Document de travail · Vérification interne | v1.0 | 2026-11-20

Le choix des textes de test change la qualité mesurée d’une traduction de l’anglais vers le créole haïtien

Comparaison des scores publiés pour des phrases générales, des textes religieux et des récits éducatifs

Adrian Dunkley | info@caribbeanaiassociation.com

Résumé

Un système de traduction peut obtenir des scores différents selon le texte à traduire. Nous avons comparé les résultats publiés d’OPUS de l’anglais vers le créole haïtien sur Tatoeba, JW300 et MIT-Haiti. Nous avons ensuite comparé deux modèles sur MIT-Haiti. Remplacer Tatoeba par des récits éducatifs fait baisser le score d’OPUS de 30,5 points BLEU et de 23,4 points chrF. Changer de modèle sur ces récits donne des écarts respectifs de 7,3 et 8,1 points. Ces chiffres indiquent un risque lors de la conception de tests pour la traduction scolaire. Ils ne permettent pas de savoir quelles traductions les élèves comprennent. [1, 2]

Question de recherche

Pour la traduction de l’anglais vers le créole haïtien, quel est l’écart entre les scores lorsque le corpus change, comparé à l’écart entre deux modèles sur le même corpus éducatif ?

Données et méthodes

Nous avons retenu un seul sens de traduction. La fiche d’OPUS donne les scores sur Tatoeba et JW300. Lent et ses collègues ont évalué OPUS et CreoleM2M sur les mêmes 1 559 lignes anglais-créole haïtien de MIT-Haiti. Le matériel éducatif multilingue provient de 60 récits de la plateforme MIT-Ayiti. Nous avons utilisé leurs résultats agrégés. Les phrases et traductions restent chez les fournisseurs d’origine. [1, 2, 3]

BLEU mesure les correspondances entre suites de mots et pénalise les traductions trop courtes. chrF compare des suites de caractères en combinant précision et rappel. La fiche d’OPUS exprime chrF sous forme de fraction. Nous avons multiplié ces valeurs par 100 pour utiliser l’échelle de CreoleVal. Aucune des deux mesures ne donne le pourcentage de phrases bien traduites. [1, 2, 4, 5]

Pour chaque mesure, nous avons calculé trois écarts. Nous avons soustrait le score éducatif d’OPUS de son score sur Tatoeba, puis sur JW300. Nous avons enfin soustrait le score éducatif d’OPUS à celui de CreoleM2M sur les mêmes textes. Les deux mesures ont été comparées séparément, chacune sur sa propre échelle.

Résultats

OPUS obtient 45,2 points BLEU sur Tatoeba, 38,3 sur JW300 et 14,7 sur MIT-Haiti. En chrF il obtient 59,2, 54,5 et 35,8. Les deux scores baissent lorsqu’on passe de l’un ou l’autre corpus de comparaison au corpus éducatif. Les écarts entre Tatoeba et ce dernier sont de 30,5 points BLEU et 23,4 points chrF. Avec JW300, ils sont de 23,6 et 18,7 points. [1, 2]

Sur MIT-Haiti, CreoleM2M obtient 22,0 points BLEU et 43,9 points chrF. Son avance sur OPUS sur ces mêmes lignes atteint 7,3 points BLEU et 8,1 points chrF. Pour chaque mesure, les deux écarts entre corpus dépassent l’écart entre modèles. Cet ordre se maintient quand on remplace Tatoeba par JW300. [1, 2]

Scores BLEU de l’anglais vers le créole haïtien, par corpus et par modèle



Un score plus élevé indique davantage de correspondances avec la référence. Les deux dernières barres portent sur les mêmes textes éducatifs. Le tableau donne aussi chrF, qui compare des caractères. Sources : CreoleVal, tableau 5 [1], et fiche d’OPUS [2].

Discussion

Une école qui choisit un outil doit connaître ses résultats sur les textes qu’elle compte traduire. Un score sur des phrases générales peut donner une impression trompeuse de ce que produira le même modèle sur des récits éducatifs. Même avec JW300 comme point de comparaison, l’écart entre les corpus dépasse celui entre les deux systèmes sur les récits.

Une prochaine évaluation pourrait soumettre des passages entiers à des enseignants haïtiens et poser deux questions. La traduction conserve-t-elle le concept ? Un élève peut-il suivre la consigne ? Lire un passage entier permettrait de détecter les erreurs de renvoi entre phrases qu’un examen phrase par phrase peut manquer.

Limites

Les corpus diffèrent par leur sujet, le choix des phrases et la formulation des références. Les écarts ne permettent donc pas d’isoler l’effet du contenu éducatif. Les sources ne garantissent pas que le logiciel et les paramètres de calcul ont été identiques dans les trois tests.

Les sources publient des mesures agrégées. Faute de traductions appariées, d’identifiants des récits et de références alignées dans cette analyse, nous n’avons ni calculé d’intervalles d’incertitude ni compté les erreurs préjudiciables. Ces systèmes et ces récits ne déterminent pas la qualité des outils actuels, des manuels ou des usages en classe.

Reproduire l’analyse

Relever les lignes Tatoeba et JW300 de la fiche d’OPUS anglais-créole haïtien [2]. Multiplier chrF par 100.

Dans le tableau 5 de CreoleVal [1], relever les deux lignes ayant l’anglais pour source et le créole haïtien pour cible. Conserver les deux scores.

Pour chaque mesure, soustraire le score d’OPUS sur MIT-Haiti à chacune des trois autres valeurs. Indiquer le nom de la mesure et arrondir à une décimale.

Tableau de données

Corpus de test	Modèle	Points BLEU	Points chrF (0–100)
Tatoeba	OPUS	45.2	59.2
JW300	OPUS	38.3	54.5
MIT-Haiti	OPUS	14.7	35.8
MIT-Haiti	CreoleM2M	22.0	43.9

Références

[1] Lent, H. et al. (2024). CreoleVal: Multilingual Multitask Benchmarks for Creoles. Transactions of the Association for Computational Linguistics, Section 4.2 and Table 5.

[2] Helsinki-NLP. OPUS-MT English-to-Haitian model card, benchmark scores for Tatoeba and JW300.

[3] Platfòm MIT-Ayiti. Educational resources in Haitian Creole.

[4] Papineni, K. et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation.

[5] Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation.

Crédits des données

Lent et al. (2024), Table 5. CAIA comparisons. CC BY 4.0