Workin piepa · In-house review | v1.0 | 2026-11-20

Di text weh yuh pick fi di test change di score fi translation from English to Haitian Kreyòl

A comparison a published score fi everyday sentence, religious text an story fi school

Adrian Dunkley | info@caribbeanaiassociation.com

Wa di study find

Di same translation system can get different score when di evaluator change weh it haffi read. Wi compare di published English-to-Haitian Kreyòl result fi OPUS pon Tatoeba, JW300 an MIT-Haiti. Den wi compare two model pon MIT-Haiti alone. Wid di school story dem in place a Tatoeba, di OPUS score drop by 30.5 BLEU point an 23.4 chrF point. When wi switch model pon di school story dem, di score change by 7.3 an 8.1 point inna dat order. Dem figure show a risk inna how a school plan im translation test. Dem cyaan tell wi which translation di student dem understand. [1, 2]

Weh di risorch a aks?

When wi test translation from English to Haitian Kreyòl, how much di score change wid di set a text weh wi pick, compared wid changing di model pon di same school material?

Data an metad

Wi keep di translation going inna one direction. Di OPUS model card give English-to-Haitian Kreyòl score fi Tatoeba an JW300. Lent an colleague dem report OPUS an CreoleM2M result pon di same 1,559 English-to-Kreyòl line from MIT-Haiti. Dem put together di education material inna more than one language from 60 story pon di MIT-Ayiti platform. Wi use di published overall score. Di original provider dem keep di sentence an di translation dem. [1, 2, 3]

BLEU check how much group a word match up, an it tek off point when a translation too short. chrF compare group a character. It check how much a di translation fit di reference an how much a di reference show up inna di translation. Di OPUS card print chrF as a fraction. Wi multiply di fraction dem by 100 so di scale match CreoleVal. Neither score tell wi di percentage a sentence weh translate right. [1, 2, 4, 5]

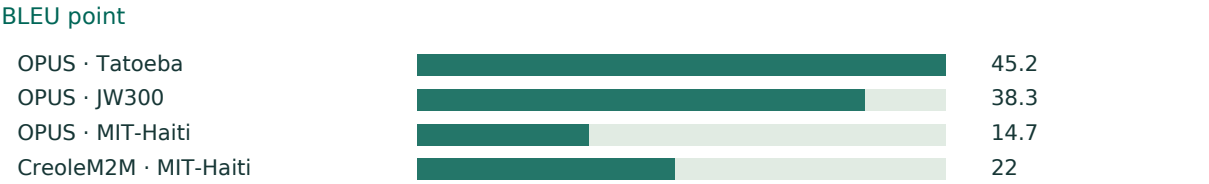
Fi each measure wi work out three gap. Wi tek weh di OPUS education score from im Tatoeba score, den do di same wid JW300. Wi also tek weh di OPUS education score from di CreoleM2M score pon di same material. Wi compare each measure by itself pon im own scale.

Rizolt dem

OPUS get 45.2 BLEU pon Tatoeba, 38.3 pon JW300 an 14.7 pon MIT-Haiti. Di chrF score dem a 59.2, 54.5 an 35.8. Both score drop when di education material tek di place a either one a di two comparison set. Di gap between Tatoeba an school material a 30.5 BLEU point an 23.4 chrF point. Wid JW300 di gap smaller: 23.6 an 18.7 point. [1, 2]

Pon MIT-Haiti, CreoleM2M get 22.0 BLEU an 43.9 chrF. Pon dem same line it lead OPUS by 7.3 BLEU point an 8.1 chrF point. Fi each measure, either change a text set gi a bigger gap than di change a model. Dat still hold when wi put JW300 in place a Tatoeba. [1, 2]

English-to-Kreyòl BLEU score by test set an model



A higher score mean more a di wording match di reference. Di last two bar use di same school material. Di table also show chrF, weh compare character. Source: CreoleVal, Table 5 [1], an di OPUS model card [2].

Wa di rizolt mean

A school weh a pick translation software need evidence pon di material it a plan fi translate. A score from a collection a everyday sentence coulda gi di wrong impression bout weh di same model woulda do wid school story. Even when wi compare wid JW300, di gap between di text set dem still bigger than di gap between di two system pon di story dem.

A next study coulda gi Haitian teacher whole passage an ask two separate question. Di translation keep di idea weh di lesson a teach? One student coulda follow di instruction? Reading di whole passage woulda help di teacher spot when a word point back to di wrong thing inna another sentence. A test weh tek one sentence at a time coulda miss dat.

Limit a di study

Di collection dem differ inna more than topic. Dem differ inna which sentence dem pick an how di reference translation word it. Dat mean di gap cyaan single out one effect weh come from school content alone. Di original report dem nuh show weh all three test use di same scoring software an setting.

Di source dem give overall score. Wi never have paired translation, story ID an matching reference together inna dis analysis, so wi never work out interval fi di uncertainty or count harmful mistranslation. Dem older model an school story cyaan settle how good di current tool dem, di textbook dem or dem use inna class is.

Run di analysis agen

Tek di Tatoeba an JW300 row from di English-to-Haitian OPUS model card [2]. Multiply chrF by 100.

Inna Table 5 a CreoleVal [1], tek di two row weh start wid English an translate to Haitian Kreyòl. Keep both score dem.

Fi each measure, tek weh di OPUS MIT-Haiti score from each a di other three row. Gi di gap to one place after di decimal point an name di measure.

Data tiebl

Text set fi testing	Model	BLEU point	chrF point (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

Suos dem

[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.

Weh di data come fram

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