Traduction Automatique Neuronale (TAN) Anglais-Russe

Projet 2 présenté à Ngoc Tan Lê

Dominique S. Loyer (TALN, DIC-9345)

Université du Québec à Montréal

24 Avril 2025

Plan

Introduction : Traduction Automatique Neuronale (TAN)

- Importance croissante de la traduction automatique (contenu multilingue).
- Révolution de la TAN vs approches précédentes (SMT).
- Architecture dominante : Transformer (Vaswani et al., 2017).

1. Ce modèle est basé sur l'architecture Transformer et implémenté à l'aide du framework MarianMT

Introduction : Traduction Automatique Neuronale (TAN)

- Importance croissante de la traduction automatique (contenu multilingue).
- Révolution de la TAN vs approches précédentes (SMT).
- Architecture dominante : Transformer (Vaswani et al., 2017).

Objectifs du Projet :

1. Développer un système TAN pour Anglais -> Russe.
2. Utiliser des outils open-source (Hugging Face) et données publiques (opus_books).
3. Affiner un modèle pré-entraîné (Helsinki-NLP). [1]
4. Évaluer (BLEU, chrF) et comparer à l'état de l'art.

1. Ce modèle est basé sur l'architecture Transformer et implémenté à l'aide du framework MarianMT

État de l'Art Des Modèles Clés

- Passage SMT -> TAN (RNNs (Bahdanau et al., 2015) -> Transformer (Vaswani et al., 2017)).
- Importance des données parallèles (WMT (Ng et al., 2019), OPUS (Tiedemann, 2012)).
- Modèles pré-entraînés bilingues : MarianMT (Helsinki-NLP) [2] (Junczys-Dowmunt et al., 2018) sur OPUS. Base solide pour affinage.
- Modèles multilingues massifs : NLLB (Costa-Jussà et al., 2022) domine souvent, scores BLEU EN-RU > 40 sur benchmarks WMT (Ng et al., 2019; Costa-Jussà et al., 2022).

2. Helsinki-NLP/opus-mt-en-ru est un modèle pré-entraîné spécifiquement pour la traduction de l'anglais vers le russe

Tendances Récentes et Positionnement

- GML pour la traduction : Certains GMM affinés rivalisent/dépassent TAN dédiés (Xu et al., 2023).
- Campagnes WMT : Utilisation de rétro-traduction [3], assemblage [4], etc. pour l'état de l'art (Molchanov et al., 2023; Barrault et al., 2020).

3. traduire un grand corpus de texte monolingue dans la langue cible vers la langue source à l'aide d'un modèle de TA, créant ainsi des données parallèles synthétiques.
4. combiner les prédictions de plusieurs modèles différents pour obtenir un résultat final plus robuste

Tendances Récentes et Positionnement

- GML pour la traduction : Certains GMM affinés rivalisent/dépassent TAN dédiés (Xu et al., 2023).
- Campagnes WMT : Utilisation de rétro-traduction [3], assemblage [4], etc. pour l'état de l'art (Molchanov et al., 2023; Barrault et al., 2020).
- Notre approche : Affinage (Fine-tuning) pragmatique de MarianMT sur l'ensemble d'opus_books.
- Attentes : Performances inférieures à l'état de l'art mais approche accessible et reproductible.

3. traduire un grand corpus de texte monolingue dans la langue cible vers la langue source à l'aide d'un modèle de TA, créant ainsi des données parallèles synthétiques.
4. combiner les prédictions de plusieurs modèles différents pour obtenir un résultat final plus robuste

Outils et Paramètres Finaux

- Modèle : Helsinki-NLP/opus-mt-en-ru (MarianMT/Transformer) (University of Helsinki, 2020).
- Données : opus_books EN-RU (Tiedemann, 2012).
  - Split Train/Val/Test (16k / 0.8k / 0.9k).
  - Utilisation de l'ensemble des données disponibles.
- Outils : Python, PyTorch, Hugging Face ('transformers', 'datasets', 'evaluate').
- Prétraitement : Tokenisation (MarianTokenizer), séquences max 128.
- Fine-tuning (Final) :
  - 'Seq2SeqTrainer'.
  - 5 époques, batch size 8, lr 2e-5, AdamW, fp16.
  - Environnement : Kaggle GPU T4 et GPU P100 (TPU quelques instants...).
- Évaluation : SacreBLEU , chrF (standard).

Le Pipeline

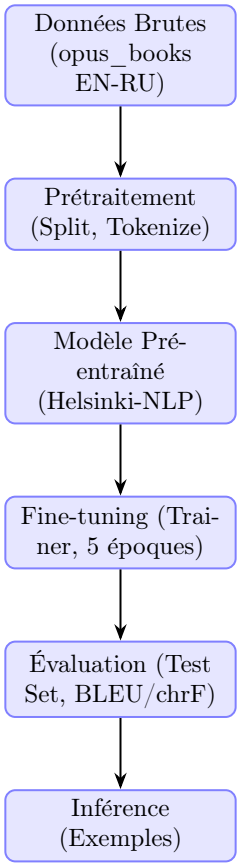

Figure – Schéma simplifié du pipeline de traduction implémenté.

Parcours Itératif et Rationnel

Configuration	Rationnel / Données / Époques	BLEU	chrF
Test Initial (Helsinki)	Validation pipeline rapide / 1k / 1 ép.	0.0	-
Augmenté (Helsinki)	1er résultat significatif / 10k / 3 ép.	22.6	48.5
Tentative NLLB	Test modèle + puissant / full / 5 ép.	22.4*	46.9*
Final (Helsinki)	Maximiser Helsinki / full / 5 ép.	24.6	50.2

- Validation progressive du pipeline.
- Confirmation de l'impact des données et des époques.
- Modèle Helsinki-NLP plus performant/stable que NLLB-600M sur ce jeu de données/configuration.
- (*) Exécution NLLB instable (trop massif pour ma configuration).

Performances Finales (Helsinki-NLP, opus_books complet)

- Entraînement sur 16k exemples, 5 époques (17 minutes sur T4).
- Perte finale (train loss) : 1.5029.

-

- SacreBLEU : 24.6
- chrF : 50.2

Note : Amélioration par rapport aux tests initiaux, mais toujours inférieur à l"état de l'art.

Positionnement par rapport à la littérature (EN->RU)

Système / Réf.	Dataset Eval.	BLEU	chrF
Facebook WMT19 (Ng et al., 2019)	WMT19 newstest	45.3*	-
NLLB-200 (Costa-Jussà et al., 2022)	FLORES-200	39-40**	-
PROMT WMT23 (Molchanov et al., 2023)	WMT23 test	38***	-
Notre Modèle (Final)	opus_books (full)	24.6	50.2

- Scores États de l'ART (WMT, NLLB) souvent $>$ 35-40 BLEU.
- Différences dues à : taille des données, taille modèle, techniques (rétro-traduction...).
- Notre approche = compromis pragmatique ressources/performance sur opus_books.

(* Ng et al. (2019) ; ** Costa-Jussà et al. (2022) ; *** Molchanov et al. (2023))

Exemples de Traduction (Modèle Final)

1

EN : Machine translation is fascinating.
RU (Modèle) : Машинный перевод увлекателен.
(Traduction correcte et fluide)

Exemples de Traduction (Modèle Final)

1

EN : Machine translation is fascinating.
RU (Modèle) : Машинный перевод увлекателен.
(Traduction correcte et fluide)

2

EN : This model was trained on the full dataset for more epochs.
RU (Modèle) : Эта модель была натренирована по всему массиву данных для новых эпох.
(Traduction globalement correcte, légère maladresse ?)

Conclusion et Perspectives

Conclusion :

- Implémentation réussie d'un pipeline TAN EN->RU (Helsinki-NLP fine-tuné).
- Évaluation finale (opus_books complet, 5 ép.) : BLEU=24.6, chrF=50.2.
- Amélioration notable par rapport aux tests initiaux (+2 points BLEU).
- Écart avec EA demeure (attendu vu les données/modèle).

Conclusion et Perspectives

Conclusion :

- Implémentation réussie d'un pipeline TAN EN->RU (Helsinki-NLP fine-tuné).
- Évaluation finale (opus_books complet, 5 ép.) : BLEU=24.6, chrF=50.2.
- Amélioration notable par rapport aux tests initiaux (+2 points BLEU).
- Écart avec EA demeure (attendu vu les données/modèle).

Perspectives (Projet Actuel) :

- Affiner mon modèle davantage sur WMT19.

Conclusion et Perspectives

Conclusion :

- Implémentation réussie d'un pipeline TAN EN->RU (Helsinki-NLP fine-tuné).
- Évaluation finale (opus_books complet, 5 ép.) : BLEU=24.6, chrF=50.2.
- Amélioration notable par rapport aux tests initiaux (+2 points BLEU).
- Écart avec EA demeure (attendu vu les données/modèle).

Perspectives (Projet Actuel) :

- Affiner mon modèle davantage sur WMT19.

Perspectives (Futures Expériences) :

- Utiliser WMT19 + GPU P100 et TPU pour viser BLEU $\sim$ 40.
- Tester NLLB sur WMT19.
- Explorer optimisation hyperparamètres.

Références I

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations (ICLR).

Loïc Barrault, Ondřej Bojar, Christian Federmann, Yvette Graham, Barry Haddow, Chris Hokamp, Philipp Koehn, Christof Monz, Lucia Specia, and Antonio Toral. 2020. Findings of the 2020 conference on machine translation (wmt20). In Proceedings of the Fifth Conference on Machine Translation, WMT20, pages 1–55. Association for Computational Linguistics.

Marta R. Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Ela Licht, Xutai Ma, Jean Maillard, and NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. In Findings of the Association for Computational Linguistics : ACL 2022. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. In Proceedings of ACL 2018, System Demonstrations, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Vladislav Molchanov, Valery Solovyev, Vasily Gerasimov, and Maria Shmatova. 2023. Promt systems for wmt23 shared general translation task. In Proceedings of the Eighth Conference on Machine Translation, WMT23, pages 125–130, Singapore. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, Sergey Edunov, David Grangier, and Marc' Aurelio Ranzato. 2019. Facebook fair's wmt19 news translation task submission. In Proceedings of the Fourth Conference on Machine Translation, volume 1 of WMT19, pages 300–309, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann. 2012. OPUS – a collection of parallel corpora for machine translation. In Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, pages 2214–2218, Istanbul, Turkey. European Language Resources Association.

Références II

University of Helsinki. 2020. Helsinki-nlp / opus-mt models. https://huggingface.co/Helsinki-NLP. Accessed : 2025-04-23.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gómez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, NeurIPS 30, pages 5998–6008.

Felix Fischer Xu, Orhan Firat, Ankur Bapna, Mona Diab, Colin Cherry, Noah Constant, and Xavier Garcia. 2023. ALMA: Advanced language model-based machine translation assistant. arXiv preprint, arXiv :2309.11674.

Discussion

Merci !

Quelques définitions

SacreBleu : Comparer la traduction machine à plusieurs traductions «humaines» de référence et en mesurer le chevauchement des n-grammes

ChrF : Mesure la similarité entre la traduction machine et les références en se basant sur les n-grammes de caractères plutôt que de mots

MarianMT : framework de Traduction Automatique Neuronale (TAN) rapide et efficace, principalement connu pour être écrit en C++