Traduction Automatique neuronale du Mandarin vers le Wu (Shanghaïen)

Simeng SONG Xiaobo WANG
INALCO, Paris, France
simeng.song@outlook.com, sibelwang17@gmail.com

Abstract

Cet article présente une étude sur la traduction automatique neuronale du Mandarin vers le Wu (Shanghaïen), une langue peu dotée utilisée dans la région de sud d'est de la Chine. Nous proposons un système basé sur l'architecture Transformer entraîné sur un corpus parallèle de 3,855 paires de phrases. Faute de ressource, nous explorons la méthode de back-translation pour augmenter les données, en générant des paires pseudo parallèles à partir d'un corpus monolingue en Wu. Nos expériences comparent le modèle de base entraîné uniquement sur les données réelles à une variante intégrant des données synthétiques. Notre étude vise également à analyser les bénéfices et les limites de la back-translation dans un contexte de langue peu dotées. Cela contribue ainsi à une meilleure compréhension des défis spécifiques à la traduction automatique des dialectes chinoises.

Keywords: Traduction automatique neuronale, Mandarin, Wu, baseline, langue peu dotée

1. Introduction

La traduction automatique neuronale a connu des avancées majeures ces dernières années, notamment grâce à l'architecture Transformer, qui a permis d'améliorer significativement la qualité des traductions dans de nombreuses paires de langues bien dotées. Cependant, ces progrès restent largement inégaux selon les langues concernées. Les langues dites *peu dotées*, et en particulier les langues régionales ou dialectales, continuent de poser des défis importants, principalement en raison du manque de ressources parallèles de grande qualité.

Le Wu, groupe de variétés sinitiques parlées principalement dans la région du delta du Yangzi (dont le shanghaïen est la variété la plus connue), constitue un exemple typique de langue peu dotée dans le domaine de la traduction automatique. Bien que le mandarin dispose aujourd'hui de ressources abondantes, la langue Wu reste faiblement représentée dans les corpus annotés et parallèles disponibles publiquement, ce qui limite fortement l'entraînement de modèles neuronaux performants pour cette paire de langues.

Dans ce projet, nous nous intéressons à la tâche de traduction automatique du mandarin vers le wu, en nous appuyant sur un corpus parallèle de taille limitée. Dans un tel contexte de rareté des données, une question centrale se pose : **Comment améliorer les performances d'un système de traduction automatique lorsque les données parallèles disponibles sont insuffisantes ?**

Une approche classique pour répondre à ce problème consiste à exploiter des données monolingues au moyen de techniques d'augmentation de données, parmi lesquelles la back-translation occupe une place centrale. Cette méthode, largement utilisée pour les langues peu dotées, consiste à générer des données pseudo-parallèles à partir de corpus monolingues, dans l'espoir d'enrichir les données d' apprentissage du système principal.

L'objectif de ce travail est double. Dans un premier temps, nous proposons un système de traduction neuronale basé sur uns architecture Transformer entraînée sur un corpus parallèle mandarin-wu. Dans un second temps, nous cherchons l'impact de l'ajout de données pseudo-parallèles obtenues par back-translation à partir d'un corpus monolingue en wu.

Ce projet s'inscrit dans une démarche expérimentale visant à analyser les bénéfices et les limites de la back-translation dans un cadre de ressources extrêmement limitées, et à mieux comprendre les facteurs qui influencent l'efficacité de cette méthode pour la traduction automatique des langues peu dotées.

2. État de l'art

2.1. Traduction automatique neuronale et architecture Transformer

La traduction automatique neuronale (Neural Machine Translation, NMT) s'est imposée comme l'approche dominante en traduction automatique au cours de la dernière décennie. Contrairement aux systèmes statistiques ou à règles, les modèles neuronaux reposent sur un apprentissage de bout en bout à partir de données parallèles, permettant une modélisation plus souple et plus contextuelle des correspondances entre langues.

Parmi les architectures proposées, le modèle Transformer constitue aujourd'hui une référence incontournable. Introduit par Vaswani et al., il se distingue par l'abandon des mécanismes récurrents

au profit de couches d'attention multi-têtes, capables de capturer efficacement les dépendances à longue distance. Cette architecture a démontré des performances supérieures sur de nombreuses paires de langues, tout en offrant une meilleure parallélisation lors de l'entraînement(Vaswani et al., 2017).

Cependant, l'efficacité du Transformer dépend fortement de la qualité et de la quantité des données parallèles disponibles. Dans les contextes où les ressources sont abondantes, ces modèles atteignent des performances élevées, mais leur généralisation reste limitée lorsque les corpus sont de taille réduite.

2.2. Traduction automatique pour les langues peu dotées

Les langues peu dotées représentent un défi majeur pour la traduction automatique neuronale. Le manque de corpus parallèles de grande taille entraîne souvent des modèles instables, sujets au surapprentissage et à des erreurs de généralisation (Koehn and Knowles, 2017). Ce problème est particulièrement aigu pour les langues régionales, dialectes ou minoritaires, qui disposent rarement de ressources standardisées ou largement diffusées.

Dans ce contexte, plusieurs stratégies ont été essayées dans la littérature, notamment le transfert inter-langues, l'apprentissage multilingue ou encore l'augmentation artificielle des données. Néanmoins, lorsque les langues concernées présentent des différences lexicales, phonologiques ou syntaxiques importantes, ces approches peuvent s'avérer complexes à mettre en oeuvre.

Le Wu, et plus particulièrement le shanghaïen dans notre projet, s'inscrit pleinement dans cette problématique. Bien qu'il soit étroitement lié au mandarin sur le plan historique, il présente des spécificités linguistiques marquées, tant au niveau lexical que morphosyntaxique. Ces différences rendent la traduction automatique directe difficile, en particulier lorsque les données disponibles sont limitées à des corpus de taille modeste.

2.3. Back-translation comme méthode d'augmentation de données

Parmi les méthodes d'augmentation de données proposées pour les langues peu dotées, la back-translation s'est imposée comme une approche simple et efficace. Le principe consiste à exploiter des corpus monolingues dans la langue cible afin de générer des données pseudo-parallèles : un modèle auxiliaire traduit les phrases monolingues vers la langue source, produisant ainsi des paires synthétiques qui peuvent être ajoutées aux données d'entraînement du système principal.

Cette méthode a montré des gains importants dans de nombreux contextes, en particulier lorsque les données parallèles initiales sont insuffisantes. Elle peut enrichir la diversité lexicale et syntaxique des données d'apprentissage, tout en exploitant des ressources monolingues souvent plus abondantes (Sennrich et al., 2016).

Cependant, plusieurs travaux soulignent également les limites de la back-translation. La qualité des données générées dépend fortement des performances du modèle utilisé pour la traduction inverse. En présence de bruits ou de traduction erronées, l'ajout massif de données pseudo-parallèles peut dégrader les performances du système final (Edunov et al., 2018). La question du filtrage et du contrôle de qualité des données synthétiques apparaît ainsi comme un enjeu central, en particulier dans les scénarios de ressources extrêmement limitées.

2.4. Positionnement du présent travail

Dans ce projet, nous nous inscrivons dans cette lignée de recherches en explorant l'usage de la back-translation pour la traduction automatique du mandarin vers le wu (shanghaïen). Contrairement aux études menées sur des langues mieux dotées, notre objectif n'est pas d'optimiser à tout prix les performances absolues, mais d'analyser de manière contrôlée l'impact de l'ajout de donnée pseudo-parallèles dans un cadre réaliste de faible disponibilité des ressources.

Nous comparons ainsi un modèle de base entraîné uniquement sur des données parallèles réelles à une variante intégrant des données synthétiques, afin d'évaluer dans quelle mesure cette stratégie permet d'améliorer (ou non) les performances du système.

3. Données

Les expériences menées dans ce travail reposent sur un corpus parallèle mandarin-wu issu de la plateforme **MagicHub**. Plus précisément, nous utilisons le corpus **ASR-SCShhiDiaDuSC**, intitulé *A Scripted Chinese Shanghai Dialect Daily-use Speech Corpus*.

Ce corpus a été initialement conçu pour des tâches de reconnaissance automatique de la parole (ASR) et se compose d'énoncés de la vie quotidienne, enregistrés et transcrits en mandarin et en wu. Dans le cadre de ce projet, seules les transcriptions textuelles alignées ont été exploitées pour la traduction automatique.

Les textes alignés sont fournis au format CSV, chaque ligne représentant une paire de phrases parallèles. Dans ce projet, la traduction est formulée dans la direction suivante :

- Langue source : Mandarin
- Langue cible : Wu (Shanghaïen)

Le corpus a été divisé en trois sous-ensembles distincts, conformément aux pratiques standards en traduction automatique neuronale :

- Train : 3,855 paires
- Dev : 481 paires
- Test : 483 paires

L'ensemble d'entraînement est utilisé pour l'apprentissage des modèles, tandis que l'ensemble de développement sert à l'ajustement des hyperparamètres et au suivi de l'entraînement. L'ensemble de test, strictement tenu à l'écart des phases d'apprentissage, est réservé à l'évaluation finale des performances.

Ce découpage permet de garantir une évaluation équitable et reproductible des différents systèmes comparés dans ce travail.

4. Systèmes proposés et points de comparaison

Comme nous avons mentionné au dessus, nous proposons deux systèmes de traduction automatique neuronale basés sur l'architecture Transformer, tous les deux entraînés pour traduire le Mandarin vers le Wu. Ces deux systèmes se distinguent par les données d'entraînement utilisées.

4.1. Transformer Baseline : traduction mandarin vers wu

Nous désignons le premier système comme baseline, qui est un point de référence. Il s'agit d'un modèle Transformer standard (encodeur - décodeur) entraîné sur le corpus parallèle réel décrit dans la section 3. Malgré la limitation des ressources, ce système nous permet d'établir une performance assez satisfaisante.

Le modèle est implémenté avec l'aide de l'API Keras. Les paires de phrases sont représentées au niveau des caractères donc pas de segmentation préalable. Les encodeurs et les décodeurs sont composés de 4 couches, chacune intégrant un mécanisme d'attention multi-têtes et un réseau feed-forward positionnel. Nous avons ajouté les embeddings lexicaux pour donner l'ordre des caractères, et également, un mécanisme de dropout pour limiter le surapprentissage.

4.2. Transformer avec back-translation

Le second système intègre des corpus pseudo-parallèles générés par la back-translation. Nous avons d'abord entraîné un modèle Transformer avec la direction inverse à partir du corpus parallèle. Et ce modèle est ensuite utilisé pour traduire un corpus monolingue de Wu. Après la back-translation, nous avons obtenu 405 paires de phrases comme notre corpus pseudo-parallèle. Au début de la constrution du modèle, nous avons essayé une stratégie de réentraînement complet, cependant, la performance n'est pas bien du tout. Nous pensons donc de fine-tuning le modèle baseline, c'est-à-dire le modèle est entraîné d'abord sur les données réelles, et puis affiné sur un corpus combinant les données réelles et les données pseudo-parallèles avec la méthode Shuffle, qui mélange aléatoirement les données. Ce système permet au modèle d'abord d'apprendre les correspondances fiables avant d'aller plus loins pour enrichir la couverture.

5. Expérience

5.1. Prétraitement

Le corpus sont fournies au format csv, qui est plus claire et plus simple, avecune paire de phrases par ligne. Comme le corpus est déjà très propre, nous n'avons pas fait le prétraitement au niveau linguistique. Nous ajoutons juste les symboles spéciaux au début et à la fdin de phrases pour permettre un décodage auto-régressif lors de la génération.

5.2. Entraînement

5.2.1. Hyperparamètres

Compte tenu de la taille du corpus d'entraînement, nous avons configuré le modèle avec une capacité réduite mais suffisante pour être efficace. Les hyperparamètres sont les suivants :

- Batch size : 64
- Loss function : Sparse Categorical Crossentropy
- Optimiseur : Adam avec learning rate schedule
- Learning rate initial : 5e-4
- Learning rate après : 1e-4
- Dropout : 0.1

5.2.2. Entraînement des modèles

Les deux systèmes sont entraînés identiquement avec les callbacks de early stopping et checkpoint. L'early stopping permet le modèle arrêter si le loss de validation n'améliore pas pendant 5 epochs consécutives. Le model checkpoint est pour sauvegarder le meilleur modèle selon le loss de validation.

Sauf cela, le modèle baseline est entraîné pendant 20 epochs maximum et puis affiné pendant 10 epochs supplémentaires sur le corpus combiné, en utilisant un learning rate plus faible. Avec la méthode de fine-tuning, elle permet au modèle d'acquérir une base solide sur les données originales, d'éviter la perturbation.

5.3. Résultats et évaluation

5.3.1. Métriques d'évaluation automatique

Nous avons choisi les métriques automatiques BLEU et chrF avec l'outil sacreBLEU sur l'ensemble de test.

Vue que nos modèles sont entraînés au niveau caractère, les scores BLEU sont calculés avec une tokenisation caractères. Cela permet de garantir une cohérence entre la représentation des données pendant l'entraînement et l'évaluation.

Nous savons que le BLEU est utilisée afin de mesurer le degré de correspondance n-grammes entres les traductions automatiques et les références humaines. Alors que dans un contexte de langue peu dotée, BLEU peut être très sensible et pénaliser les petites erreurs acceptables.

Nous complétons donc BLEU par chrF, qui peut capture plus finement les similarités au niveau linguistique. Le chrF est particulièrement adaptée aux langues non segmentée telles que le chinois. Et c'est pour ça que nous avons choisi le chrF.

5.3.2. Métrique d'évaluation manuelle

Pour voir la performance au niveau de linguistique, une évaluation manuelle a été réalisée sur un sous-ensemble de pharses extraites aléatoirement de l'ensemble de test.

5.3.3. Résultats des scores

Comme présenté dans la figure 1 et la table 1, les résultats montrent que l'intégration de données permet une légère amélioration sur deux mértiques. Le système avec back-translation a obtenu des scores un peu supérieurs à ceux du modèle baseline.

Cette phénomène suggère que les données synthétiques générées donnent les informations supplémentaires utiles au modèle, surtout dans la couverture de lexique et de structure. Malgré la progression est très légère, il est quand même important de noter que la progression est obtenu uniquement par le fine-tuning avec 405 paires de back-translation. Ce qui nous confirme l'intérêt de la stratégie que nous avons choisie.

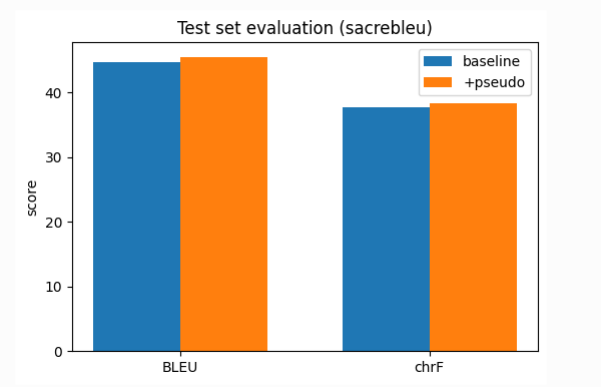

Figure 1: Les scores BLEU et chrF sur l'ensemble de test

Modèle	BLEU	chrF
Baseline	44.8	37.6
+pseudo	45.8	38.4

Table 1: Table de scores BLEU et chrF sur l'ensemble de test

5.3.4. Résultats des évaluations manuelles

Le figure 2 présente quelques exemples comparant les traductions générées par deux systèmes, en regard de la référence humaine.

Du point de vue d'une locuteure du Wu, les différences entres les deux systèmes sont assez perceptibles. Les sorties de la baseline présentent souvent des problèmes de cohérence sémantique et de naturalité. Les phrases produites ont des erreurs lexicales ou des substitutions de mots dans lien avec le contexte (souvent remplacer par les mots qui ont la prononciation similaire). Il s'agit même des mots ont la signification inverse, comme par exemple "伐急撒" (pas de précipiation) dans la référence, mais le modèle baseline les traite totalement différent "急了伐" (urgent, précipiation) Cela rend la traduction difficile à comprendre.

En revanche, le modèle intégrant la back-translation produit des phrases globalement proche de la référence. Les structures sont plus stables, et les traductions sont similaire à ce qu'on utilise dans l'oral réel. Surtout les particules modales telles que "伐", "啊", "呢", sont utilisée de manière plus appropriée. Les phrases sont très compréhensible pour un locuteur natif. Cependant, il s'agit quand même certaines limites. Nous observons encore des choix lexicaux parfois trop standardisés (proche de Mandarin), par exemple le gros mots de "册那" n'est pas bien traduit, mais le modèle utilise le mots en Mandarin "靠". Mais ces erreurs restent toutefois localisées et ne nous empêchent pas la compréhension globale du message.

En conclusion, cette évaluation manuelle con-

mandarin (src)	wu_ref (gold)	wu_baseline	wu_pseudo
我参加同舟共济都不参加齐心协力	吾参加同舟共济都伐参加齐心协力!	吾调加同意思今朝伐敢教师全部车子!	吾认之后头同意侪伐参与安排侬。
是你的意思吧？好好看看信息	是侬呃意思伐？好好看看消息	是侬呃意思伐？好伐？好看看信息	是侬呃意思伐？好好好看看信息
靠！，我给贾照玲打电话呢!	册那！，吾帮贾照玲打电话呢!	靠！吾拨李严重拨电话呢!	靠！，吾拨拜六百块打电话呢!
你说吧我们都用一颗诚意交谈好吗	侬讲哇，阿拉侪用一颗诚意交谈好哇?	侬讲伐阿拉侪用一颗减肥安排好伐?	侬讲伐阿拉侪用一颗减肥安吃好伐?
晕 快了快了不急撒	晕，快了快了，伐急撒	晕哦 快了快点急了伐	晕 快了快了伐急啥?
我好朋友我想你	吾好朋友吾想侬	吾好朋友吾想侬	吾老朋友吾想侬
哦 那你别喝太多了	哦，葛侬伐吃太多了	哦 那侬伐要吃多了	哦 那侬伐要吃多了
好的，我明天打电话，晚安	好呃，吾明朝打电话，夜安	好呃，吾明朝打电话，晚安	好呃，吾明朝打电话，晚安
如果你不肯原谅我就算勒!	假使侬伐肯原谅吾就算了!	如果侬伐肯原谅吾就算了!	假使讲侬伐肯原谅吾就算了!
你什么时候结婚啊	侬撒辰光结婚啊	侬啥辰光结婚啊	侬啥辰光结婚啊
8 30就已回来 你来陪我就不痛了	八三零就已回来 侬来陪吾就伐痛了	八 突便就回来，侬来寻吾就是伐痛了	八 零零就回来，侬来寻吾就伐
好冷酷现在开始	好冷酷现在开始	好叫那现在开始	好意思现在开始
不接就不接吧	伐接就伐接吧	伐接就伐接伐	伐接就伐接伐
弟弟睡着吗	弟弟困着吗	阿弟困着伐	阿弟困着伐
给我打电话撒事	拨吾打电话啥事体?	拨吾打电话撒事体	拨吾打电话撒事体
好让我感动耶!	老让吾感动耶!	好让吾感动伐!	好让吾感动呢!
你不要走好吗	侬伐走好吗	侬伐要走好伐	侬伐要走好伐
怎么会 你忙萨呢	哪能会 侬忙啥呢	哪能会 侬忙啥呢	哪能会的 侬忙啥呢
你怎么就觉得不对劲了呢	侬哪能就觉着伐对劲了呢	侬哪能就觉得伐对了呢	侬哪能就觉得伐对了呢
我又不想睡觉你在复习吗?	吾又伐想困高，侬勒复习伐?	吾又伐想困觉侬辣海复习吗?	吾又伐想困觉侬辣复习伐?

Figure 2: Évaluation manuelle sur l'ensemble de test

firme les tendances observées dans l'évaluation automatique. Même si la progression selon BLEU et chrF demeure modérée, l'analyse manuelle montre que la back-translation contribue à beaucoup améliorer la lisibilité et la cohérence de la langue Wu. Les résultats soulignent aussi l'importance de l'évaluation manuelle dans le cas de langues peu standardisées.

6. Discussion

Cet article permet de tirer un peu enseignements sur l'utilisation de la back-translation dans un contexte de traduction automatique pour une langue peu dotée. D'abord, les résultats obtenu dans la section précédente montrent que l'ajout de donnée pseudo-parallèles conduit à une amélioration mais modérée des performances.

Étant donnée que le nombre de nos données est restreint, il peut explique le phénomène des améliorations limitées. Néanmoins, le fait que la progression soient cohérente sur les deux mértiques automatiques nous confirme l'avantage de la back-translation. Sauf cela, la comparaison entre BLEU et chrF met en évidence les limites des mértiques automatiques dans la traduction de la langue peu dotée. Normalement chrF se révèle plus robuste et refléter mieux les améliorations observées, mais dans notre cas, il n'est pas très évident (BLEU +1 score ; chrF +0.8 score, il n'y a pas beaucoup de différence).

Par ailleurs, les résultats montrent que la stratégie de fine-tuning du modèle baseline est un choix plus stable. Le réentraînment complet ont conduit à une dégradation des performances, nous supposons que ça pourrait due au bruit introduit par les données synthétiques.

Enfin, il existe certaines erreurs dans les deux systèmes, surtout pour les expressions ou des choix lexicaux. Nous pouvons constater un défis inhérents à la traduction automatique des diaclectes, où la variabilité linguistique et la langue plutôt orale compliquent à la fois l'entraînement et l'évaluation des modèles.

7. Conclusion et perspectives

Dans cet article, nous avons présenté une étude sur la traduction automatique neuroale du Mandarin vers le Wu, en nous appuyant sur une architecture Transformer. Faute des données, nous avons exploré l'utilisation de la back-translation comme la façon d'augmentation de données, en comparant un modèle baseline qui est entraîné uniquement sur les données réelles à un système intégrant des données pseudo-parallèles.

Ce travail présente néanmoin plusieurs limites malgré la progression des résultats cohérente. La taille réduite du corpus restreint l'apprestissage intiale et l'amélioration des modèles. De plus, l'évaluation automatique reste pas suffisante pour refléter la qualité des traductions des dialectes, ce qui rend l'évaluation humaine (coûteuse) indispensable.

Il y aura plusieurs perspectives pour prolonger et améliorer ce travail. Premièrement, il s'agit une piste consisterait à exploiter un volume plus important de données originales. Deuxièmement, l'exploration de cadre multilingues ou la traduction par pivot, en tirant parti des similarités entre différentes variétés. Finalement, l'utilisation des autres métriques d'évaluation plus sensible aux dialectes, ou une évaluation manuelle à plus grande échelle, peuvent permettre d'améliorer l'analyse des résultats.

8. Acknowledgments et éthiques

Les données utilisées dans ce projet proviennent de la plateforme MagicHub, distribué sous la licence Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International. L'utilisation de ces données respecte les conditions

de la licence, et aucune information personnelle ou sensible n'est impliquée dans ce travail.

Des outils d'IA générative ont été utilisés uniquement comme assistance linguistique lors de la rédaction du rapport. En tant que locuteurs non natifs du français, les auteurs ont eu recours à ces outils pour la correction grammaticale, la reformulation stylistique et l'amélioration de la clarté rédactionnelle.

9. Références

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*.

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