Traduction automatique du Chinois classique vers l'Anglais: Comparaison Pivot vs Direct

LIU Yongcan, ZHUGE Mélanie
M2 TAL, INALCO, Paris, FRANCE
yongcanliu19@gmail.com, zhugemelanie@gmail.com

Abstract

Ce projet s'intéresse à la traduction automatique neuronale d'une langue ancienne: le chinois classique. Nous proposons une architecture Transformer et évaluons notre hypothèse: la stratégie de traduction par pivot via le chinois moderne surpasse la traduction directe dans un contexte de forte complexité syntaxique. Trois modèles sont entraînés en intégrant une contextualisation historique: W2M (chinois classique vers chinois moderne), M2E (chinois moderne vers l'anglais) et W2E (chinois classique vers l'anglais). Les performances sont évaluées quantitativement à l'aide du score BLEU et qualitativement par l'analyse humaine.

Keywords: Traduction automatique neuronale, Traduction par pivot, Architecture Transformer, Chinois classique (Wenyan)

1. Introduction

La traduction du chinois classique présente un double défi: la rareté des corpus parallèles et la forte ambiguïté sémantique des caractères, qui évolue selon les dynasties. L'objectif de ce projet est de comparer deux stratégies. La première consiste en une cascade avec un pivot, en traduisant d'abord le chinois classique vers le chinois moderne puis vers l'anglais, nous supposons que l'utilisation du chinois moderne comme intermédiaire permet de réduire le fossé sémantique. La seconde est une traduction directe du chinois classique vers l'anglais. Nous analysons la performance des Transformers construits avec Keras ainsi que l'effet du prétraitement enrichi par des métadonnées temporelles, tout en comparant les résultats de ces deux méthodes de traduction automatique à l'aide du score BLEU et par l'analyse humaine. De plus, des métadonnées relatives aux dynasties de la Chine sont injectées sous forme de prompts temporels en entrée du modèle afin de contextualiser la traduction.

2. Etat de l'art

Depuis l'introduction de l'architecture Transformer (Vaswani et al., 2017), les modèles basés sur l'attention sont devenus la norme en traduction automatique, surpassant les anciens modèles récurrents (RNN/LSTM). Par ailleurs, la traduction en cascade (Pivot) est une technique standard pour les langues low-resource, bien qu'elle soit sujette à la propagation d'erreurs. Dans le contexte du chinois classique, l'approche pivot via une langue intermédiaire est souvent utilisée pour compenser le manque de données parallèles directes, tandis que la traduction directe est plus rapide mais peut manquer de contexte historique ou lexical. Plus récemment, l'état de l'art s'est déplacé vers les LLM pré-entraînés sur des corpus massifs. C'est le cas de modèles comme Tencent Hunyuan, que nous avons tester sa performance dans un autre cours. Ils ne sont pas seulement entraînés à traduire, mais à "comprendre" le contexte global grâce à des milliards de paramètres. Contrairement à notre approche supervisée qui nécessite un corpus parallèle aligné, ces modèles excellent dans le Zero-shot ou Few-shot learning. Toutefois, ils nécessitent des ressources de calcul colossales.

3. Données

Au départ, ce projet avait pour but la traduction du chinois classique vers le français. Mais malheureusement nous avons rencontré un manque de corpus parallèles bien alignés pour cette paire de langues. Face à ce problème, nous avons changé l'orientation du travail vers l'anglais, et nous avons alors choisi le jeu de données WenYanWen_English_Parallel, créé par KaifengGGG.

Le jeu de données utilisé est présent sur la plateforme Hugging Face, ce dataset contient au total 972 467 phrases, il s'agit d'un fichier organisé au format JSON et contient trois parties parallèles: la première partie correspond au chinois classique, la deuxième partie correspond au chinois moderne, la troisième partie correspond à l'anglais.

De plus, une métadonnée info fournit le titre de l'œuvre, que nous exploitons pour la contextualisation temporelle.

4. Systèmes proposés et points de comparaison

Pour notre tâche, nous avons construit plusieurs modèles basés sur l'architecture Transformer, implémentés à l'aide de Keras en utilisant l'API fonctionnelle de tensorflow.keras et les couches de bas niveau (layers). Trois variantes de modèle ont été expérimentées. Le premier modèle, W2M, traduit le chinois classique vers le chinois moderne et constitue la première étape de l'approche par pivot. Le second, M2E, traduit le chinois moderne vers l'anglais, complétant ainsi la chaîne cascade. Le troisième modèle, W2E, traduit directement du chinois classique vers l'anglais.

L'architecture commune à ces trois modèles reprend la structure canonique du Transformer (Vaswani et al., 2017) de type Encodeur-Décodeur.

Une spécificité de notre approche réside dans la représentation des données d'entrée pour les modèles traitant le Wenyan (W2M et W2E).

Nous avons pris en compte l'évolution du sens des caractères classiques selon les époques, a donc aussi enrichi les séquences de texte avec des informations temporelles. Chaque phrase source commence par une indication de la dynastie d'origine du texte, cette approche permet de guider l'encodeur car elle l'aide à mieux interpréter le sens des caractères selon leur contexte historique. Pour le système pivot, le processus se fait par étapes: le système effectue d'abord une première phase de décodage, d'où le modèle W2M produit une traduction intermédiaire en chinois moderne. Cette traduction sert ensuite de base pour la suite du traitement.

Cette sortie est ensuite utilisée comme entrée pour le second modèle (M2E) afin de produire la traduction finale en anglais. Cette méthode permet d'exploiter la proximité syntaxique entre le chinois moderne et l'anglais lors de la seconde étape.

5. Expériences

5.1. Prétraitements

Après nettoyage et filtrage des phrases invalides, 971 853 phrases ont été conservées, soit 99,9% des données. Pour l'ensemble des expériences, nous avons utilisé la totalité des données. Chaque texte a été encodé à l'aide de tokenizers spécifiques: le chinois classique dispose d'un vocabulaire de 4000 tokens, le chinois moderne de 5000 tokens et l'anglais de 3137 tokens.

En amont de la tokenisation, les titres des œuvres sont convertis en marqueurs dynastiques via un dictionnaire de règles, puis concaténés en début de séquence. Il est à noter que cette étiquette reflète la période narrative (le sujet de l'histoire) et non nécessairement la date de rédaction. Par exemple, une œuvre rédigée sous les Ming mais relatant les Trois Royaumes sera classée "Trois Royaumes".

Ensuite, chaque texte a été encodé à l'aide de tokenizers spécifiques. Afin de maîtriser la complexité computationnelle et la taille des matrices d'embeddings, nous avons construit des tokenizers spécifiques limités aux tokens les plus fréquents. Le Tableau 1 résume les caractéristiques des données après tokenisation.

Langue	Taille du Vocabulaire	Longueur Max. de Séquence
Chinois Classique	4000	150
Chinois Moderne	5000	200
Anglais	3137	300

Table 1: Statistiques du corpus et des vocabulaires.

Pour garantir une évaluation robuste et éviter le sur-apprentissage, notre corpus a été divisé de manière aléatoire en trois sous-ensembles disjoints selon une répartition 80%/10%/10% sur l'ensemble d'entraînement, de validation et de test.

En amont de la tokenisation, une étape de prétraitement spécifique a été appliquée pour intégrer le contexte historique. Le dataset initial fournit le titre de l'œuvre, mais cette information n'est pas directement exploitable par le modèle. Nous avons donc utilisé un dictionnaire pour mapper les titres d'œuvres aux périodes historiques correspondantes. Cette information est ensuite injectée sous forme de prompt textuel concaténé au début de chaque séquence source. Le jeu de données initial fournit le titre de l'œuvre, mais cette information n'est pas directement utilisable par le modèle. Nous avons donc mis en place un dictionnaire de règles simples qui permet d'associer chaque titre à une période historique donnée. Cette information est ensuite ajoutée au début de chaque séquence source sous la forme d'un prompt textuel, afin de fournir un contexte temporel au modèle.

Cette syntaxe explicite permet au tokenizer de traiter le marqueur temporel comme une partie intégrante de la séquence d'entrée.Il convient de noter que cette attribution repose sur une approximation heuristique. Nous associons chaque texte à la période historique mentionnée dans le titre ou dans le contenu narratif. Cette période ne correspond pas toujours à la date réelle de rédaction de l'œuvre. Par exemple, un texte écrit sous la dynastie Ming mais qui raconte des événements de la période des Trois Royaumes est étiqueté selon la période décrite. Bien que cela introduise un biais

philologique, cette simplification permet de fournir un point d'ancrage contextuel au modèle sans nécessiter une datation carbone de chaque segment textuel.

5.2. Entraînement

Tous les modèles sont entièrement construits par nos soins avec les couches standards de Keras, sans recourir à des modèles pré-entraînés, ce qui permet de contrôler précisément l'entraînement et l'architecture. Chaque modèle est configuré avec les hyperparamètres suivants, choisis pour offrir un compromis entre capacité d'apprentissage et coût computationnel sur notre infrastructure : une taille d'embedding de 256, un réseau Feed-Forward interne de 1024 unités, un empilement de 4 couches pour l'encodeur et 4 couches pour le décodeur, 8 têtes d'attention par couche. Afin de gérer l'ordre des séquences sans récurrence, nous avons implémenté une couche de *Positional Encoding* personnalisée utilisant des fonctions sinusoïdales fixes. De plus, nous exploitons la gestion des masques de Keras pour ignorer dynamiquement les tokens de remplissage (*padding*) à travers toutes les couches du réseau, optimisant ainsi le calcul de la fonction de perte.

L'optimisation des poids est réalisée par l'algorithme Adam avec ses paramètres par défaut, couplée à un scheduler personnalisé. Ce dernier augmente linéairement le taux d'apprentissage au début (warmup) pour stabiliser les gradients, puis le réduit progressivement. Les calculs sont accélérés par l'utilisation de la précision mixte et une taille de lot (*batch size*) de 256.

Nous avons utilisé deux callbacks essentiels pour superviser l'entraînement, EarlyStopping (patience=5) arrête l'entraînement si la perte de validation ne s'améliore plus, et ModelCheck point pour sauvegarder la meilleure version du modèle.L'entraînement de trois modèles sur 15 époques a duré environ 7 heures sur le NVIDIA GeForce RTX 3090 (Ertix).

5.3. Résultats

Afin de valider la stabilité de l'entraînement et la capacité de généralisation des modèles, nous avons tracé l'évolution de la fonction de perte (*loss*) et de la précision (*accuracy*) pour les ensembles d'entraînement et de validation.

L'analyse des courbes d'apprentissage est caractérisée par une convergence accélérée durant les cinq premières époques. Pour les modèles M2E et W2E, la fonction de perte subit une chute drastique, passant de plus de 1.5 à moins de 0.5 en seulement trois itérations, tandis que la précision bondit de 67% à environ 90% sur ce même

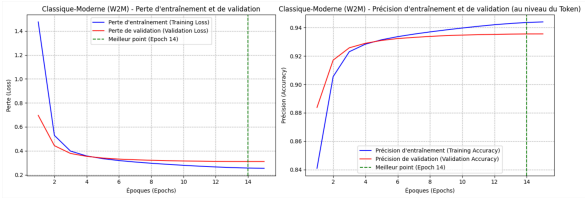

Figure 1: Classique-Moderne (W2M), Meilleur Epoch : 14, Val Loss = 0.3107

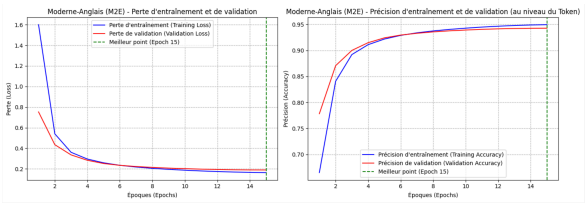

Figure 2: Moderne-Anglais (M2E), Meilleur Epoch : 15, Val Loss = 0.1891

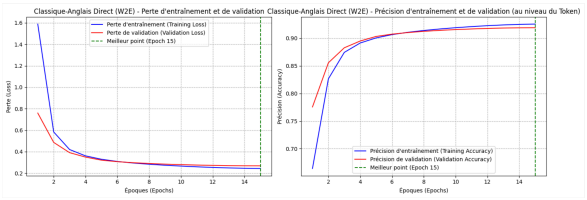

Figure 3: Classique-Anglais Direct (W2E), Meilleur Epoch : 15, Val Loss = 0.2681

intervalle. Cette phase initiale est suivie d'un ralentissement marqué entre la cinquième et la dixième époque, où les gains deviennent incrémentaux, la précision progressant péniblement de 90% à 93%. Au-delà de la 10ème époque, bien que la courbe semble atteindre un plateau asymptotique avec des variations inférieures à 0.5% jusqu'à 15 époques.

Par ailleurs, une évaluation qualitative s'est posée sur une comparaison entre les traductions générées par le pipeline en cascade (W2M -> M2E) et le modèle direct (W2E) sur des échantillons du jeu de test. Deux évaluateurs ont examiné chacun un échantillon aléatoire de dix phrases distinctes (soit 20 phrases au total) et ont attribué une note de 1 à 5 pour juger la qualité de chaque traduction.

Les exemples présentés montrent que les traductions en cascade produisent des sorties globalement plus cohérentes sur le plan syntaxique.

Enfin, une évaluation quantitative a été réalisée à l'aide du score BLEU, calculé sur un sous-ensemble aléatoire de 100 phrases issues du jeu de test, en comparant les sorties des modèles aux traductions de référence humaines.

Le modèle classique → moderne atteint un score BLEU de 30.11, tandis que le modèle mod-

PIV.	DIR.	Source	W2M	M2E	W2E
5	4	五月甲戌朔	五月初一为甲戌日	The first day of May was Jiaxu Day.	The first day of May is Jiaxu.
3	3	郑忽出奔卫	郑国的忽然出逃到卫国	Zheng suddenly fled to Wei.	Zheng Hu fled to Wei.

Table 2: Evaluation humaines.

Modèles	Scores BLEU
W2M	30.11
M2E	19.62
W2E	10.64

Table 3: Scores BLEU des modèles.

Source	Référence	Prédiction (Erreur)	Analyse
后复结陈向城	Later, the rebel soldiers lined up to attack the city again.	**M2E:** Later, he made friends with Chen Juzhen and advanced in the city. **W2E:** Later, he colluded with the Chen Dynasty again.	En raison des caractéristiques grammaticales du Wenyan, le sujet « the rebel soldiers » est omis dans le texte source. Mais le sujet doit être explicite en anglais, ce qui conduit le modèle à ajouter le sujet erroné « he » dans la traduction anglaise, causant ainsi une erreur sémantique.

Table 4: Analyse des erreurs (Hallucination du sujet).

erne → anglais obtient un score de 19.62. En comparaison, le modèle direct classique → anglais présente un score BLEU plus faible, égal à 10.64.

6. Discussion

L'objectif de cette comparaison est d'évaluer l'impact de la traduction pivot sur la qualité des résultats et de déterminer si la traduction directe peut atteindre des performances comparables. Le score final du pipeline pivot (BLEU 11.64) n'est que légèrement supérieur à celui du modèle direct (BLEU 10.64). À première vue, cet écart de +1.0 BLEU peut sembler modeste. Cependant, il convient de le considérer en tenant compte du phénomène de propagation d'erreurs propre aux systèmes en cascade. Dans une approche pivot, les erreurs générées par le premier modèle (W2M) sont amplifiées par le second (M2E). Le fait que le système pivot parvient à surpasser le système direct, malgré cette double perte d'information, valide notre hypothèse du fossé sémantique. Le modèle direct (W2E) échoue à capturer les structures profondes du Wenyan et produit souvent une traduction "mot-à-mot" inintelligible, bien que certains mots-clés corrects lui permettent de maintenir un score BLEU artificiellement proche. L'évaluation humaine corrobore cette analyse : les sorties du système pivot sont syntaxiquement plus robustes. À l'inverse, le modèle direct traite souvent les termes métaphoriques de manière phonétique (*pinyin*) ou génère des phrases grammaticalement brisées.

Le problème de l'hallucination des sujets (Pro-drop) est le défi majeur. En effet, le chinois classique est une langue omission du sujet, alors que l'anglais exige un sujet explicite. Nos modèles tendent à "halluciner" ou inventer des sujets pour satisfaire cette contrainte grammaticale, créant parfois des contresens historiques.

Ces exemples illustrent que si le modèle Pivot réussit mieux à construire une phrase grammaticale Sujet + Verbe + Objet, il n'est pas immunisé contre les erreurs de référence contextuelles.

Au niveau de l'entraînement des modèles, l'étude des courbes de perte montre une convergence très rapide (dès la cinquième époque). Si cela témoigne de la stabilité de notre implémentation, le plafonnement rapide des performances suggère aussi une limitation architecturale. Avec une taille d'embedding de 256 et un vocabulaire restreint (5000 tokens), nos modèles saturent rapidement. De plus, l'utilisation du Greedy Search pour l'inférence est clairement identifiée comme un goulot d'étranglement : elle empêche le modèle de corriger une erreur initiale, conduisant aux boucles de répétitions observées sur les phrases longues dans le modèle direct.

7. Conclusion et perspectives

Pour conclure, bien que le risque de propagation d'erreurs soit réel, il est compensé par la capacité du système pivot à rétablir une structure cohérente. L'approche directe échoue souvent à produire des phrases grammaticalement correctes et porteuses de sens. Le passage par le chinois moderne simplifie considérablement la tâche de traduction vers l'anglais, ou même vers d'autres langues. Pour la traduction d'une langue ancienne, l'usage d'une langue moderne apparentée s'avère être une stratégie pertinente.

Au-delà des améliorations techniques, nous voudrions bien faire une extension vers le français via le pivot. En effet, l'un des obstacles majeurs de ce projet a été l'absence de corpus aligné français suffisamment grand, une prochaine étape sera donc remplacé M2E par un modèle traduisant du chinois moderne vers le français.

Une autre perspective stimulante serait d'inverser la direction de la traduction pour entraîner un modèle capable de traduire du chinois moderne, de l'anglais ou du français vers le chinois classique. Cette tâche est plus complexe que le sens inverse car le chinois classique est une langue extrêmement concise et implicite.

Dans le cadre de notre projet, nous avons

délibérément opté pour une approche basée sur un dictionnaire pour l'extraction des dynasties. Ce choix visait à concentrer nos ressources sur l'architecture de la traduction automatique sans complexifier le pipeline avec une tâche secondaire de classification. Toutefois, une perspective d'amélioration serait de remplacer ce système de règles par un véritable module de reconnaissance d'entités nommées, cela permettrait d'identifier la période historique à partir du style littéraire et d'adapter le style de la traduction en conséquence.

Acknowledgments et éthique

Ce projet est lié au projet du cours Traduction automatique avec Maxime FILY, où nous allons faire des évaluations sur des modèles pré-entraînés, en comparant avec les modèles que nous avons fait dans ce cours avec tencent/Hunyuan-MT-7B, en utilisant les mêmes jeu de données, mais en plus avec un petit corpus parallèle trilingue (chinois classique, chinois moderne et français) faites par nous même basé sur https://github.com/NiuTrans/Classical-Modern.

Nous tenons à remercier les auteurs du jeu de données KaifengGG pour la mise à disposition de ce corpus trilingue précieux.

Dans un souci de transparence et d'intégrité académique, nous déclarons avoir eu recours à des assistants basés LLM (GPT et DeepSeek) au cours de ce projet. Leur utilisation a été strictement encadrée et limitée à des tâches de débogage et des corrections d'erreurs de syntaxe: Nous avons observé une saturation précoce de la précision (atteignant 100% dès la 10ème époque), GPT nous a expliqué que ce phénomène était dû à une mauvaise propagation du masque dans la couche PositionalEncoding. De plus, la qualité des traductions a initialement souffert d'un décodage glouton où une seule erreur en début de phrase compromettait la cohérence globale, l'ajout de pénalités de répétition ont été proposé par l'agent pour améliorer la robustesse des modèles.

DeepSeek a servi d'aide à la compréhension du chinois classique, et à la structuration des métadonnées de la dynastie de Chine.