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

Combien d’exemples de créole jamaïcain rapprochent les résultats de deux modèles de langue ?

Analyse secondaire de neuf conditions d’entraînement dans le test d’inférence JamPatoisNLI

Adrian Dunkley | info@caribbeanaiassociation.com

Résumé

Un modèle entraîné en anglais conserve-t-il son avance sur un modèle multilingue lorsque les deux reçoivent des exemples annotés en créole jamaïcain ? Nous avons reconstitué les écarts de taux de bonnes réponses dans neuf conditions du test JamPatoisNLI. L’avance de RoBERTa en anglais sur XLM-RoBERTa passe de 11,50 points de pourcentage sans paires locales à 1,33 point avec 249 paires. Elle augmente légèrement à certaines étapes intermédiaires. L’étude porte sur l’inférence écrite. Elle ne mesure directement ni la reconnaissance vocale ni la qualité d’un dialogue. [1]

Question de recherche

Pour les mêmes phrases de test en créole jamaïcain, comment l’écart entre RoBERTa en anglais et XLM-RoBERTa multilingue évolue-t-il selon le nombre de paires locales annotées ?

Données et méthodes

Armstrong, Hewitt et Manning ont créé des paires de phrases pour un test à trois réponses. Le modèle lit une prémisse et un second énoncé. Il décide si ce dernier découle de la prémisse, la contredit ou reste indéterminé. Les prémisses proviennent de textes publiés, surtout sur les réseaux sociaux. Une locutrice du créole jamaïcain a rédigé les seconds énoncés. La documentation des données décrit les catégories et la répartition des exemples. [1, 2]

Nous avons extrait du tableau 9 les deux colonnes RoBERTa dont les paramètres restaient ajustables. Les deux modèles avaient déjà appris à faire de l’inférence sur des textes anglais. Nous avons retenu tous les effectifs publiés pour ces modèles. Chaque unité apporte un exemple de chacune des trois classes. Nous avons donc multiplié l’effectif indiqué par trois. Chaque taux est la moyenne publiée de trois essais sur les mêmes 200 paires réservées au test. [1] La dernière ligne indique 83 triplets, soit 249 exemples selon cette règle. Le tableau 5 recense 250 paires disponibles (84, 83 et 83 par classe), et les auteurs décrivent aussi la condition complète comme comprenant 250 exemples. Nos totaux comptent les triplets équilibrés. Le tableau ne précise pas si la paire supplémentaire a servi dans un essai. [1]

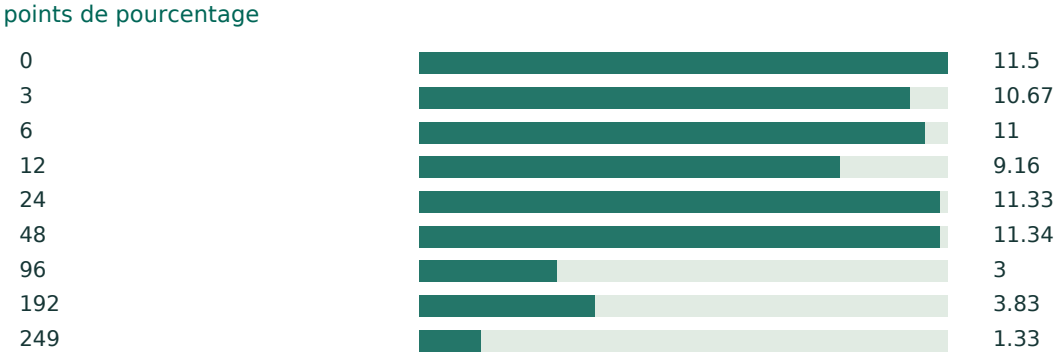
Pour chaque effectif, nous avons soustrait le taux de bonnes réponses du modèle multilingue à celui du modèle anglais. Nous avons aussi relevé le premier effectif testé auquel chaque moyenne atteint 70 % puis 75 %. Ces repères décrivent la courbe observée. Ils ne fixent aucun seuil pour une application.

Résultats

Sans paire locale, RoBERTa en anglais obtient 67,50 % et XLM-RoBERTa 56,00 %. Avec 249 paires, ils obtiennent 76,50 % et 75,17 %. Le premier gagne 9,00 points et le second 19,17. La différence entre ces gains explique la réduction de 10,17 points de l’avance initiale. [1]

L’écart atteint 11,34 points avec 48 paires, descend à 3,00 avec 96, remonte à 3,83 avec 192 puis tombe à 1,33 avec 249. Le modèle anglais atteint pour la première fois 70 % avec 48 paires et 75 % avec 192. Le modèle multilingue y parvient avec 96 et 249 paires. Un effectif intermédiaire non testé aurait pu atteindre ces niveaux. [1]

Avance du modèle anglais selon le nombre de paires locales



Les étiquettes indiquent le nombre de paires d’entraînement en créole jamaïcain. Chaque barre montre le taux de bonnes réponses de RoBERTa en anglais moins celui de XLM-RoBERTa. Le jeu de test est identique. Source : Armstrong et coll., tableau 9 [1].

Discussion

Une sélection fondée sur le seul essai sans paire locale favoriserait nettement le modèle anglais. Une équipe qui peut annoter 249 exemples observerait un écart bien plus réduit. L’étude invite à tester plusieurs budgets d’annotation avant de choisir un modèle pour cette tâche. Elle ne permet pas de choisir un modèle pour reconnaître la parole au téléphone.

La plus forte baisse entre deux conditions voisines survient entre 48 et 96 paires. Une réplication pourrait resserrer les effectifs dans cet intervalle. Tirer plusieurs échantillons d’entraînement indépendants permettrait de distinguer l’effet de leur taille de celui des exemples retenus dans chaque essai.

Limites

Les modèles ont répondu aux mêmes questions et les moyennes publiées regroupent des essais répétés. Les prédictions individuelles et les identifiants des essais seraient nécessaires pour calculer un intervalle de confiance apparié de l’écart. L’avance finale de 1,33 point reste descriptive.

La tâche porte sur un petit corpus écrit sélectionné. Le corpus ne représente pas les locuteurs de la Jamaïque selon leur paroisse, leur âge ou leurs usages. Les modèles conversationnels récents, la parole spontanée, les variantes orthographiques et l’alternance de langues demandent des tests distincts. [1, 2]

Reproduire l’analyse

Dans le tableau 9 de [1], relever les deux dernières colonnes de modèles pour toutes les lignes en conservant leur ordre.

Multiplier la première colonne par trois. Soustraire le taux de XLM-RoBERTa à celui de RoBERTa en anglais et arrondir à deux décimales.

Dans chaque colonne, repérer la première ligne qui atteint 70 %, puis celle qui atteint 75 %. Comparer les nombres totaux de paires locales.

Tableau de données

Exemples par classe	Nombre total de paires locales	Bonnes réponses de RoBERTa en anglais (%)	Bonnes réponses de XLM-RoBERTa (%)	Écart (points de pourcentage)
0	0	67.5	56.0	11.5
1	3	68.17	57.5	10.67
2	6	69.17	58.17	11.0
4	12	66.83	57.67	9.16
8	24	68.83	57.5	11.33
16	48	70.17	58.83	11.34
32	96	73.0	70.0	3.0
64	192	76.33	72.5	3.83
83	249	76.5	75.17	1.33

Références

[1] Armstrong, R.-A., Hewitt, J. and Manning, C. (2022). JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset. Findings of EMNLP, Tables 5 and 9.

[2] Armstrong, R.-A. JamPatoisNLI dataset and documentation.

Crédits des données

Armstrong, Hewitt & Manning (2022), Table 9. CAIA calculations. CC BY 4.0