Evaluating Gender Bias in Machine Translation

Chatri Chansuk[1], Ahmed Dekkar[2], David El Rais[3], and Marion Mollard[4]

[1,2,3,4] *Étudiant en master II MLDS, Université de Paris, France*

INTRODUCTION

Dans un monde où les innovations sont en perpétuelle évolution, la traduction automatique a connu un essor considérable ces dernières années dans plusieurs domaines notamment dans le Web avec la possibilité de traduire automatiquement et en quelques secondes des pages Web ou des textes de plus en plus longs.

La traduction automatique désigne la traduction d'un texte (en direct ou en différé) entièrement réalisée par un ou plusieurs programmes informatiques, sans qu'un traducteur humain n'ait à intervenir.

Les performances et la rapidité des moteurs de recherches sont très bons, en effet, la barrière de la langue disparait de plus en plus et on se rapproche de la traduction en temps réel.

Cependant certaines faiblesses persistent, on peut notamment noter la présence de préjugés sexistes. Cette faiblesse existe car les modèles neuronaux sont formés sur de grands corpus de textes qui contiennent des biais et des stéréotypes ce qui fait que les modèles héritent de ces biais sociaux.

Ce qui peut réduire considérablement la qualité de traduction. Un exemple de ce préjugé sexiste est le mot ''friend" dans la phrase '' She works in a hospital, my friend is a nurse '' doit être correctement traduit par ''amiga" (amie en espagnol) en Espagnol, tandis que ''She works in a hospital, my friend is a doctor '' sera traduit incorrectement par ''amigo" (ami en espagnol) en Espagnol. Nous considérons que cette traduction contient des préjugés sexistes car elle ne tient pas compte du fait que, dans les deux cas, "ami" est féminin et traduit en se concentrant sur les stéréotypes professionnels, c'est-à-dire en traduisant médecin par homme et infirmière par femme (Escude et al., 2019).

Le reste du papier est organisé comme suit. La section 2 présente les différentes solutions existantes (état de l'art). La section 3 est un résumé de l'étude reproduite. Les résultats, les problèmes et la discussion sont inclus dans la section 4, la section 5 et la section 6. Enfin la section 7 présente les principales conclusions et idées pour la suite des travaux.

1. STATE OF THE ART

Plusieurs techniques on été développées pour réduire les préjugés sexistes, on peut notamment citer une méthode basée sur des techniques de word embeddings (Escude et al., 2019), le principe est assez simple il consiste à entrainer un ensemble de word embeddings à partir de l'algorithme GloVe. Ensuite ils suppriment le biais en utilisant une méthode de post-processing et entraine une version neutre de genre. Puis ils fusionnent l'ensemble des modèles dans un Transformer. Ils montrent une amélioration de la performance sur un jeu de données Anglais-Espagnol.

Une autre façon populaire est la mise en place d'un modèle deep learning fine tuning sur un dataset équilibré entre les genres (Scota et al., 2019). Les résultats sont intéressants si le dataset provient d'un différent domaine d'application on a une amélioration de la qualité de la traduction et peut atténuer les préjugés sexistes dans une large mesure.

2. SUMMARY OF THE REPLICATED STUDY

L'approche proposée par les auteurs utilise deux ensembles de données récents de résolution de coréférence composé de phrases anglaises qui projettent les participants dans des rôles de genre non stéréotypés (par exemple, "Le médecin a demandé à l'infirmière de l'aider dans l'opération").

Cette méthode, conçue pour huit langues cibles ayant un genre grammatical, est basée sur une analyse morphologique (par exemple, l'utilisation de l'inflexion féminine pour le mot "doctor").

Plus précisément, le processus se découpe en trois étapes. À partir de l'ensemble test WinoMT ils estiment les biais en commençant par la traduction des phrases de WinoMT. La deuxième étape consiste à aligner les traductions entre la source et la cible, en utilisant fast_align (Dyer et al., 2013). Puis à faire correspondre l'entité anglaise annotée dans les ensembles de données de la coréférence à sa traduction. La troisième étape consiste à extraire le sexe de l'entité côté cible à l'aide d'une heuristique simple par rapport à une analyse morphologique spécifique à la langue.

Dans la suite de l'article, les auteurs présentent leur mé-

thode d'évaluation et les résultats qu'ils ont obtenus. Afin de comparer leur méthode à d'autres déjà existantes, ils ont testé sur les mêmes données six autres algorithmes de traduction automatique (comme celui de Google ou de Microsoft).

Leurs résultats leur ont permis de dire que sur huit langues cibles différentes, les systèmes testés sont sensiblement enclins à traduire selon des stéréotypes sexistes plutôt qu'en fonction d'un contexte plus significatif.

3. RESULT

Dans la Table 1 nous présentons les résultats obtenus en utilisant les différents systèmes populaires (Google, Microsoft et Amazon), ainsi que celui de la méthode appliquée dans cette étude pour la traduction en français.

TABLE 1 – Résultats de l'étude

Accuracy	Google Translate	Microsoft Translator	AWS	Model [1]
Spanish	66.0	60.9	73.4	–
Italian	49.7	46.1	49.4	–
French	76.3	55.5	64.5	76.2
Ukrainian	45.5	49.4	–	–
Russian	44.3	42.2	45.6	–
Hebrew	71.7	61.5	73.5	–
Arabic	56.9	54.9	60.6	–

Les résultats que l'on a obtenu montre que la précision des traductions n'est pas affectée par le changement d'alphabet, par exemple de l'anglais vers l'hébreux on obtient des résultats aussi bons que de l'anglais vers l'espagnol, et meilleurs que vers l'italien.

4. ISSUES

Nous avons initialement choisit de reprendre l'article "Reducing Gender Bias in Neural Machine Translation as a Domain", mais après avoir rencontré plusieurs difficultés nous avons finalement travaillé sur la première version de ce projet. Les problèmes que nous avons renconté sont les suivant.

Premièrement, nous avons essayé d'exécuter le code disponible sur le "github" des auteurs pour avoir une visibilité sur le déroulement de l'exécution et le résultat final, mais des problèmes de compatibilité entre la version 1 et 2 de tensorflow nous ont causées des erreurs, comme par exemple le fait que certaines librairies soit compatibles avec la dernière version de tensorflow et d'autres avec l'ancienne.

Deuxièmement, des problèmes au niveau des installations sur les environnements créés comme "sgmn_env", "tf_env_main" qui sont deux environment différents.

Troisièmement, nous n'avons pu retrouver les scripts permettant de reproduire un nouveau modèle pour une traduction Anglais-Français.

Enfin, nous avons eu des difficultés pour comprendre la structure du projet et les liens entre les différents scripts le composant.

Le code n'est pas déployé sur Docker car nous avons des problèmes au niveau de la création d'image (installation des packages et permission).

WORK REPARTITION

La répartition du travail s'est faite principalement par équipe.

Au démarrage du projet Ahmed et Marion ont été chargé de la partie codage, tandis que David et Chatri ont été assignés à la rédaction du rapport (Introduction, State of the Art et Summary of the replicated study).

Puis suite aux difficultés rencontrées, nous avons changer d'article, et les équipes se sont interverties, Ahmed et Marion se sont occupés de la partie rapport, tandis que David et Chatri ont pris en charge la partie code.

CONCLUSION

Nous avons tenté de reproduire l'étude menée dans l'article de recherche que nous avons présenté [1].

Nous avons pu remarquer d'après nos résultats, que la traduction n'était pas impactée par un changement d'alphabet.

Notre code est disponible sur github via le lien suivant : `https://github.com/Masimisa/Projet_NLP`

RÉFÉRENCES

[1] Gabriel Stanovsky and Noah A. Smith and Luke Zettlemoyer *Evaluating Gender Bias in Machine Translation*. ACL, Association for Computational Linguistics, Florence, Italy, June 2019. `https://github.com/gabrielStanovsky/mt_gender`

[2] Joel Escud e Font, Marta R.Costa-jus *Equalizing Gender Bias in Neural Machine*, 2019.

[3] Joel Escud e Font, Marta R.Costa-jus *Fine-tuning Neural Machine.*, 2019.