSL) basé sur l'analyse morphologique et les techniques d'apprentissage profond. Les mots sont représentés par des modèles morphologiques qui sont transformés en vecteurs scalaires, et les caractéristiques morphologiques sont codées à l'aide de la représentation One-Hot. Les techniques d'apprentissage profond comprennent un réseau neuronal appelé Feed-Forward (FFNN) formé à l'aide de l'algorithme de rétropropagation, qui possède quatre couches en cache et utilise la fonction d'activation ELU (Exponential Linear Unit) dans les couches internes et une fonction sigmoïde dans les couches de sortie. Cette architecture est utilisée à la fois pour la classification morphologique et la génération de traductions. La sortie est ensuite traduite en langue des signes à l'aide d'un système de transformation basé sur des règles qui utilise les codes SiGML et HamNoSys pour créer des animations 3D représentant des phrases en langue des signes. Le système combine l'apprentissage profond et le traitement symbolique pour produire une traduction précise et efficace (Brour & Benabbou, 2019).

- **JEZIA ZAKRAOUI, MOUTAZ SALEH, SOMAYA AL-MAADEED & JIHAD MOHAMED ALJA'AM**
  Cet article traite de la traduction automatique (TA), en se concentrant sur la traduction entre l'arabe et les langues européennes, tout en examinant les difficultés et les contributions à l'amélioration de la précision et de la qualité de la traduction interlinguale. Un certain nombre d'ensembles de données qui soutiennent la formation et l'évaluation des modèles sont présentés, y compris Arab-Acquis, Tashkeela, et Arabic-SQuAD. L'accent a également été mis sur l'importance des procédures d'évaluation telles que la post-édition (PE) et la combinaison de la traduction neuronale et de la mémoire de traduction pour l'amélioration des résultats. La recherche a démontré que la NMT surpasse la SMT en termes de précision de traduction, en

particulier pour l'arabe. En outre, le modèle Transformer suscite un intérêt croissant pour l'amélioration des performances. Cependant, la traduction entre les dialectes arabes et les domaines spécialisés reste un défi important, et il est impératif de faire progresser les méthodes telles que l'apprentissage par transfert et la traduction multilingue pour résoudre ces problèmes à l'avenir (Zakraoui et al., 2021).

- **Lamis Ismail Omar * & Abdelrahman Abdalla Salih**
  Dans un premier temps, les chercheurs se concentrent sur l'évaluation et l'analyse des systèmes de traduction automatique anglais/arabe afin d'identifier les erreurs que ces systèmes produisent et de proposer des solutions pour améliorer leur précision à l'aide de techniques telles que la traduction automatique basée sur des règles et la traduction automatique statistique. Par la suite, entre 2010 et 2020, la recherche a progressé pour intégrer des méthodes de pointe comme la traduction automatique neuronale (NMT), en se concentrant sur l'amélioration de la précision de la traduction dans les textes compliqués et spéciaux. Après la pandémie de COVID-19 en 2020, l'intérêt des chercheurs pour ce domaine a considérablement changé, avec une augmentation de la recherche sur la traduction automatique en raison du passage à l'apprentissage en ligne et de la nécessité croissante d'une traduction rapide pour fournir des informations sur la pandémie. De 2020 à 2023, l'évaluation de l'efficacité de la traduction automatique pour la traduction de textes spécialisés, tels que les textes juridiques et littéraires, et l'importance de l'intégration de la traduction automatique avec la post-édition pour améliorer la qualité ont suscité beaucoup d'intérêt. Malgré ces avancées dans le domaine des technologies de traduction, il y a encore peu de recherches qui soutiennent l'intégration réelle des technologies de TA dans la formation universitaire des traducteurs, et il y a un manque d'études sur la façon de former les traducteurs à l'utilisation des technologies de TA dans les programmes d'enseignement (Salih & Omar, 2023).

- **Arwa Alqudsi, Nazlia Omar & Khalid Shaker**
  Les chercheurs mentionnés dans le livre ont apporté des contributions significatives au domaine de la traduction automatique en se concentrant sur l'amélioration des méthodes de traduction entre l'arabe et un certain nombre d'autres langues à l'aide d'une variété d'approches différentes. Par exemple, Alsharaf et al. (2004) ont introduit une nouvelle approche de la traduction du français vers l'arabe en combinant des techniques traditionnelles telles que la grammaire et les statistiques avec un accent sur les phénomènes linguistiques entre les deux langues. Habash et Hu (2009) ont comparé deux approches de la traduction de l'arabe vers le chinois en utilisant l'anglais comme médiateur, démontrant ainsi l'impact des modèles multilingues sur le traitement de la traduction. En utilisant une approche grammaticale pour optimiser la traduction entre l'arabe et l'hindi, Mark et al. (2004) ont développé une comparaison analytique entre les deux langues. Ils se concentrent sur l'utilisation de données parallèles et de corpus multilingues pour améliorer les méthodes de traduction. Bien que cette recherche ne soit pas spécifiquement axée sur l'apprentissage profond, elle a eu un impact significatif sur le développement de modèles neuronaux contemporains pour la traduction de textes en offrant des environnements idéaux pour la construction de bases de données que les chercheurs utilisent aujourd'hui pour créer des modèles de traduction automatique basés sur l'apprentissage profond

(Alqudsi et al., 2014).

2.2 L'apprentissage profond (en anglais : deep learning)

2.2.1 L'historique de l'apprentissage profond

mobiskill « équels sont les algorithmes de deep learning ? », 26 mai 2021
Le développement des réseaux de neurones artificiels depuis leur proposition initiale en 1943. Il mentionne le développement du « perceptron » en 1957, qui marque le début du réseau de neurones artificiels, et l'apparition du premier modèle binaire de « neurone formel ». Ensuite, il mentionne « l'hiver »de l'« IA » et des développements ultérieurs, comme le « perceptron multicouche » dans les années 1980.L'expression « apprentissage profond » a été utilisée pour la première fois en référence aux recherches de Yann LeCun sur les réseaux neuronaux convolutés. Néanmoins, son utilisation a été limitée en raison des exigences élevées en matière de puissance de calcul.En 2012, l'apprentissage profond a connu un regain d'intérêt grâce à un programme apparu lors d'une tâche de reconnaissance visuelle. Depuis, la grande majorité des entreprises utilisant l'apprentissage automatique sont passées à l'apprentissage profond dans de nombreux domaines (DE SONS, 2016).

2.2.2 Définition

L'apprentissage profond, parfois appelé « deep learning », est un ensemble d'approches d'apprentissage automatique qui a permis des avancées significatives dans le domaine de l'intelligence artificielle au cours des dernières années. Dans l'apprentissage automatique, un programme examine un ensemble de données pour en extraire les règles qui permettront de tirer des conclusions à partir de nouvelles données règles qui permettront de tirer des conclusions à partir de nouvelles données. L'apprentissage profond repose sur ce que l'on a appelé, par analogie, des « réseaux de neurones artificiels », composés de millions d'unités (les « neurones ») qui exécutent des tâches individuelles simples.Les résultats de la première couche de « neurones » sont utilisés pour les calculs de la deuxième couche, et ainsi de suite. Pour la reconnaissance visuelle, par exemple, les premières couches d'unités identifient les lignes, les courbes, les angles, etc.Les canapés de niveau supérieur qui identifient les formes, les combinaisons de formes, les objets, les contextes, etc. Les progrès de l'apprentissage profond ont été rendus possibles, en partie, par le développement de grandes bases de données (souvent appelées « big data ») et l'augmentation de la puissance des ordinateurs (Boughaba et al., 2017).

2.2.3 Domaines d'application de l'apprentissage profonde

Ces méthodes sont développées dans le domaine des applications informatiques aux NTIC (reconnaissance visuelle, comme un panneau de signalisation utilisé par un robot ou un véhicule autonome), à la bio-information, à la reconnaissance ou à la comparaison

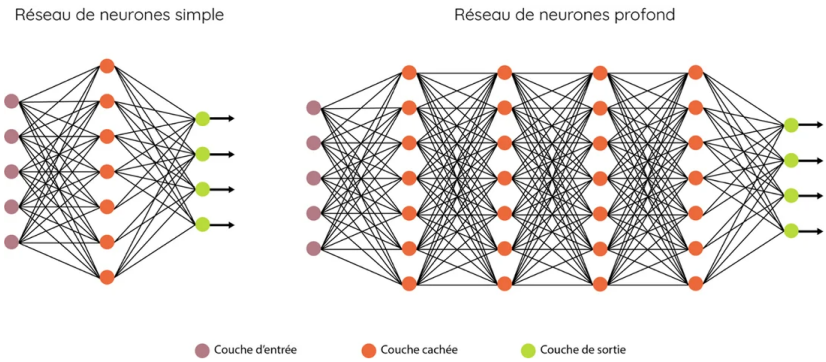

FIG. 2.1 : l'architecture de d'apprentissage profond (DILEPIX, 2023)

de formes, à la sécurité,l'éducation assistée par ordinateur, et plus généralement de l'intelligence artificielle.
Elle aide à prédire certaines propriétés : les propriétés d'un sol filmé par un robot , la reconnaissance ou la comparaison de formes ou d'objets hautement déformables, l'analyse de mouvements et positions des doigts d'une main, ce qui peut être utile pour traduire les langues signées , le positionnement automatique d'une caméra , la télédétection (notamment en imagerie satellitaire) , l'art numérique , l'histoire de l'art ; par exemple la création d'œuvres artistiques à partir d'une photo, la reconnaissance vocale de voix humaine ou de signaux sonores , la robotique , la sécurité , la pédagogie assistée par l'informatique , l'intelligence artificielle en général , des jeux de société complexes (échecs , go , shogi) , la traduction automatique (moteurs tels que Google traduction , DeepL, Pons) , l'analyse d'émotions révélées par un visage photographié ou filmé .

2.3 Principes de fonctionnement

Le cerveau humain se compose de millions de neurones interconnectés. Les algorithmes de l'apprentissage profond imitent cette configuration naturelle du cerveau humain en utilisant des réseaux de neurones artificiels composés de multiples couches de nœuds interconnectés, appelés neurones artificiels ou unités. Chaque couche traite l'information venant de la couche précédente pour la transmettre à la couche suivante, ce qui permet une extraction plus fine des caractéristiques et des schémas.
Les modèles d'apprentissage profond sont entraînés à partir de grandes quantités de données étiquetées, appelées ensembles d'entraînement, pour apprendre comment reconnaître et classer des schémas. Il peut s'agir d'ensembles d'entraînement simples, comme un ensemble de données « chiens et chats » où la vision artificielle doit classer des photos selon qu'elles contiennent un chien ou un chat. Ces ensembles de données peuvent aussi être plus sophistiqués et couvrir par exemple l'appréciation du vin ou la détection de fausses informations dans des rapports authentiques (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO), s. d.).

2.3.1 Les différents types de modèles de deep learning

Dans le domaine de l'apprentissage profond, plusieurs variétés de réseaux de neurones sont utilisées, chacune ayant des caractéristiques uniques qui la rendent adaptée à différents types d'applications. Parmi les modèles les plus connus et les plus utilisés figurent :

- Les réseaux neuronaux convolutifs (CNN), principalement utilisés dans le traitement des images ;
- Les réseaux neuronaux récurrents (RNN), adaptés aux données séquentielles ;
- Les réseaux à mémoire à long terme (LSTM), une variante avancée des RNN utilisée dans des tâches telles que la traduction automatique.

Le réseau de neurones convolutif (CNN)

Les réseaux neuronaux convolutifs (CNN), souvent appelés ConvNets, sont des réseaux neuronaux profonds conçus pour traiter des données structurées en grille, telles que des images. Ils sont largement utilisés dans des domaines tels que la vision artificielle, l'analyse d'images médicales, la prédiction de séries temporelles et la détection d'anomalies (Sadoon & Ali, 2021a).

Les CNN comportent plusieurs couches qui traitent et extraient les caractéristiques des données :

- **Couche de convolution** : le CNN utilise une couche de convolution avec de nombreux filtres pour effectuer l'opération de convolution.
- **Unité linéaire rectifiée (ReLU)** : les CNN utilisent la couche ReLU pour effectuer des opérations sur les éléments. Le résultat est une carte caractéristique corrigée.
- **Couche de pooling** : placée entre deux couches de convolution, elle applique la mise en commun sur de nombreuses cartes de caractéristiques. Elle divise l'image en blocs et conserve le maximum pour chaque bloc, réduisant ainsi la taille de l'image tout en préservant les caractéristiques les plus importantes.
- **Couche fully connected** : cette couche accepte un vecteur d'entrée et produit un nouveau vecteur de sortie via une combinaison linéaire suivie éventuellement d'une fonction d'activation.

Réseaux de neurones récurrents (RNN)

Les RNN sont conçus pour traiter des données séquentielles comme les séries temporelles ou les textes. Contrairement aux réseaux neuronaux classiques, les RNN ont des connexions récurrentes qui leur permettent de conserver la mémoire des états précédents, facilitant la modélisation des dépendances temporelles. Ils sont utilisés dans des domaines tels que la prévision météo, la finance, la traduction automatique, la reconnaissance audio et l'analyse des sentiments (Wikipedia contributors, s. d.-b).

Principe de fonctionnement des RNN

Un RNN se distingue par sa capacité à traiter des séquences de données grâce à des connexions récurrentes. Contrairement aux réseaux feedforward, les RNN disposent d'une mémoire interne permettant de garder la trace des états précédents.

L'architecture typique d'un RNN comprend une couche d'entrée, une ou plusieurs couches de cache récurrentes et une couche de sortie. À chaque instant temporel t, un neurone récurrent reçoit l'entrée actuelle $x(t)$ et sa propre sortie précédente $y(t-1)$. Cela permet de modéliser les dépendances temporelles des données séquentielles.

Chaque neurone récurrent a deux types de poids :

- Des poids notés W reliant les entrées à la sortie (comme pour un réseau classique) ;
- Des poids notés R entre la sortie et l'entrée de la couche, qui sont les connexions récurrentes.

Fig. 2.2 : RNN et sa version dépliée dans le temps

$X(t)$ et $Y(t-1)$ déterminent $Y(t)$, qui à son tour détermine $X(t-1)$ et $Y(t-1)$, qui à son tour détermine $X(t-2)$ et $Y(t-3)$, … À partir de l'instant $t = 0$, $Y(t)$ est donc une fonction de toutes les entrées. Pendant la première phase temporelle, lorsque $t = 0$, les sorties précédentes n'existent pas (en général, elles sont supposées nulles).

Ètant donné que la sortie d'un neurone récurrent dépend de toutes les entrées des étapes précédentes, on dit qu'il possède une sorte de mémoire.

La partie du réseau RNN qui conserve son état sur plusieurs étapes temporelles est appelée **cellule mémoire**. L'état d'une cellule à l'instant t, désigné par $h(t)$ (h pour couche cachée), est une fonction de certaines entrées à cet instant et de son état à l'instant précédent :

$$h(t) = f(h(t-1), x(t))$$

Par conséquent, la sortie dépend de l'état précédent et des entrées actuelles.

2.3.2 Les réseaux LSTM

Les réseaux à mémoire à long terme (LSTM) sont une conception unique de réseaux neuronaux récurrents (RNN) introduite par Hochreiter et Schmidhuber en 1997. Ces réseaux ont été conçus pour résoudre le problème de disparité de gradient auquel sont

confrontés les RNN en leur permettant de mémoriser des informations sur de longues séquences temporelles. La cellule LSTM contient des mécanismes de porte qui contrôlent le flux d'informations, permettant aux anciennes informations de passer par plusieurs étapes temporelles sans être perdues (Sadoon & Ali, 2021b).

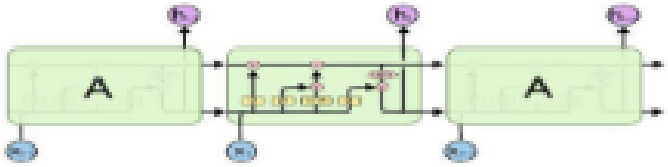

Fig. 2.3 : Chaîne de cellules LSTM
(CleanPNG, s. d.)

Les calculs se déroulent comme suit :(Olah, 2015)

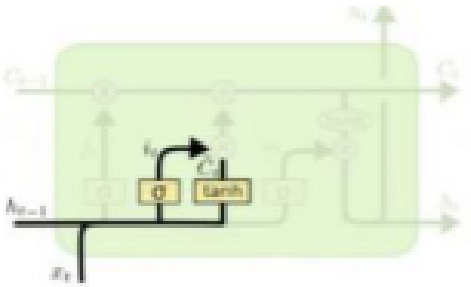

Fig. 2.4 : Opérateur d'oubli d'informations
(CleanPNG, s. d.)

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.1}$$

Tel que :

- h_{t-1} : Sortie à l'instant $t-1$
- x_t : Entrée courante à l'instant t
- b_f : Biais associé à la porte d'oubli
- W_f : Poids associés à la porte d'oubli
- σ : Fonction d'activation sigmoïde

- **Porte d'entrée (Input gate)** :

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.2}$$

- **Valeur candidate de l'état de la cellule** :

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.3}$$

Tel que :

- tanh : Fonction d'activation tangente hyperbolique
- $\tilde{C}_t$: Valeur candidate pour l'état de la cellule

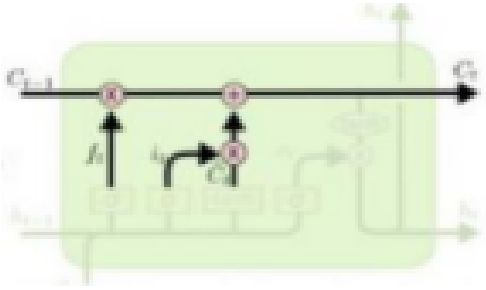

FIG. 2.5 : Mise à jour de la mémoire ct
(CLEANPNG, s. d.)

- **Mise à jour de l'état de la cellule** :

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.4}$$

Tel que :

- C_t : État interne de la cellule à l'instant t

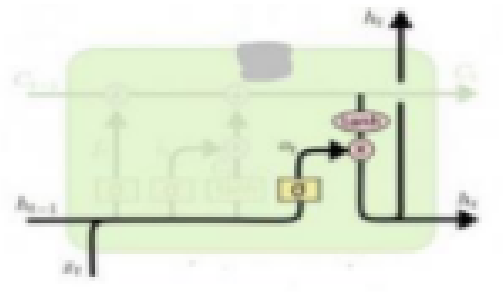

FIG. 2.6 : Sortie de la couche cachée
(CLEANPNG, s. d.)

- **Porte de sortie (Output gate)** :

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.5}$$

- **Sortie finale de la cellule** :

$$h_t = o_t * \tanh(C_t) \tag{2.6}$$

Tel que :

– h_t : Sortie de la cellule à l'instant t

2.3.3 Réseaux de neurones Artificiels (ANN)

Un réseau de neurones artificiels est une architecture composée de couches consécutives de nœuds (ou neurones formels) qui sont interconnectées et peuvent modéliser des transformations non linéaires de vecteurs d'entrée en vecteurs de sortie. La topologie du réseau, en particulier le nombre de couches et l'emplacement des neurones, influence les performances du modèle. Ces réseaux sont souvent utilisés pour la classification et l'apprentissage supervisé (Goodfellow et al., 2016).

Un neurone formel est caractérisé par :

- Le type des entrées et des sorties.
- Une fonction d'entrée.
- Une fonction de sortie.

2.3.4 Transformer

Le modèle Transformer pour la traduction automatique est un cadre entièrement basé sur le mécanisme d'auto-attention. Le codeur et le décodeur calculent les poids d'attention entre chaque point d'une séquence de mots, ce qui permet au modèle de capturer efficacement les dépendances à longue distance sans s'appuyer sur des réseaux récurrents. Cette conception permet un apprentissage totalement parallèle sur des séquences, ce qui accélère l'apprentissage et améliore la qualité de la traduction, en particulier pour les modèles linguistiques compliqués (Vaswani et al., 2017b).

Attention model

L'auto-attention est un mécanisme de réseau neuronal qui transforme chaque élément d'une séquence d'entrée en tenant compte de tous les autres composants de la séquence. Il est particulièrement utile dans les problèmes de modélisation de séquences. Cela permet au modèle de capturer dynamiquement les dépendances indépendamment de leur distance dans l'entrée en calculant une somme pondérée des éléments d'entrée, où les poids sont définis par la similarité des éléments dans un espace de caractéristiques apprises".

« L'auto-attention peut être considérée comme une forme d'adressage basé sur le contenu au sein d'une séquence, où la représentation de chaque élément est affinée par les informations provenant de tous les éléments, ce qui améliore la capacité du modèle à comprendre les relations contextuelles et les dépendances à long terme (Wikipedia contributors, s. d.-a).

Multi-Head Attention

La technique d'attention multi-têtes permet au modèle de s'intéresser simultanément à des informations provenant de divers sous-espaces de représentation à différents endroits. Au lieu de remplir une seule fonction d'attention, le modèle exécute plusieurs mécanismes d'attention, ou « têtes », simultanément. Chaque tête possède une collection unique de projections linéaires apprises pour les requêtes, les clés et les valeurs. Le résultat final est obtenu en concaténant les sorties de toutes les têtes et en les projetant à nouveau (Vaswani et al., 2017a).

L'architecture codeur-décodeur (encoder-decoder)

L'architecture codeur-décodeur est une conception de réseau neuronal qui fonctionne particulièrement bien pour les tâches de séquence à séquence telles que la traduction automatique. Le codeur lit et traite la séquence d'entrée, puis l'encode dans un vecteur de contexte de taille fixe. Le décodeur utilise ensuite ce vecteur de contexte pour construire la séquence de sortie, un élément à la fois".

"Dans la traduction automatique neuronale, le codeur prend la phrase source et compresse ses informations dans un vecteur (ou une séquence de vecteurs), et le décodeur construit la phrase cible (Sutskever et al., 2014).

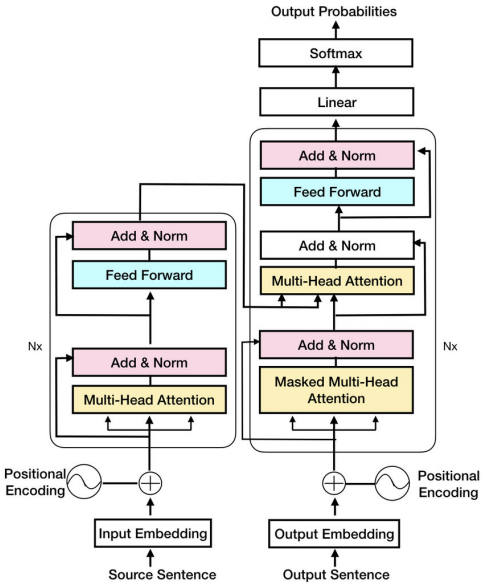

Fig. 2.7 : L'architecture encoder-decoder (Varagh, 2020)

2.4 Conclusion

En conclusion, il est clair que la traduction automatique est un domaine de recherche en constante évolution, grâce aux progrès de l'intelligence artificielle et de l'apprentissage profond. L'utilisation de modèles neuronaux, tels que les réseaux convolutifs et récurrents, a considérablement amélioré la qualité de la traduction, en particulier dans le traitement de langues compliquées comme l'arabe. Les études précédentes examinées dans ce chapitre démontrent un intérêt croissant pour l'adaptation des méthodologies modernes aux caractéristiques linguistiques de la langue arabe. Par conséquent, l'intégration de l'apprentissage profond dans les systèmes de traduction ouvre la voie à des performances plus fiables et à des opportunités de développement prometteuses dans le domaine de la traduction automatique.

Chapitre 3

Méthodologie de travaile

3.1 Introduction

Les fondements théoriques de la traduction automatique ainsi que les particularités linguistiques des langues impliquées — arabe, anglais et dialecte algérien — ont été abordés dans les chapitres précédents.

Dans ce chapitre, nous décrivons en détail la méthodologie adoptée pour la construction et l'évaluation de notre propre modèle de traduction. Cette méthodologie comprend :

- la constitution d'un corpus bilingue aligné entre la darija algérienne et l'anglais,
- les étapes de prétraitement des données textuelles,
- le choix de l'architecture du modèle (BERT),
- les critères d'évaluation appliqués pour mesurer la qualité des traductions produites.

Le modèle a été entraîné sur un ensemble de données représentatif de la diversité linguistique de la darija. Les performances du système ont ensuite été évaluées à l'aide d'indicateurs objectifs, permettant d'identifier ses points forts ainsi que les limitations rencontrées dans la traduction de cette langue peu standardisée.

3.2 Notre projet

L'objectif principal de cette étude est de développer un modèle de traduction automatique performant, capable de traduire des phrases en darija algérienne vers l'anglais.

Pour ce faire, nous avons constitué un corpus bilingue de plus de 24 000 phrases soigneusement alignées, couvrant différents domaines et registres linguistiques représentatifs de l'usage réel de la darija.

Le modèle proposé, basé sur l'architecture BERT, a été entraîné sur ce corpus et évalué à l'aide de métriques standard de qualité de traduction. L'étude met en évidence les défis liés à la traduction de cette langue peu standardisée, et analyse dans quelle mesure un modèle de type Transformer peut y répondre efficacement.

3.2.1 Collecte de données

Dans un premier temps, nous avons collecté des données sur *Kaggle*, où nous avons sélectionné un ensemble de données contenant des phrases anglaises traduites en arabe. Ces données ont été téléchargées au format `JSON` ou `CSV` et contenaient les champs suivants :

- `english_translation` : La phrase en anglais.
- `formal_arabic` : La traduction en arabe standard.
- **Nom du dataset** : `translation_corrected_json`

Format : JSON

Source : téléchargé depuis Kaggle

Contenu : paires de phrases alignées (anglais darija algérienne)

Objectif : entraîner un modèle de traduction automatique basé sur Transformer

taille total : 24000 paires de phrases

Notre objectif étant de traduire en *darija* algérienne, nous avons converti la traduction en arabe standard en *darija* algérienne de manière manuelle et semi-manuelle, en respectant le contexte général de chaque phrase.

Afin d'enrichir les données et d'explorer d'autres possibilités par la suite, telles que la traduction émotionnelle ou l'analyse de l'humeur, nous avons ajouté une étiquette de sentiment à chaque phrase :

- positif
- négatif
- neutre

3.2.2 Conversion JSON finale

Après avoir traduit toutes les phrases en darija algérienne et classé les sentiments, nous avons compilé toutes les données dans un seul fichier : `translation_corrected.json`.

Le fichier contient environ 18 487 exemples, ce qui est suffisant pour entraîner le modèle *Transformer*.

3.3 nettoyage des données

Dans un premier temps, nous avons obtenu une base de données contenant des traductions de l'anglais vers la *darija* et la *fusha* algériennes, avec les sentiments associés à chaque phrase. Ces données ont été stockées dans un fichier `JSON` nommé : `translation_corrected.json`.

Nous avons téléchargé le fichier et commencé à nettoyer les textes des espaces supplémentaires, des symboles étranges, et à les convertir en minuscules.

3.4 Division des données

Après avoir nettoyé les transcriptions, nous avons divisé les données en trois catégories :

- 70% pour l'entraînement
- 15% pour la validation
- 15% pour le test

3.5 Construction du vocabulaire

Nous avons extrait tous les mots les plus fréquents des phrases anglaises et familières et créé un vocabulaire qui comprend des mots ainsi que des symboles spéciaux tels que `<pad>`, `<sos>`, `<eos>` et `<unk>`.

3.6 Tokénisation des phrases en nombres

Dans cette étape, nous avons converti les phrases en une séquence de nombres en utilisant le vocabulaire que nous avons créé.

3.6.1 Mise en place d'un DataLoader pour alimenter le modèle

Nous avons encapsulé les données dans un objet `Dataset`, puis créé un `DataLoader` qui distribue les données par petits lots pendant la formation.

3.6.2 Construire un modèle de transformateur

Grâce à notre compréhension de la théorie des transformateurs, nous avons construit notre propre modèle en utilisant plusieurs couches de codage et de décodage, avec une couche d'intégration et une couche linéaire pour la sortie.

3.6.3 Entrainement du modèle

Nous avons commencé le processus de formation en utilisant `Adam Optimiser` et `CrossEntropy Loss`. Nous avons suivi la perte dans chaque cycle d'apprentissage (époques) pour contrôler la performance du modèle.

3.6.4 Évaluation et tests

Enfin, nous avons évalué le modèle sur l'ensemble de test en utilisant l'échelle `BLEU` pour mesurer la qualité de la traduction.

3.6.5 L'attention, l'encodeur et le décodeur dans le modèle de transformation de la traduction

Dans notre projet de traduction de textes de l'anglais vers la darija algérienne, nous nous sommes appuyés sur l'architecture `Transformer`, une architecture puissante qui a émergé pour remplacer les réseaux récurrents (`RNN`) dans les tâches de traitement des langues, en particulier la traduction automatique.

Encodeur (Compréhension de la phrase source)

L'encodeur est la partie responsable de la lecture de la phrase originale et de la compréhension de son sens, en convertissant les mots en représentations numériques appelées *embeddings*, puis en les faisant passer par une série de couches contenant le mécanisme d'attention, ce qui permet au modèle de se concentrer sur les relations entre les mots.

- Chaque mot est représenté sous forme de vecteur à l'aide d'une couche d'intégration.
- Des informations sur l'ordre des mots sont ensuite ajoutées par le biais du codage positionnel.
- Ces représentations passent ensuite par une série de couches, chacune ayant sa propre couche :
  - **Auto-attention multi-têtes** : Permet à chaque mot de tenir compte de tous les autres mots de la phrase.
  - **Réseau feedforward** : un petit réseau neuronal appliqué à chaque mot.
  - **LayerNorm et Dropout** pour améliorer la stabilité.

Équations importantes dans l'encodeur

Dans la couche *self-attention*, elle est calculée comme suit :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

Où :

- **Q, K, V** : Représente la requête, la clé et la valeur (calculées à partir de la même entrée).
- d_k : Le nombre de dimensions de la clé (utilisé pour minimiser l'explosion des valeurs).
- La sortie est la nouvelle représentation du mot, en tenant compte du contexte complet de la phrase.

Décodeur (Production de la traduction)

Le décodeur est la partie qui génère la traduction mot par mot, sur la base de ce que l'encodeur a compris de la phrase originale.

Il se compose également de plusieurs couches, et dans chaque couche, il y a :

- **Masked Multi-Head Attention** : Pour empêcher le modèle de voir les mots à venir pendant la traduction.
- **Attention codeur-décodeur** : Permet au décodeur de se concentrer sur les mots originaux pertinents pendant la traduction.
- **Feedforward + LayerNorm + Dropout** : est identique à l'encodeur.

Chaque étape de la traduction est basée sur les étapes précédentes et sur les informations tirées de la phrase originale.

Encodeur-Décodeur Attention

Même équation que pour *Attention* mais :

- Q = output from decoder, K, V = output from encoder

Et c'est là que réside la magie : Le modèle apprend où chercher dans la phrase originale pour générer la traduction appropriée.

3.6.6 Le rôle de l'attention dans la traduction

Elle permet au modèle de traiter les phrases longues plus efficacement que le `RNN`. Le modèle ne s'appuie pas sur une séquence étape par étape, mais examine tous les mots à la fois. Cela lui permet de comprendre les relations distantes entre les mots (par exemple, la relation entre le pronom et le verbe d'origine, même s'ils sont séparés par plusieurs mots).

3.6.7 Optimiseur Adam

`Adam` (*Adaptive Moment Estimation*) est un algorithme d'optimisation basé sur le suivi des moyennes mobiles des gradients et de leurs carrés. C'est l'un des algorithmes les plus populaires en apprentissage automatique car il combine les caractéristiques d'`AdaGrad` et de `RMSProp`.

Cela signifie qu'il est responsable de l'ajustement des poids du modèle pendant l'apprentissage d'une manière intelligente et rapide, de sorte que le modèle apprend à traduire plus efficacement.

3.7 Conclusion

Ce chapitre a présenté en détail la méthodologie adoptée pour concevoir et mettre en œuvre notre système de traduction automatique basé sur l'architecture Transformer. Nous avons d'abord décrit les étapes de préparation des données, en insistant sur le choix des corpus, leur nettoyage et leur structuration afin de garantir une qualité optimale pour l'entraînement. Ensuite, nous avons expliqué la mise en place du modèle, depuis la création du DataLoader jusqu'à la construction de l'architecture complète, en passant par le choix des fonctions de perte et de l'optimiseur.

La phase d'entraînement a été menée de manière rigoureuse, avec un suivi précis de la performance du modèle à travers plusieurs époques. L'évaluation, quant à elle, a été assurée à l'aide de métriques standard comme le score BLEU, permettant d'apprécier la qualité des traductions générées.

Enfin, une attention particulière a été portée sur les composants clés de l'architecture Transformer, notamment le mécanisme d'attention, l'encodeur et le décodeur, ainsi que leur rôle dans l'amélioration de la compréhension et de la génération linguistique.

Les fondations méthodologiques posées dans ce chapitre constituent ainsi un socle solide pour l'expérimentation et l'analyse des résultats, qui seront présentés dans le chapitre suivant.

Chapitre 4

Conception, Implémentation

4.1 Introduction

Ce chapitre se concentre principalement sur les expériences menées dans le cadre de notre projet de traduction automatique. Son objectif est de présenter les pratiques expérimentales permettant de comparer la traduction humaine, la traduction effectuée par notre propre modèle basé sur le BERT, et la traduction produite par les modèles automatiques existants. Pour ce faire, nous décrirons tout d'abord les outils techniques utilisés, tels que le langage de programmation, l'environnement de développement et les ressources matérielles. Ensuite, nous expliquerons la méthode de traitement des données multilingues, y compris les étapes de prétraitement, d'entraînement du modèle et d'évaluation. Nous présentons ensuite les résultats obtenus lors de plusieurs tests de traduction, ainsi qu'une analyse comparative et des discussions détaillées. Enfin, nous clôturons ce chapitre par une conclusion synthétisant les observations majeures issues de ces expérimentations.

4.2 Outils matériels et logiciels

4.2.1 Configuration matérielle

Ce travail a été implémenté sur un PC, caractérisé comme suit :

- **Processeur :** Intel Core i3-6006U CPU @ 2.00GHz 1.99GHz
- **Mémoire RAM :** 8 GB
- **Système d'exploitation :** Windows 10 Pro 64 bits

4.2.2 Environnement logiciel

Nous avons utilisé **Python**, un langage de programmation bien connu pour le traitement automatique des langues en raison de son abondance de bibliothèques spécialisées

telles que `transformers`, `torch` et `datasets`. Python est un langage *open source*, flexible et intuitif qui permet un prototypage rapide et efficace des algorithmes d'apprentissage profond.

Au lieu d'une installation locale traditionnelle, nous avons utilisé la plateforme **Kaggle** de Google, qui fournit des ressources matérielles gratuites telles que des GPU (en particulier *NVIDIA Tesla T4*). Cette plateforme fournit un environnement de développement intégré basé sur `Jupyter Notebook`, permettant un codage structuré et collaboratif, l'exécution, la sauvegarde et la documentation de nos expériences.

En utilisant Kaggle, nous avons pu surmonter les restrictions physiques de nos propres ordinateurs et bénéficier d'un écosystème prêt à l'emploi pour l'exécution de modèles sophistiqués. Dans cet environnement, nous avons mis en œuvre notre modèle **BERT**, traité des données et évalué la performance de la traduction à travers plusieurs expériences.

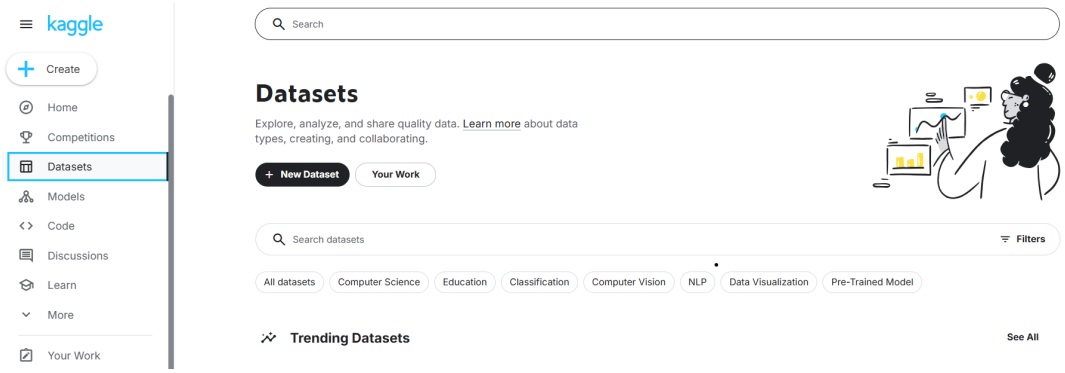

FIG. 4.1 : Environnement logiciel

Jupyter Notebook est une plateforme *open source* pour l'exécution et le partage de code. Il est principalement utilisé pour le développement et l'exécution de code Python, mais il fonctionne également avec d'autres langages de programmation.

Jupyter permet aux utilisateurs de créer des carnets de notes interactifs qui intègrent du code, des images, du texte et des résultats en temps réel. Ces carnets peuvent être partagés et collaborer avec d'autres utilisateurs, ce qui en fait un outil précieux pour l'exploration des données, l'apprentissage automatique, l'analyse statistique et d'autres domaines (KLUYVER et al., 2016).

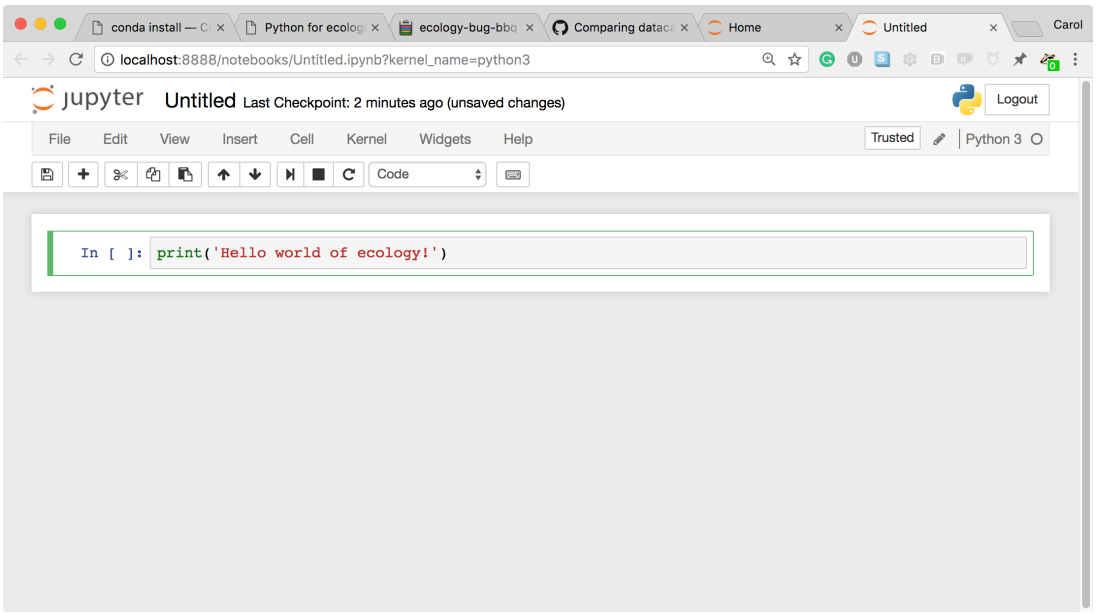

FIG. 4.2 : interface de Jupyter Notebook

4.3 Bibliothèques utilisées

Dans cet environnement, nous avons installé plusieurs packages qui ont facilité la programmation, parmi lesquels :

4.3.1 PyTorch

PyTorch est une bibliothèque d'apprentissage profond open source développée par Meta AI. Elle est largement utilisée dans le domaine de l'intelligence artificielle, en particulier pour la vision par ordinateur et le traitement du langage. Elle se distingue par son approche dynamique de la conception des graphes de calcul, qui offre un haut degré de flexibilité lors du développement et du débogage des modèles. L'interface intuitive de PyTorch et sa compatibilité avec CUDA permettent un prototypage rapide et une exécution efficace des algorithmes d'apprentissage profond, ce qui en fait un outil privilégié dans la recherche académique et industrielle (PASZKE et al., 2019).

`NLTK` (Natural Language Toolkit) est une bibliothèque open-source largement utilisée pour le traitement automatique du langage naturel (TAL) en Python ; elle fournit un ensemble complet d'outils et de ressources linguistiques, y compris des corpus annotés, des fonctions de tokenisation, d'étiquetage grammatical, d'analyse syntaxique et d'extraction d'entités, ce qui en fait un outil privilégié dans les contextes universitaires et de recherche pour l'analyse et la modélisation (BIRD et al., 2009).

4.3.2 Rouge score (Recall-Oriented Understudy for Gisting Evaluation)

est une technique largement utilisée dans le traitement du langage naturel pour évaluer la qualité des CV automatisés en les comparant à des CV humains. Cette métrique évalue la similarité entre les textes générés et les textes de référence principalement en calculant la récupération des n-grammes, des phrases ou des mots, en mettant l'accent sur le rappel (recall) des éléments inclus dans les références. Il existe de nombreuses variantes, telles que ROUGE-N (basée sur les n-grammes) et ROUGE-L (basée sur la plus longue sous-séquence commune), qui permettent une évaluation quantitative de la précision et de l'exhaustivité du CV automatisé par rapport au texte original (Lin, 2004).

4.3.3 Matplotlib

Matplotlib est une bibliothèque Python open source utilisée pour créer des visualisations graphiques bidimensionnelles flexibles et avancées. Elle fournit une large gamme d'outils pour créer de nombreux types de graphiques, tels que des courbes, des histogrammes, des diagrammes de dispersion et des représentations sophistiquées, avec un haut niveau de personnalisation visuelle. Matplotlib est largement utilisé dans les domaines scientifiques et techniques pour l'analyse des données et la présentation visuelle des résultats, et c'est un outil essentiel dans l'écosystème de la science des données en raison de son intégration avec d'autres bibliothèques telles que NumPy et Pandas (Hunter, 2007).

4.3.4 JSON (JavaScript Object Notation)

est un format de texte léger utilisé pour l'échange de données entre systèmes, facile à lire et à écrire pour les humains, ainsi qu'à analyser et à générer par les machines. Il repose sur une structure de données basée sur des paires clé-valeur et des listes ordonnées, ce qui lui permet de représenter des données complexes de manière organisée et standardisée. JSON est largement utilisé dans les applications web, en particulier pour la communication entre les serveurs et les navigateurs, et constitue une alternative populaire à des formats tels que XML en raison de sa simplicité et de son efficacité (Bray, 2014).

4.3.5 RANDOM

se réfère à une valeur ou à un phénomène qui se produit sans modèle connu ou prévisible. Cette notion est fondamentale en mathématiques, en statistiques et en informatique pour décrire des événements qui suivent une distribution de probabilité spécifique. En programmation et en science des données, le terme « aléatoire » est utilisé pour générer des nombres ou des événements imprévisibles, en particulier dans la simulation, les expériences stochastiques, l'échantillonnage aléatoire et l'amélioration des algorithmes d'apprentissage automatique grâce à des techniques telles que la randomisation ou la sélection (Gentle, 2009).

4.3.6 Tqdm

est une bibliothèque Python open source qui est utilisée pour afficher des barres de progression simples et esthétiques pendant l'exécution de boucles ou de longues tâches dans les logiciels. Elle permet aux utilisateurs de voir facilement la progression des opérations en affichant le pourcentage d'achèvement, le temps écoulé et le temps restant estimé, améliorant ainsi le contrôle des performances et l'expérience de l'utilisateur dans des environnements interactifs ou non. La Tqdm est largement utilisée dans des domaines tels que l'analyse de données et l'apprentissage automatique, où les processus peuvent être longs (da Costa-Luis, 2019).

4.3.7 Os

est une bibliothèque Python commune qui fournit une interface pour interagir avec le système d'exploitation. Elle vous permet d'effectuer diverses opérations liées au système, telles que la gestion de fichiers et de répertoires, l'obtention d'informations sur l'environnement d'exécution, l'exécution de commandes système et la manipulation de chemins et de variables. Les programmes Python peuvent devenir multiplateformes grâce à `os` tout en ayant accès aux fonctionnalités essentielles du système d'exploitation (Foundation, 2024).

4.4 Les résultats des algorithmes

Dans cette partie, nous allons afficher les résultats obtenus dans chaque algorithme, avec un résultat final comme les meilleurs paramètres pour cet algorithme

4.4.1 Évolution des performances du modèle Transformer au cours des époques entraînement

Époque	Loss	BLEU	METEOR	ROUGE-L F1	BERTScore F1
1	8.3552	0.0000	0.0000	0.0000	0.0000
2	8.2456	0.0000	0.0000	0.0000	0.0000
3	8.1254	0.0000	0.0000	0.0000	0.0000
4	6.7651	3.8471	0.1093	0.3045	0.7721
5	6.3421	5.9012	0.1780	0.4053	0.7980
6	6.0003	7.3271	0.2215	0.4447	0.8133
7	5.8672	6.9823	0.2152	0.4388	0.8110
8	5.7219	7.4120	0.2329	0.4473	0.8145
9	5.5993	7.6931	0.2381	0.4522	0.8179
10	5.5814	7.5914	0.2399	0.4534	0.8175

Tab. 4.1 : Évolution des métriques de performance au fil des époques

Le tableau présente l'évolution des performances du modèle de traduction sur 10 époques, évaluées à l'aide de plusieurs métriques : **BLEU**, **METEOR**, **ROUGE-L F1** et **BERTScore F1**.

- Durant les trois premières époques, toutes les métriques restent nulles (0.0000), indiquant que le modèle n'a pas encore commencé à apprendre efficacement.
- À partir de l'époque 4, on observe une amélioration significative des scores, en parallèle avec une diminution progressive de la perte (*loss*).
- Les meilleurs scores BLEU et BERTScore sont atteints à l'époque 9 (BLEU = 7,6931, BERTScore F1 = 0,8179), tandis que les scores METEOR et ROUGE-L F1 culminent à l'époque 10.

Ce résultat indique que le modèle à l'époque 9 offre le meilleur compromis entre la fidélité lexicale (BLEU = 7,6931), la cohérence grammaticale (ROUGE-L) et la similarité sémantique (BERTScore F1 = 0,8179), tout en maintenant des scores élevés sur les métriques METEOR et ROUGE-L.

Par conséquent, le modèle entraîné à l'époque 9 a été sélectionné comme le **meilleur modèle**.

Évolution de la perte d'entraînement et de validation au cours des époques

4.4.2 Évolution de la perte d'entraînement et de validation au cours des époques

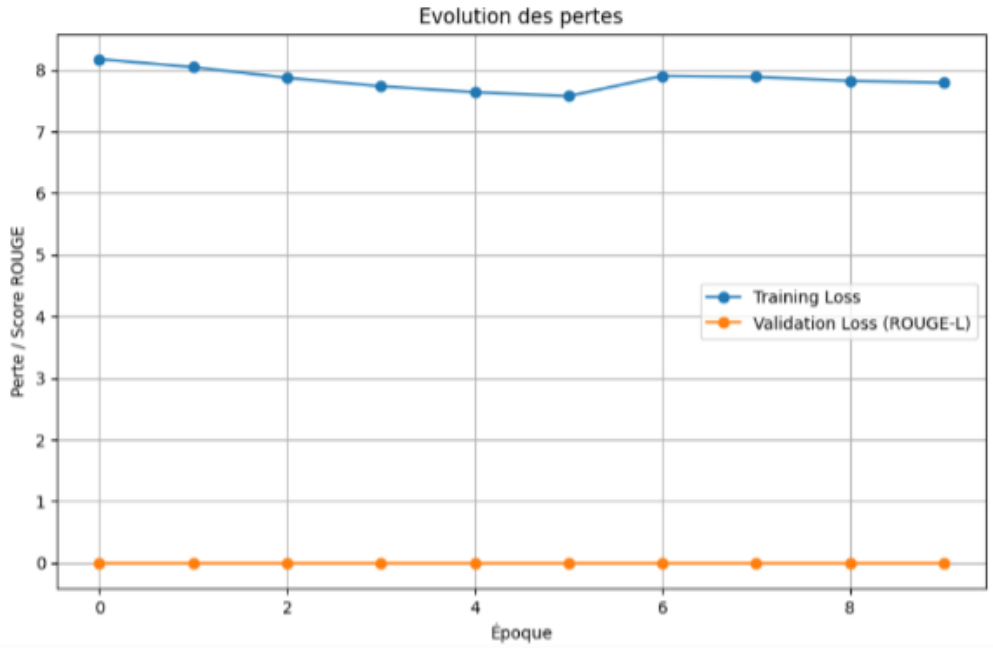

Fig. 4.3 : Évolution de la perte d'entraînement et de validation au cours des époques

La perte passe de **8,35** à **5,58**, ce qui indique que le modèle apprend efficacement à partir des données d'entraînement.

Parallèlement, la perte de validation suit une tendance similaire à la baisse, ce qui signifie que le modèle ne se contente pas de mémoriser les données (*pas d'overfitting*), mais généralise bien sur des exemples non vus.

4.5 Conclusion

Dans ce chapitre, nous avons implémenté et évalué un modèle Transformer pour la traduction automatique. Les résultats expérimentaux ont montré une amélioration progressive des performances au fil des époques, avec une bonne généralisation et une convergence stable. Le modèle entraîné à l'époque 9 a été retenu comme le plus performant, confirmant l'efficacité de l'architecture Transformer pour la traduction entre l'anglais et la darija algérienne.

Conclusion Générale

Au cours des dernières années, la croissance fulgurante des contenus multilingues sur internet, notamment sur les réseaux sociaux, a mis en évidence le besoin pressant de systèmes de traduction automatique performants, capables de comprendre et de produire du texte dans différentes langues, y compris les langues peu représentées comme la darija algérienne.

Dans ce contexte, notre mémoire s'inscrit dans la problématique de la traduction automatique, avec un focus particulier sur la traduction de l'anglais vers la darija algérienne, une tâche encore peu explorée. Pour cela, nous avons eu recours à des techniques de pointe en apprentissage profond, notamment le modèle Transformer, reconnu pour sa capacité à capturer les dépendances contextuelles grâce au mécanisme d'attention.

Nous avons utilisé un corpus de 2400 paires de phrases alignées (anglais darija), récupéré depuis Kaggle au format JSON. Les données ont été réparties en 70% pour l'entraînement et 30% pour la validation/test. Après avoir appliqué un prétraitement adéquat, nous avons entraîné un modèle Transformer et évalué ses performances à l'aide de métriques standards comme BLEU, METEOR, ROUGE-L et BERTScore.

Les résultats expérimentaux ont montré une amélioration progressive au fil des époques, avec un pic de performance à la neuvième époque. Ces résultats soulignent l'efficacité du modèle Transformer même sur un corpus de taille modeste, tout en montrant le potentiel du deep learning dans la traduction vers des dialectes.

Enfin, ce travail ouvre des perspectives intéressantes pour l'avenir, notamment l'utilisation de bases de données plus larges, l'intégration de modèles multilingues pré-entraînés, et l'extension de la tâche de traduction vers d'autres variantes dialectales ou contextes d'usage plus complexes.

Bibliography

Almaaytah, S. A., & Alzobidy, S. A. (2023). Challenges in rendering Arabic text to English using machine translation : a systematic literature review. *IEEE Access*, *11*, 94772-94779.

Alqudsi, A., Omar, N., & Shaker, K. (2014). Arabic machine translation : a survey. *Artificial Intelligence Review*, *42*, 549-572.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT : Transformer-based Model for Arabic Language Understanding. *arXiv preprint arXiv :2003.00104*. https://arxiv.org/abs/2003.00104

Bai, X., & Yu, S. (2014). Rule-Based Machine Translation. In *Routledge Encyclopedia of Translation Technology*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003168348-11/rule-based-machine-translation-xiaojing-bai-shiwen-yu

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python : Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc. https://www.nltk.org/book/

Boughaba, M., Boukhris, B., & Meflah, M. S. (2017). L'apprentissage profond (Deep Learning) pour la classification et la recherche d'images par le contenu. *Revue de Statistique Appliquée et d'Informatique*, *14*(1), 1-12.

Bouziane, A., Attia, M., et al. (2021). DZiriBERT : A pre-trained Language Model for the Algerian Dialect. *arXiv preprint arXiv :2109.12346*. https://arxiv.org/abs/2109.12346

Bray, T. (2014). The JavaScript Object Notation (JSON) Data Interchange Format [Accessed : 2025-06-07]. *RFC 7159*. https://www.rfc-editor.org/rfc/rfc7159

Brour, M., & Benabbou, A. (2019). ATLASLang MTS 1: Arabic text language into Arabic sign language machine translation system. *Procedia computer science*, *148*, 236-245.

DE SONS, D. E. (2016). *Ecole Doctorale Sciences, Technologies et Santé* [thèse de doct., Muséum National d'Histoire Naturelle].

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv :1810.04805*. https://arxiv.org/abs/1810.04805

Dilepix. (2023). *Quelle est la différence entre Intelligence Artificielle, Deep Learning et Machine Learning en agriculture ?* Récupérée juin 13, 2025, à partir de https://www.dilepix.com/blog/difference-intelligence-artificielle-deep-learning-agriculture

Firat, O., Chen, S., Xu, J., & et al.. (2022). No Language Left Behind : Scaling Human-Centered Machine Translation. *arXiv preprint arXiv :2207.04672*. https://arxiv.org/abs/2207.04672

Gentle, J. E. (2009). *Computational Statistics* [Accessed : 2025-06-07]. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-98144-4

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Hunter, J. D. (2007). Matplotlib : A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Junczys-Dowmunt, M. (2018). Marian : Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, 116-121. https://aclanthology.org/P18-4020/

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing (3rd ed. draft)* [Consulté le 6 juin 2025]. https://web.stanford.edu/~jurafsky/slp3/

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing : Players, Agents and Agendas*, 87-90.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press. https://epdf.pub/statistical-machine-translation.html

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177-180. https://aclanthology.org/P07-2045

Lakhfif, A., & Laskri, M. T. (2017). L'analyse et l'annotation à base de FrameNet : contribution à l'étude contrastive des événements de mouvement en arabe et en anglais. *Traitement Automatique des Langues*, *58*(3), 67-96. https://aclanthology.org/2017.tal-3.0.pdf

Lin, C.-Y. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, 74-81. https://aclanthology.org/W04-1013

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *arXiv preprint arXiv :2001.08210*. https://arxiv.org/abs/2001.08210

M., S. (2025). *Qu'est-ce que la Traduction Automatique* [Consulté le 5 juin 2025]. https://www.bureauworks.com/fr/blog/qu-est-ce-que-la-traduction-automatique

Olah, C. (2015). Understanding LSTM Networks [Accessed : 2025-06-06]. *colah's blog*. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch : An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*. https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Rezeg, K., & Laskri, M. (2014). Une Approche connexionniste pour la traduction automatique des textes arabe en français. *Courrier du Savoir scientifique et technique*, *8*(8), 59-67.

Sadoon, T. A., & Ali, M. H. (2021a). Deep learning model for glioma, meningioma and pituitary classification. *Int. J. Adv. Appl. Sci. ISSN*, *2252*(8814), 8814.

Sadoon, T. A., & Ali, M. H. (2021b). Deep learning model for glioma, meningioma and pituitary classification. *Int. J. Adv. Appl. Sci. ISSN*, *2252*(8814), 8814.

Salah, M. H., Vial, L., Blanchon, H., Zrigui, M., & Schwab, D. (2018). Traduction automatique de corpus en anglais annotés en sens pour la désambigüisation lexicale d'une langue moins bien dotée, l'exemple de l'arabe (Automatic Translation of English Sense Annotated Corpora for Word Sense Disambiguation of a Less Well-endowed Language, the Example of Arabic). *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, 329-336.

Salih, A. A., & Omar, L. I. (2023). Reflective glimpses of culture in EFL online classes during COVID-19 pandemic in Oman. *Sustainability*, *15*(13), 9889.

Sanscrit. (2023). *Les trois principaux types de traduction* [Consulté le 5 juin 2025]. https://www.sanscrit.net/fr/types-traduction-professionnelle-post-edition-automatique/

Song, X., Cheng, Y., & Zeng, T. (2024). Using Cross-Cultural Machine Translation Technology to Promote Communication and Cooperation in English Courses [Consulté le 5 juin 2025]. *Journal of Electrical Systems*, *20*(6s). https://journal.esrgroups.org/jes/article/view/2665

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/1409.3215

Team, P. B. (2023). *Translation Technology : Definition, Types, and Examples* [Consulté le 5 juin 2025]. https://phrase.com/blog/posts/translation-technology/

Tiedemann, J. (2017). EC SYSTRAN : The Commission's Machine Translation System. *Traitement Automatique des Langues*, *58*(3), 111-130. https://aclanthology.org/2017.tal-3.0

Varagh. (2020). *Attention is All You Need – Notes* [Consulté en juin 2025]. Récupérée juin 9, 2025, à partir de https://medium.com/@varagh2/attention-is-all-you-need-notes-8c418ab7570c

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017a). Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, *30*, 5998-6008. https://arxiv.org/abs/1706.03762

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017b). Attention is all you need. *Advances in neural information processing systems*, *30*. https://arxiv.org/abs/1706.03762

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's Neural Machine Translation System : Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv :1609.08144*. https://arxiv.org/abs/1609.08144

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic machine translation : A survey with challenges and future directions. *IEEE Access*, *9*, 161445-161468.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O., & Callison-Burch, C. (2012). Machine translation of Arabic dialects. *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics : Human language technologies*, 49-59.