Rapport de projet

Évaluation du moteur de traduction neuronale *OpenNMT* sur des corpus parallèles anglais-français

Dans le cadre du cours

Traduction Automatique et Assistée

Présenté par

Keming YI

Jing LIU

Lidan ZHANG

Encadré par

M. Nasredine SEMMAR

Sommaire

I. Introduction

La traduction automatique neuronale est une technologie basée sur les réseaux de neurones artificiels. Elle a fait des progrès importants ces dernières années grâce à l'intelligence artificielle, par rapport à d'autres formes de traduction automatique, permettant des traductions plus précises et contextuellement appropriées par rapport aux approches traditionnelles basées sur des règles ou des statistiques (Barbin, 2020). Elle s'appuie sur des modèles de deep learning, en particulier des réseaux de neurones récurrents (RNN) et des transformateurs, pour comprendre et générer du texte de manière plus fluide et contextuellement appropriée.

Notre projet se concentre donc sur l'évaluation du moteur de traduction neuronale OpenNMT, en utilisant des corpus parallèles anglais-français tirés d'Europarl et d'EMEA. L'objectif est de démontrer notre capacité à installer, entraîner des modèles de traduction à l'aide de OpenNMT, et à les évaluer selon les résultats obtenus.

II. Présentation OpenNMT

OpenNMT est un outil de traduction automatique neuronale open-source, développé initialement par le groupe TAL d'Havard et SYSTRAN. The OpenNMT initiative consists of several projects to assist researchers and developers in their NMT journey, from data preparation to inference acceleration. It supports a wide range of model architectures and training procedures for neural machine translation as well as related tasks such as natural language generation and language modeling.(Klein G, Hernandez F, Nguyen V, et al, 2020) OpenNMT se distingue par sa capacité à utiliser des architectures avancées de réseaux de neurones, notamment les réseaux de neurones récurrents (RNN) et les transformateurs.

OpenNMT a utilisé des architectures basées sur des RNN, telles que les LSTM (Long Short-Term Memory) et les GRU (Gated Recurrent Units), pour modéliser les séquences de texte. Ces architectures sont capables de capturer les dépendances temporelles dans les données séquentielles, ce qui est crucial pour les tâches de traduction. Les LSTM sont conçus pour résoudre le problème de la disparition du gradient dans les RNN classiques, ce qui les rend inefficaces pour apprendre des dépendances à long terme. Les LSTM introduisent des mécanismes de régulation du flux d'information à travers la cellule LSTM, en utilisant des portes. (Hochreiter et Schmidhuber, 1997). Les GRU sont une variante simplifiée des LSTM,

introduite pour combiner les avantages des LSTM tout en réduisant la complexité computationnelle. Les GRU utilisent moins de portes et n'ont pas d'état de cellule séparé.(Chung et al., 2014)

Contrairement aux RNN et aux LSTM, les transformateurs n'utilisent pas de structure récurrente. Ils reposent plutôt sur des mécanismes d'attention pour traiter les données séquentielles en parallèle. L'entrée est transformée en vecteurs d'embedding et combinée avec un codage positionnel. Ces vecteurs passent par plusieurs couches d'encodage où chaque couche applique un mécanisme d'attention multi-tête et un réseau de neurones feedforward. Le décodeur utilise les représentations générées par l'encodeur et, avec ses propres mécanismes d'attention, génère la séquence de sortie un élément à la fois.(Vaswani A, Shazeer N, Parmar N, et al, 2017)

III. Évaluation OpenNMT sur un corpus en formes fléchies

- Le corpus d'apprentissage et d'évaluation

Nous avons récupéré nos corpus parallèles anglais-français depuis Europarl et EMEA. Ce sont tous les deux des corpus parallèles, Europarl représente un domaine général avec une grande diversité de sujets, tandis qu'Emea offre un contexte spécialisé dans la médecine.

Voici un tableau qui démontre le nombre de phrases que nous avons extraits de chaque corpus :

	Apprentissage	**Développement**	**Test**
Europarl	100k phrases	3750 phrases	500 phrases
Emea	10k phrases	/	500 phrases

- Les métriques d'évaluation: Score BLEU

Le Score BLEU est une métrique standard utilisée pour évaluer la qualité des textes générés automatiquement. Il compare les phrases consécutives de la traduction automatique avec les phrases consécutives qu'il trouve dans la traduction de référence, et il compte le nombre de correspondances de manière pondérée. Variant de 0 à 1, un degré de concordance

plus élevé indique un degré de similitude plus élevé avec la traduction de référence, donc un score plus élevé (laujan, 2023).

Voici le résultat d'évaluation que nous avons obtenu après le nettoyage des corpus (tokenisation, processus de true-case etc.):

N° du run	**Apprentissage (nombre de phrases)**	**Tuning (nombre de phrases)**	**Test (nombre de phrases)**	**Score Bleu**
1	100K (Europarl)	3,75K (Europarl)	500(Europarl)	**0.07**
2	100K+10K (Europarl+Emea)	3,75K (Europarl)	500(Europarl)+ 500(Emea)	**0.13**

Selon le score BLEU que nous avons obtenu de notre corpus après le nettoyage, nous constatons que le score a augmenté de 0.07 à 0.13, ce qui suggère une très faible similarité entre les traductions générées par le modèle et les traductions de référence. Néanmoins, cette augmentation de score suggère une amélioration de notre modèle de traduction entraîné. Nous supposons que ce changement est dû à l'augmentation de la taille du corpus entraîné. Cela souligne l'importance d'un corpus d'entraînement diversifié et suffisamment large pour améliorer la performance des modèles de traduction automatique.

IV. Évaluation OpenNMT sur un corpus en lemmes

- Le lemmatiseur NLTK pour l'anglais

NLTK est une bibliothèque puissante pour le traitement du langage naturel et est largement utilisée pour la manipulation de textes en anglais. Le script qu'on écrit pour lemmatiser le corpus en anglais commence par importer les modules nécessaires de NLTK, notamment pour la tokenisation, l'étiquetage grammatical (POS tagging), et la lemmatisation précise via le 'WordNetLemmatizer'. Cela permet de normaliser le texte, facilitant des analyses linguistiques plus poussées en réduisant la diversité morphologique des mots tout en préservant leur sens fondamental.

Le processus de lemmatisation de NLTK repose sur la base de données WordNet, un vaste dictionnaire lexical anglais qui inclut des définitions, des synonymes, des antonymes, etc. Chaque mot dans WordNet se voit attribuer une ou plusieurs catégories grammaticales et est lié à d'autres mots lexicalement apparentés. Avant de pouvoir procéder à la lemmatisation, il est nécessaire de marquer les mots avec leur catégorie grammaticale. NLTK utilise la méthode ‘pos_tag’ pour taguer automatiquement la catégorie grammaticale de chaque mot dans le texte, puis détermine la catégorie grammaticale de WordNet la plus appropriée basée sur les tags Treebank. Une fois la catégorie grammaticale correcte de WordNet obtenue, le ’WordNetLemmatizer’ utilise cette catégorie pour ramener le mot à sa forme de base. Par exemple, le verbe "saw" serait lemmatisé en "see" selon le contexte, et le nom "leaves" redeviendrait "leaf".

Selon les résultats d'exécution du script, on peut trouver que NLTK réalise très efficacement la tokenisation, l'étiquetage grammatical et la lemmatisation pour le corpus anglais.

- Le lemmatiseur NLTK pour le français

NLTK est principalement conçu pour traiter l'anglais, donc il dispose de fonctionnalités limitées pour le français et d'autres langues non anglaises. Pour la lemmatisation en français, NLTK peut être utilisé mais avec plusieurs ajustements et limitations.

Pour faire la lemmatisation en français, on a utilisé un lemmatiseur spécifique au français ‘FrenchLefffLemmatizer’, mais selon le résultat qu’on a obtenu, les verbes ne sont pas lemmatisés à leur forme en infinitif. Bien que NLTK propose des outils de tokenisation et d'étiquetage grammatical (POS tagging) qui peuvent être adaptés au français, ces outils ne sont pas aussi robustes ni aussi précis que ceux conçus spécifiquement pour le français, telles que ses règles de conjugaison et de genre, qui sont plus complexes qu'en anglais.

Enfin, on se tourne vers l’outil SpaCy qui est conçu spécialement pour traiter le français.

- Tableau des résultats

Voici le résultat d’évaluation sur les corpus parallèles en lemmes :

N° du run	Apprentissage (nombre de	Tuning (nombre de	Test (nombre de phrases)	Bleu

	phrases)	phrases)		
3	100K (Europarl)	3,75K (Europarl)	500(Europarl)	0.22
4	100K+10K (Europarl+Emea)	3,75K (Europarl)	500(Europarl)+ 500(Emea)	0.18

Entre le run 3 et le run 4, le score BLEU a diminué de 0.22 à 0.18, ce qui indique que l'ajout du jeu de données Emea a entraîné une baisse de la qualité de traduction. Cela pourrait être dû au fait que le jeu de données Emea est un corpus du domaine de spécialité, avec un style linguistique, des termes ou une structure qui diffèrent de ceux des données Europarl, ce qui a entraîné une réduction de la capacité de généralisation du modèle sur le jeu de données de test.

Cependant, par rapport aux run 1 et run 2, les deux valeurs BLEU ont connu une amélioration, en particulier pour la traduction des données Europarl après la lemmatisation, où la qualité de traduction a relativement augmenté. Cela indique que OpenNMT est plus performant dans la traduction des textes lemmatisés.

V. Points forts, limitations et difficultés rencontrées

i. Difficultés rencontrées

a. L'entraînement du modèle de traduction

Pour surmonter la lenteur de l'entraînement des modèles de traduction neuronale sur des machines locales, nous avons utilisé l'abonnement Pro de Google Colab, qui offre un accès à des GPU puissants comme le NVIDIA L4 GPU. En configurant notre environnement sur Google Colab, nous avons pu réduire considérablement le temps d'entraînement de notre modèle OpenNMT, passant de plusieurs heures à seulement 2 minutes par session. Cette utilisation de Google Colab a pu largement accélérer notre workflow afin d'obtenir des résultats rapides.

b. La lemmatisation du corpus en français

On a utilisé d'abord la bibliothèque NLTK pour la tokenisation et le POS-Tagging du corpus en français et un outil spécifique au français 'FrenchLefffLemmatizer' pour la lemmatisation. Mais les mots dans le résultat n'est pas en forme de lemme, bien qu'on modifiait le script en plusieurs fois, on trouvait toujours des verbes en conjugaison dans le texte obtenu. Enfin, on a utilisé le SpaCy pour la tokenisation du corpus en français.

ii. Points forts

En tant qu'outil open-source, OpenNMT est accessible à tout le monde. Il est bien documenté et bénéficie d'une communauté active. Sa documentation complète, incluant des guides détaillés et des tutoriels, facilite grandement sa prise en main. La simplicité d'utilisation d'OpenNMT est également remarquable : avec des fichiers de configuration YAML et des données prétraitées, les utilisateurs peuvent facilement définir les paramètres d'entraînement et laisser l'outil gérer le processus de bout en bout. De plus, OpenNMT supporte un large éventail de langues, ce qui le rend encore plus utile pour des projets multilingues.

Cependant, dans le cadre de notre projet, nous n'avons pas eu l'opportunité de comparer OpenNMT avec d'autres outils de traduction automatique, ce qui limite notre capacité à évaluer pleinement ses avantages relatifs. De plus, les résultats de nos traductions n'ont pas été à la hauteur de nos attentes, avec des scores BLEU relativement bas indiquant des performances médiocres.

iii. Limitations

L'entraînement de modèles de traduction automatique de haute qualité avec OpenNMT peut nécessiter des ressources computationnelles significatives, notamment des GPU puissants et une grande quantité de mémoire. Nous avons dû payer Google Colab pro afin d'entraîner nos modèles d'économiser du temps.

Comme tous les outils de traduction automatique neuronale, il est dépendant de la quantité des données d'entraînement. Si les données sont biaisées, cela peut affecter la performance du modèle. Les scores BLEU obtenus très bas nous empêchent de déterminer clairement les points forts d'OpenNMT dans un contexte comparatif. Les résultats décevants peuvent être attribués à plusieurs facteurs, notamment la qualité et la taille des corpus d'entraînement utilisés, ainsi que les configurations spécifiques des modèles. Sans une analyse comparative approfondie et des ajustements optimisés des hyperparamètres, il est

difficile de tirer des conclusions définitives sur l'efficacité d'OpenNMT par rapport à d'autres solutions disponibles sur le marché.

VI. Organisation de tâches

Nous avons pu répartir assez équitablement les tâches pour notre projet. Chacun a pu contribuer sa partie de code et de données à notre repository sur Github.

Keming : Récupération des corpus en ligne et la séparation des corpus. Tentation d'utilisation d'OpenNMT mais a échoué à installer à cause du confit de dépendance. Rédaction de la partie 1, 2 et 5 du rapport.

Jing : La partie de la lemmatisation des corpus en français et en anglais pour remplacer dans les corpus parallèles par leur lemme. Et la rédaction de la partie 4 du rapport.

Lidan : Test de fonctionnement de OpenNMT, construction de vocabulaire, entraînement de modèle de traduction, traduction de texte des corpus Europarl, Emea avant et après la lemmatisation, rédaction de la partie 3 du rapport.

VII. Annexes

Nous avons utilisé le fichier YAML par défaut du tutoriel disponible sur le site d'OpenNMT. Cependant, en ajustant certains paramètres, nous pourrions obtenir de meilleurs résultats. De plus, nous avons rencontré des difficultés lors de l'étape de lemmatisation des corpus. Si nous raffinons la lemmatization, le modèle sera plus performant.

Voici les résultats de traduction générés par nos différents modèles (par ordre : Europarl après nettoyage, Europarl et Emea après nettoyage, Europarl après nettoyage et lemmatisation, Europarl et Emea après lemmatisation).

```
1   en ce qui concerne la Commission, nous devons tous la proposition.
2   en ce qui concerne les États membres, les États membres doivent être <unk>.
3   je crois que nous devons faire que nous nous le savons.
4   en ce qui concerne les États membres, nous devons tous la proposition.
5   en ce qui concerne les États membres, l'Union européenne n'a pas encore été
6   Monsieur le Président, Monsieur le Commissaire, l'Union européenne, nous devons tous la proposition.
7   en ce qui concerne les États membres, nous devons tous que nous nous le savons.
8   en ce qui concerne les États membres, l'Union européenne n'a pas été <unk>.
9   j'invite que j'ai déjà dit à la Commission.
10  j'invite que la Commission n'a pas le fait.
```

```
Monsieur le Président, la Commission n'a pas fait à la question de la Commission.
la Commission n'a pas la question de la Commission en matière de la Commission en matière de la Commission e
je pense que la Commission n'a pas fait à la question de la Commission en matière de la Commission en matièi
en ce qui concerne la Commission en matière de la Commission en matière de la Commission en matière de la Cc
la Commission n'a pas la question de la Commission.
la Commission n'a pas la question de la Commission en matière de la Commission en matière de la Commission e
la Commission n'a pas la question de la Commission en matière de la Commission en matière de la Commission e
en ce qui concerne la Commission en matière de la Commission en matière de la Commission en matière de la Cc
je pense que la Commission n'a pas fait à la question de la Commission.
je pense que la Commission n'a pas fait à la question de la Commission en matière de la Commission en matièi
```

```
Monsieur le commissaire , il ne s&apos ; agir pas d&apos ; agir de l&apos ; union européen .
le rapport de l&apos ; union européen , c&apos ; est le fait qu&apos ; il ne s&apos ; agir pas de l&apos ; c
j&apos ; espérer que le rapport de l&apos ; union européen , mai il ne falloir pas que le rapport de l&apos
dans le cadre de l&apos ; UE , il ne s&apos ; agir pas d&apos ; agir de l&apos ; union de l&apos ; union .
l&apos ; union européen , c&apos ; est le fait qu&apos ; il ne s&apos ;
Monsieur le commissaire , Monsieur le commissaire , Monsieur le commissaire , Monsieur le commissaire , Mons
le rapport de l&apos ; union européen , nous ne devoir pas que le rapport de l&apos ; union de l&apos ; unic
le rapport de l&apos ; union européen , c&apos ; est le fait qu&apos ; il ne s&apos ; agir pas de l&apos ; c
j&apos ; espérer que le commissaire , Monsieur le commissaire , Monsieur le commissaire , Monsieur le commis
j&apos ; espérer que le rapport de l&apos ; union européen de l&apos ; union de l&apos ; union européen de
```

```
en effet , la commission ne pouvoir pas faire la commission à la commission .
la commission n&apos ; avoir pas que la commission ne pouvoir pas faire la commission à l&apos ; avenir de
je penser qu&apos ; il n&apos ; y avoir pas d&apos ; abord que la commission ne pouvoir pas faire la commis
en effet , la commission ne pouvoir pas faire la commission à la commission , en tant que la commission ne p
la commission ne pouvoir pas se faire .
la commission n&apos ; avoir pas que la commission ne pouvoir pas donner la commission à l&apos ; avenir de
en effet , la commission devoir donc faire la commission à l&apos ; avenir de la commission en uvre de la cc
en effet , la commission ne pouvoir pas faire la commission à l&apos ; objet de la commission , la commissic
je penser qu&apos ; il n&apos ; y avoir pas d&apos ; abord que la commission ne pouvoir pas faire la commis
je penser qu&apos ; il n&apos ; y avoir pas d&apos ; abord que la commission ne pouvoir pas faire la commis
```

VIII. Références

1. Europarl Corpus: https://opus.nlpl.eu/Europarl.php
2. EMEA Corpus: : https://opus.nlpl.eu/EMEA.php
3. laujan, Ericurban. Qu'est-ce qu'un Score BLEU? 20/07/2023
4. Klein G, Hernandez F, Nguyen V, et al. The OpenNMT neural machine translation toolkit: 2020 edition[C]//Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track). 2020: 102-109.
5. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
6. Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
7. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

8. Franck Barbin. La traduction automatique neuronale, un nouveau tournant ?. Palimpseste. Sciences, humanités, sociétés , 2020, 4, pp.51-53. ffhalshs-03603588f