

IFT6285 – Devoir 8

Traduction Automatique

Maxime MONRAT
Juan Felipe DURAN

Université de Montréal
Automne 2021

Objectifs

Nous explorons dans ce devoir le processus de traduction automatique à l'aide de modèles transformers. Nous évaluerons ce processus avec le score BLEU du corpus de test WMT-14 de la librairie *sacrebleu*.

1. Traduction à l'aide de la librairie *transformers*

Nous utilisons dans ce devoir la librairie Python *transformers* de HuggingFace, qui permet d'implémenter rapidement des modèles pré-entraînés existants. Dans notre cas nous avons utilisé le modèle par défaut de la librairie : T5-base.

Avec ce modèle, et pour une taille de phrase maximale de 120 mots (la taille maximale des phrases sources et des phrases de référence du corpus de test est de 75 mots avec une tokenisation simple (*split*), nous avons élargi un peu la fenêtre par mesure de sécurité), traduire les 3003 phrases du corpus de test prend **372 minutes**. C'est très long, mais le code tourne de manière non optimisée depuis un Notebook Jupyter directement dans l'IDE.

2. Score BLEU

2.1. Score général

En utilisant le fichier `wmt14.fr` comme référence, nous avons eu un score de **38.11** ($\mu = 38.0903 \pm 0.6644$). La signature de notre programme était **`nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0`**.

Sans réelle surprise, le nombre de phrases considérées a une influence non-négligeable sur la validité du score. La taille maximale des phrases a aussi une importance, ce qui n'est pas non plus étonnant puisque la traduction d'une phrase ne contient que rarement le même nombre de mot, et qu'une phrase incomplète perd aussi son contexte. On remarque que l'on atteint un plateau autour de 40 mots.

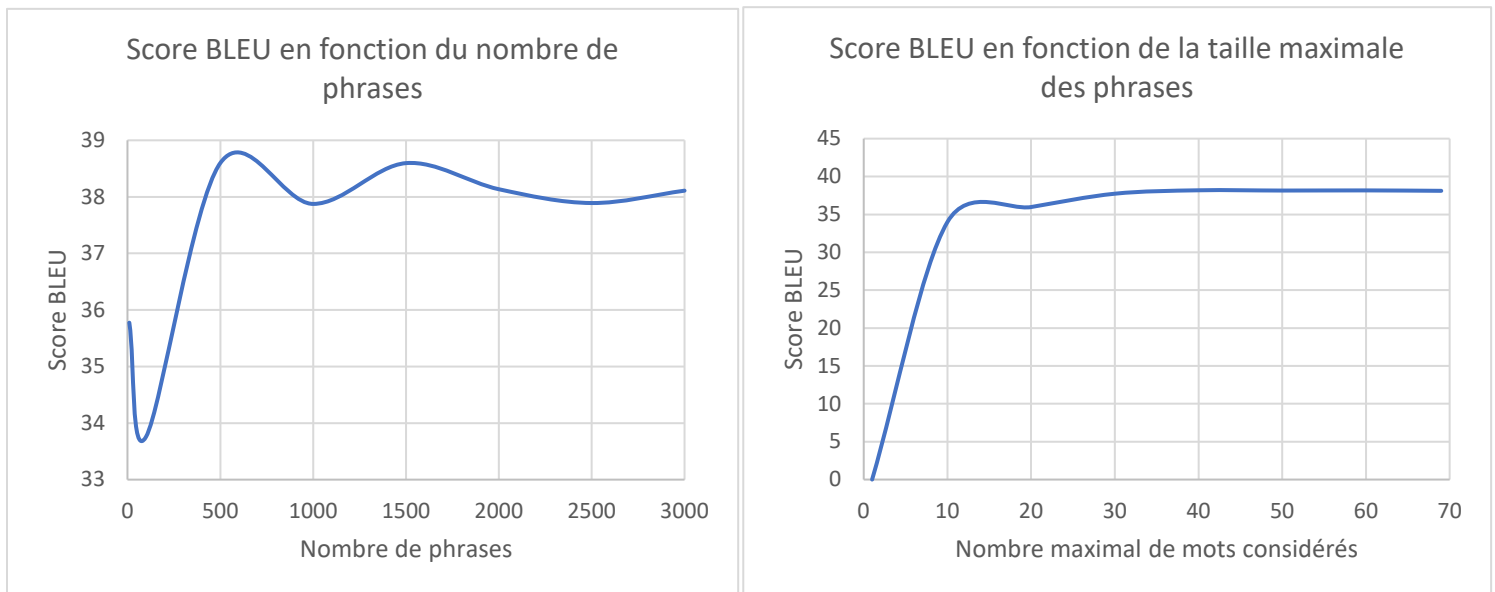


Figure 1: Influence du nombre de phrases et de leur longueur maximale sur le score BLEU

2.2. Scores individuels

En examinant de manière plus fine les scores individuels on remarque que l'étendue des valeurs est très grande (de 0 à 100). La moyenne est à **34.99**, la médiane à **32.74**, et l'écart-type est de **19.62**.

2.3. Phrases Problématiques

Voici 6 exemples de phrases dont le score BLEU est très pauvre :

Phrase 1 :

Output: Outre le contrôle, il se livre à des parties, à des femmes et à la drogue.

Reference: Sans limites, l'homme enchaîne les fêtes, les femmes, les drogues.

Score: 2.863

Phrase 2 :

Output: Trekking through mud, rivers and jungle to provide free medical care

Reference: Il marche dans la boue, la jungle et traverse des rivières pour offrir une assistance médicale gratuite

Score: 2.4426

Phrase 3 :

Output: La solennité authentique de l'histoire risque d'en faire une épreuve semblable à celle d'une chemise à cheveux pour les spectateurs.

Reference: La consciencieuse solennité de la narration pourrait également devenir une pénitence pour le public.

Score: 2.8423

Phrase 4 :

Output: Toronto Mayor Chases Off Journalists Seeking to Interview Him on Drug Affair

Reference: Le maire de Toronto chasse des journalistes venus l'interroger sur une affaire de drogue

Score: 2.8666

Phrase 5 :

Output: Les nouveau-nés ne seront plus rigide ment affectés à des hommes ou à des femmes.

Reference: Le sexe masculin ou féminin des nouveaux nés ne leur sera plus attribué, de manière rigide, à la naissance.

Score: 2.5792

Phrase 6 :

Output: Ce jour de l'Halloween, Pia a connu une soirée troublée marquée par l'éclatement d'un brawl impliquant un certain nombre de personnes autour du bureau de poste.

Reference: En ce jour d'Halloween, la soirée a été agitée du côté de Pia où une rixe impliquant plusieurs personnes a éclaté du côté de la Poste.

Score: 2.4854

Discussion:

Certaines différences qu'on remarque entre le résultat et le référence c'est l'utilisation de synonymes. Par exemple, « Outre le contrôle » et « sans limite » pour la phrase 1 et « l'histoire » et « la narration » pour la phrase 3. On peut aussi remarquer les anglicismes dans nos traductions qui ne sont pas présents dans le corpus de référence (*brawl* et *parties*. Sont-ce des anglicismes, ou bien des mots absents du vocabulaire qui n'auraient pas été traduits?). Les idiomes sont également problématiques : ce sont des phrases hautement symboliques qui ne peuvent pas se traduire de manière littérale (phrase 3).

Notons également les phrases 2 et 4 qui n'ont même pas réussi à se faire traduire : ce sont des phrases en anglais.

Répartition du score BLEU sur les données de traduction

