Développement et Évaluation d'un Système de Traduction Automatique Neuronale Anglais–Russe

Dominique S. Loyer
Travail présenté le 28 avril 2025
à Ngoc Tan Lê, dans le cadre d'un
cours de doctorat, TALN DIC-9345, à l'UQAM
loyer.dominique@courrier.uqam.ca

Abstract

Ce rapport présente le développement et l'évaluation d'un système de traduction automatique neuronale (TAN) Anglais→Russe (EN→RU). En utilisant la bibliothèque Hugging Face Transformers(Wolf et al., 2020) et un modèle pré-entraîné `Helsinki-NLP/opus-mt-en-ru`(Junczys-Dowmunt and Grundkiewicz, 2018) basé sur l'architecture Transformer(Vaswani et al., 2017), nous avons fine-tuné le modèle sur un sous-ensemble de 10 000 paires de phrases du corpus parallèle `opus_books`(Tiedemann, 2012). Le système a été entraîné pendant 3 époques dans un environnement Kaggle (GPU T4). L'évaluation sur un jeu de test réservé du même corpus donne SacreBLEU 22.6 et chrF 48.5.

Mots-clés : Traduction Automatique Neuronale, EN→RU, Transformer, Fine-tuning, SacreBLEU, chrF.

1 Introduction

La traduction automatique (TA) vise à traduire du texte d'une langue source vers une langue cible sans intervention humaine. Historiquement dominée par des approches statistiques (SMT), la TA a été révolutionnée par la traduction automatique neuronale (NMT) au milieu des années 2010(Bahdanau et al., 2014). Ce projet :

implémente un pipeline complet (données, fine-tuning, évaluation), utilise le modèle `Helsinki-NLP/opus-mt-en-ru`(Junczys-Dowmunt and Grundkiewicz, 2018), s'appuie sur 10 000 paires de `opus_books`(Tiedemann, 2012).

Nous mesurons la qualité via SacreBLEU et chrF et comparons à l'état de l'art.

2 État de l'art

Les premières architectures NMT combinaient RNN et attention(Bahdanau et al., 2014). Le Transformer(Vaswani et al., 2017) a ensuite imposé son efficacité. Les modèles MarianMT(Junczys-Dowmunt and Grundkiewicz, 2018), pré-entraînés sur OPUS, offrent une base solide. Les campagnes WMT et le modèle NLLB-200(Team, 2022) atteignent BLEU 30–45 pour EN→RU.

3 Méthodologie

3.1 Pipeline général

Le processus :

1. **Chargement :** Corpus `opus_books`(Tiedemann, 2012) EN–RU via `datasets`(Lhoest et al., 2021).
2. **Prétraitement :** Nettoyage et tokenisation (max len=128 tokens).
3. **Modèle :** Chargement de `Helsinki-NLP/opus-mt-en-ru`(Junczys-Dowmunt and Grundkiewicz, 2018).
4. **Fine-tuning :** 3 époques, batch=8, lr=2e-5, fp16 via `Seq2SeqTrainer`.
5. **Évaluation :** SacreBLEU, chrF avec `evaluate`(Lhoest et al., 2021).

3.2 Analyse de la méthodologie

Le pipeline adopté suit une approche standard pour le développement d'un système de traduction automatique neuronale basée sur le fine-tuning. Le choix du corpus `opus_books` est pertinent car il offre un large éventail de textes traduits, bien que le sous-ensemble de 10 000 paires utilisé pour l'entraînement représente une fraction modeste de la taille totale du corpus (17 496 paires selon la section Données, mais potentiellement plus vaste si on considère l'ensemble du corpus OPUS). L'utilisation du modèle pré-entraîné `Helsinki-NLP/opus-mt-en-ru` est judicieuse car il bénéficie déjà d'une connaissance linguistique acquise sur de plus grands volumes de données, ce qui permet un apprentissage plus rapide et potentiellement de meilleurs résultats avec un jeu de données limité.

Le prétraitement, limité au nettoyage et à la tokenisation avec une longueur maximale de 128 tokens, est une étape essentielle pour garantir que les données sont dans un format approprié pour le modèle. La longueur maximale choisie peut influencer la capacité du modèle à traduire des phrases plus longues dans leur intégralité.

Les hyperparamètres de fine-tuning (3 époques, taille de batch de 8, taux d'apprentissage de 2e-5) représentent une configuration typique pour ce type de tâche. L'utilisation de la précision mixte (fp16) via `Seq2SeqTrainer` permet d'accélérer l'entraînement et de réduire la consommation de mémoire sur le GPU.

L'évaluation à l'aide de SacreBLEU et chrF fournit des métriques standardisées et largement reconnues pour évaluer la qualité des systèmes de traduction automatique. SacreBLEU, en particulier, est une métrique robuste qui prend en compte plusieurs facteurs pour éviter les biais d'évaluation.

3.3 Environnement

L'utilisation de l'écosystème Hugging Face Transformers simplifie considérablement le processus de développement en fournissant des outils et des modèles pré-implémentés. L'exécution sur un Notebook Kaggle avec un GPU Tesla T4 offre une puissance de calcul adéquate pour le fine-tuning du modèle dans un délai raisonnable.

4 Données

Le corpus `opus_books`(Tiedemann, 2012) est un choix pertinent pour ce projet compte tenu de sa disponibilité et de sa nature parallèle. La division en ensembles d'entraînement, de validation et de test est cruciale pour évaluer la capacité de généralisation du modèle et éviter le surapprentissage. La taille de l'ensemble d'entraînement (10 000 paires) est relativement petite par rapport aux standards actuels pour les tâches de traduction automatique à grande échelle.

5 Résultats

Les scores obtenus sur l'ensemble de test (SacreBLEU 22.6 et chrF 48.5) fournissent une mesure quantitative de la performance du système.

5.1 Analyse des résultats et limitations

Le SacreBLEU de 22.6 indique une qualité de traduction modérée. Bien que ce score ne soit pas comparable aux résultats de pointe rapportés pour EN→RU (BLEU 30–45 voire 45.3 pour WMT19 Best), il est important de considérer la quantité limitée de données d'entraînement utilisée (10 000 paires). Les modèles plus performants mentionnés dans l'état de l'art ont été entraînés sur des corpus beaucoup plus vastes.

Le score chrF de 48.5, qui prend en compte la similarité au niveau des caractères, est également un indicateur de la qualité de la traduction. Comparer ce score directement avec les systèmes de l'état de l'art qui rapportent principalement le BLEU est difficile sans les valeurs chrF correspondantes.

Limitations :

— **Taille des données d'entraînement :** L'utilisation de seulement 10 000 paires de phrases est une limitation majeure. Un corpus plus grand permettrait potentiellement au modèle d'apprendre des traductions plus complexes et d'améliorer la fluidité et l'exactitude.

— **Complexité du corpus :** Le corpus `opus_books` peut contenir une variété de styles et de complexités de langage. Il serait intéressant d'analyser les performances du modèle sur différents sous-ensembles du corpus (par exemple, en fonction du genre littéraire) pour identifier des points forts et des faiblesses.

— **Longueur maximale des séquences :** La limite de 128 tokens pourrait avoir un impact sur la qualité des traductions de phrases plus longues, qui pourraient être tronquées ou traduites de manière incomplète.

— **Nombre d'époques d'entraînement :** Seulement 3 époques d'entraînement pourraient être insuffisantes pour que le modèle converge complètement. Un entraînement plus long avec une stratégie d'arrêt précoce basée sur les performances sur l'ensemble de validation pourrait être bénéfique.

— **Absence d'analyse qualitative :** Le rapport se concentre sur les métriques quantitatives. Une analyse qualitative des traductions produites par le modèle (en examinant des exemples de traductions correctes et incorrectes) fournirait des informations plus approfondies sur les types d'erreurs commises par le système.

6 Perspective et travaux futurs

Bien que les résultats obtenus soient encourageants compte tenu de la quantité limitée de données, plusieurs pistes d'amélioration peuvent être explorées.

6.1 Augmentation des données d'entraînement

Comme suggéré dans la conclusion, l'augmentation de la taille du corpus d'entraînement est une voie évidente pour améliorer les performances du modèle. L'utilisation de corpus plus importants tels que WMT(Bojar) et ParaCrawl(Barrault et al., 2019) pourrait significativement augmenter la couverture linguistique et la robustesse du modèle. L'exploration de techniques de données augmentées spécifiques à la traduction automatique, comme le back-translation(Sennrich et al., 2016), où un modèle est utilisé pour traduire le corpus cible vers la langue source afin de créer des données d'entraînement supplémentaires, pourrait également être bénéfique.

6.2 Exploration de modèles plus avancés

Le modèle `Helsinki-NLP/opus-mt-en-ru` est un bon point de départ, mais l'expérimentation avec des modèles plus grands et plus récents, tels que NLLB-200(Team, 2022) (mentionné dans l'état de l'art) ou des modèles de la famille Transformer encore plus performants, pourrait conduire à des améliorations substantielles. L'utilisation de grands modèles de langage (LLM) pré-entraînés sur de vastes quantités de texte multilingue et leur fine-tuning pour la tâche de traduction pourrait également être explorée. Des modèles comme mBART(Liu et al., 2020), T5(Raffel et al., 2020), ou même des modèles plus récents comme Gemma(Team, 2024) pourraient offrir de meilleures performances.

6.3 Optimisation des hyperparamètres et de la stratégie d'entraînement

Une recherche plus approfondie des hyperparamètres de fine-tuning (taux d'apprentissage, taille de batch, nombre d'époques, fonctions d'activation, etc.) pourrait permettre d'optimiser les performances du modèle sur le corpus actuel et sur des corpus plus grands. L'implémentation d'une stratégie d'arrêt précoce basée sur la performance sur l'ensemble de validation aiderait à prévenir le surapprentissage et à optimiser le nombre d'époques d'entraînement.

6.4 Amélioration du prétraitement

L'exploration de techniques de prétraitement plus avancées, telles que l'application de la tokenisation SentencePiece(Kudo and Richardson, 2018) ou WordPiece(Wu et al., 2016) au lieu d'une simple tokenisation par espace, pourrait améliorer la gestion du vocabulaire et des mots rares. L'augmentation de la longueur maximale des séquences pourrait permettre au modèle de traiter des phrases plus longues.

6.5 Analyse qualitative des erreurs

Une analyse qualitative des erreurs de traduction produites par le modèle actuel permettrait d'identifier les types d'erreurs les plus fréquentes (par exemple, erreurs lexicales, erreurs syntaxiques, problèmes de cohérence, erreurs de traduction de termes spécifiques). Cette analyse pourrait ensuite guider les efforts futurs en ciblant les aspects spécifiques où le modèle doit être amélioré.

6.6 Évaluation plus poussée

Outre SacreBLEU et chrF, d'autres métriques d'évaluation, telles que TER (Translation Edit Rate) ou des évaluations basées sur le jugement humain, pourraient fournir une image plus complète de la qualité des traductions. L'évaluation du modèle sur des domaines spécifiques ou des types de textes différents (par exemple, textes techniques, conversations) pourrait également révéler des forces et des faiblesses spécifiques.

Références

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Loïc Barrault, Thomas Etchegoyhen, Pascale Bougares, Benjamin Cartoni, Laurent Maillard, Julien Carré, Samuel Bordes, Carolina Scarton, Xavier Tannier, Laurent Guérin, and Marek Šimko. 2019. https ://doi.org/10.18653/v1/E19-1099 ParaCrawl : Large Scale Extracted Parallel Corpora. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar. Findings of the 2016 conference on machine translation (WMT16).

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. Marian : Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Taku Kudo and John Richardson. 2018. Sentencepiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 66–71.

Quentin Lhoest et al. 2021. Datasets : A community library for natural language processing. `https://github.com/huggingface/datasets`.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xuan Li, Smith, , Luke Zettlemoyer, and Veselin Stoyanov. 2020. Multilingual denoising pre-training for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8232–8242.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matthis, Timothy Newlan, Germán Guu, Pangwei Huang, Pooja Prenger, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140) :1–67.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 86–96.

Google Team. 2024. Gemma : Open models based on gemini architecture. `https://ai.google.dev/gemma`.

NLLB Team. 2022. No language left behind : Scaling human-centered machine translation. *arXiv preprint arXiv :2207.04671*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, and et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, and et al. 2020. Transformers : State-of-the-art natural language processing. *Proceedings of EMNLP 2020 : System Demonstrations*, pages 38–45.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Navdeep Jaitly, Geoffrey Hinton, , et al. 2016. Google's neural machine translation system : Bridging the gap between human and machine translation.

Table des matières