Système de traduction automatique neuronale français-mongol (Historique, mise en place et évaluations)

Shuai Gao

Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Système de traduction automatique neuronale français-mongol
Historique, mise en place et évaluations

Shuai GAO[1]
(1) ERTIM, INaLCO, 2 rue de Lille, 75007 Paris, France
shuai_gao@outlook.fr

RESUME

La traduction automatique (abrégé ci-après TA) connaît actuellement un développement rapide, pendant lequel les langues peu dotées semblent pourtant moins développées. En effet, il existe moins de recherches sur ces dernières. Notamment, aucune recherche publiée n'a été trouvée sur la paire de langues français-mongol. Cet article entame une nouvelle étape dans les recherches en TA pour cette paire de langues peu dotée. Nous décrivons l'historique de la TA et en établissons un état de l'art pour le mongol. Ensuite, nous nous employons à mettre en place notre propre système de TA à partir des outils et ressources *open source*. Outre l'évaluation automatique comme méthode pour apprécier sa performance, nous concevons une méthode d'évaluation humaine originale nommée « *IFF* » permettant de mieux connaître les forces et les faiblesses de notre système par rapport à des moteurs de traduction commerciaux.

ABSTRACT

French-Mongolian Neural Machine Translation System (History, Implementation, and evaluations)
Machine Translation (hereafter abbreviated MT) is currently undergoing rapid development, during which less-resourced languages nevertheless seem to be less developed. Indeed, there is less research on the latter. Notably, no published research was found on the French-Mongolian language pair. This article begins a new stage in MT research for this less-resourced language pair. We describe the history of MT and establish a state of the art for Mongolian. Next, we endeavor to set up our own MT system using open-source tools and resources. In addition to automatic evaluation as a method for assessing its performance, we have designed an original human evaluation method called "*IFF*" to better understand the strengths and weaknesses of our system compared to commercial translation engines.

MOTS-CLES : traduction automatique neuronale, langues peu dotées, langue mongole, évaluation automatique, évaluation humaine, score BLEU

KEYWORDS: Neural Machine Translation, less-resourced languages, Mongolian language, automatic assessment, human assessment, BLEU Score

1 Introduction

De nos jours, la traduction automatique fait des progrès technologiques rapides en réponse au besoin des échanges internationaux. Néanmoins, elle ne fonctionne de manière satisfaisante que pour des langues parlées par un grand nombre de locuteurs et représentant un marché économique pour les grandes entreprises technologiques telles que l'anglais, le chinois et le français. En ce qui concerne les langues moins parlées ou minoritaires, la langue mongole dans le cas de notre article, son

Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL), pages 97–110.

développement fait face à de nombreux défis car il existe moins de recherches dessus, par exemple, à notre connaissance, aucune recherche n'a été publiée sur la paire de langues français-mongol.

L'objectif de notre travail est donc d'entamer une nouvelle étape dans les recherches en traduction automatique pour cette paire de langues peu dotée, tout en utilisant les techniques d'apprentissage automatique, les outils et les données *open source* existants afin de construire notre propre système de traduction français-mongol.

Nous proposons d'abord en section 2 un récapitulatif de l'historique de la TA et nous établissons en section 3 un état de l'art de la TA du mongol. Ensuite, nous décrivons en section 4 une expérimentation qui consiste à entraîner des modèles de traduction automatique neuronale pour la paire de langues français-mongol à partir d'un corpus parallèle *open source* récupéré sur *OPUS*[1]. Les modèles sont appris avec l'outil *OpenNMT*[2]. Après cela, nous réalisons deux évaluations automatiques et concevons une méthode d'évaluation humaine nommée « *IFF* » pour rapporter en section 5 les forces et les faiblesses de nos modèles de traduction par rapport à des moteurs commerciaux.

2 Historique de la traduction automatique

Le développement de la traduction automatique a connu des hauts et des bas. Nous proposons dans la FIGURE 1[3] un historique des travaux importants avec les méthodologies[4] correspondantes. À l'heure actuelle, les méthodes à base de *Transformer* sont l'état de l'art dans le domaine de la TA en termes de qualité et d'efficacité. Il est à noter que pour l'instant les travaux qui suivent se basent fondamentalement sur *Transformer* et n'apportent pas de grands changements sur l'architecture.

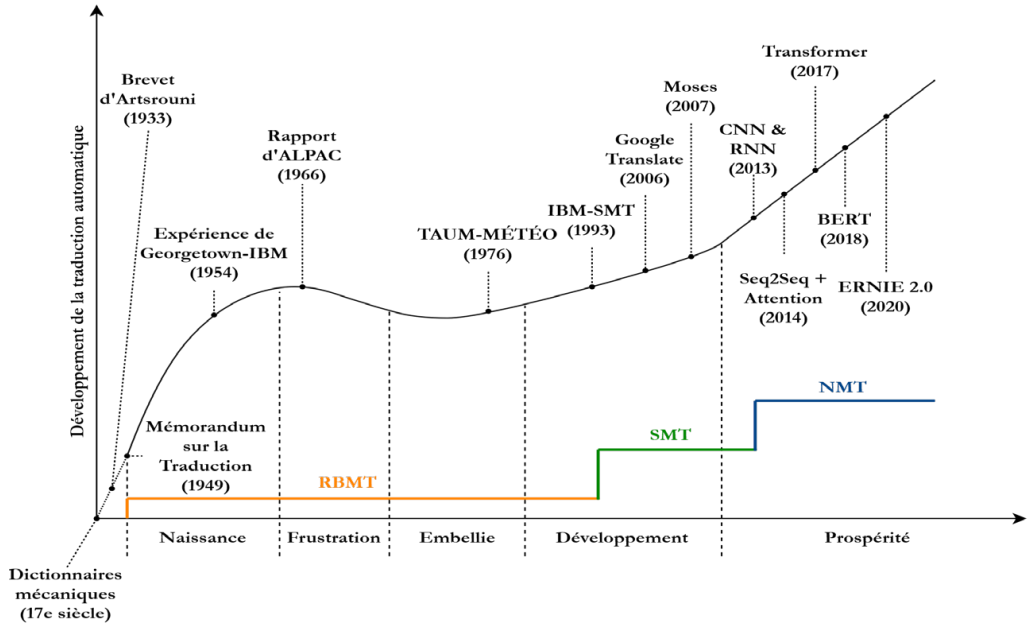

FIGURE 1 : Historique de la traduction automatique

[1] *OPUS* : *The Open Parallel Corpus*, disponible sur : `https://opus.nlpl.eu/`

[2] *OpenNMT*, disponible sur : `https://opennmt.net/`

[3] *Nota bene* : La FIGURE 1 est un des fruits de notre travail après la lecture de nombreux travaux liés. Ceci a pour but de visualiser l'évolution de la TA.

[4] Méthodologies : Rule-Based MT (RBMT), Statistical MT (SMT) et Neural MT (NMT)

3 État de l'art de la TA du mongol

3.1 Langue mongole

La langue mongole (cyrillique : Монгол хэл ; classique : ᠮᠣᠩᠭᠣᠯ ᠬᠡᠯᠡ) appartient au groupe mongolique de la famille altaïque, l'une des principales familles linguistiques d'Asie centrale et du nord-est (Janhunen, 2005). Parlée par plus de 7 millions de locuteurs, elle fait référence à plusieurs langues et dialectes similaires. La plupart des variétés de la langue mongole partagent quelques caractéristiques importantes, telles que l'ordre des phrases SOV (sujet-objet-verbe), l'agglutination et l'harmonie des voyelles (Gaunt, 2004).

La variété Khalkha (en écriture cyrillique) sert aujourd'hui de langue officielle pour la République populaire de Mongolie, soit l'ancienne région chinoise de la Mongolie extérieure. D'autres variétés très similaires (en écriture classique, aussi appelée *Mongol Bichig*) sont utilisées par les Mongols vivant dans la région autonome chinoise de la Mongolie intérieure avec un statut officiel de deuxième langue de la région après le mandarin (Janhunen, 2005).

3.2 Recherches sur la TA

Nous avons évoqué dans l'introduction le manque de recherches sur la TA du mongol, en particulier pour la paire de langues français-mongol. En effet, nous constatons que jusqu'à présent aucune recherche sur la TA n'a été publiée pour cette paire, alors qu'il existe des travaux pour les paires mongol-anglais et mongol-chinois qui servent de références.

Ochir and Serguleng (2003) ont développé un système de RBMT anglais-mongol (classique) et l'ont amélioré avec une méthode basée sur des corpus tandis que Hou and Liu (2007) ont proposé une méthode de traduction mongol-chinois basée sur des exemples et que Su (2014) a, de son côté, étudié les caractéristiques de la paire mongol-chinois et a proposé un modèle de SMT mongol-chinois basé sur l'approche hiérarchique (*Hierarchical Phrase-based System*).

Ensuite, une série de travaux sur la traduction automatique neuronale mongol-chinois ont été menés avec le développement de l'apprentissage profond. Wu (2017), Shen (2017) et Wang (2018) ont respectivement réalisé leur modèle de NMT mongol-chinois avec RNN, *Word2Vec* et CNN, et des gains de performances ont été constatés. Ensuite, FAN et al. (2018) ont proposé un modèle avec de l'information préalable permettant d'enrichir des *features* du modèle pour améliorer la performance. Liu (2018) a proposé un modèle de NMT LSTM basé sur le codage de morphèmes, qui s'avère être plus performant, surtout en termes de dépendance à long terme. Cao (2020) a exploré l'application du corpus monolingue en tant que complément du corpus parallèle dans la NMT mongol-chinois et a grandement amélioré la performance. Wang et al. (2020) ont transféré des paramètres d'un modèle de NMT anglais-chinois préentraîné à un autre système de NMT mongol-chinois, les résultats montrent que cette stratégie semble améliorer la qualité de traduction.

3.3 Corpus

Le développement de la TA est étroitement lié à la construction des corpus. En 1990, l'Institut de recherche sur la langue mongole de l'Université de Mongolie intérieure a construit la Base de données sur la langue mongole moderne qui comprend 1,2 millions de mots en écriture classique (Hua, 1997). Jaimai and Chimeddorj (2008) ont construit de façon semi-automatique un corpus à 5

millions de mots. Zhang (2009) a mis en place une plate-forme expérimentale qui permet de maintenir un corpus bilingue chinois-mongol et d'y rechercher des informations selon les étiquettes sur les caractéristiques linguistiques des phrases. Bao (2016) a créé à partir des émissions télévisées et des actualités un corpus mongol classique qui compte 1,75 millions de mots. La *China Conference on Machine Translation* (CWMT) crée des corpus parallèles mongol-chinois pour l'évaluation annuelle de la TA à l'échelle nationale, lesquels sont renouvelés à chaque édition du *workshop* (Yang et al., 2019). Крылов (2017) a développé *Mongolian Corpus* permettant d'étudier le lexème, le morphème et la distance entre les mots.

Tous ces efforts de recherche ont ainsi peu à peu doté la langue mongole de corpus. Selon Fei et al. (2019), il en existe aujourd'hui en classique et en cyrillique, pour de genres variés (livres anciens, historique, proverbes, expressions idiomatiques, corpus parallèles mongol-chinois, mongol-russe, mongol-anglais, mongol-japonais, etc.). Par ailleurs, un corpus mongol classique de 200 millions de mots est en cours de construction (Fei et al., 2019). Il est pourtant à noter qu'aucun des corpus susmentionnés n'est *open source*. Ils sont en effet, soit les propriétés intellectuelles de certaines organisations protégées par le droit d'auteur, soit les données exclusives accessibles uniquement sur l'inscription pour l'évaluation organisée par CWMT (2017) (toute utilisation en dehors de son cadre est sanctionnée). En conséquence, il ne serait possible de trouver des corpus mongols accessibles que sur *OPUS*, la collection des corpus parallèles *open source* (Tiedemann, 2012), et en volume limité.

3.4 Moteurs, plates-formes et outils

Google Traduction[5] est aujourd'hui sans conteste le moteur de traduction le plus utilisé dans le monde. Ce succès est dû non seulement à la gratuité et à la précision de l'outil, mais aussi au fait qu'aucun autre moteur ne le surpasse en termes de prise en charge des langues, y compris minoritaires. Ainsi, nous y retrouvons notre paire de langues (français-mongol cyrillique), que seuls deux autres moteurs commerciaux payants incluent *NiuTrans*[6] (小牛翻译) et *PoliLingua*[7]. Toujours est-il qu'aucune publication n'existe pour décrire le fonctionnement ou l'évaluation du système pour cette paire de langues en particulier.

Certaines entreprises en Mongolie intérieure ont développé des plates-formes et outils pour promouvoir le développement de la TA du mongol :

- *Oyun*[8] (奥云) est une plate-forme multifonctionnelle pour chinois-mongol (classique et cyrillique), qui fournit la TA textuelle et vocale, l'OCR, la synthèse de la parole, etc. ;
- *Yijinyun*[9] (毅金云) est une autre plate-forme similaire à *Oyun* offrant presque les mêmes fonctionnalités au niveau logiciel, mais elle se concentre davantage sur le développement des matériels tels que les enceintes intelligentes, les robots et les traducteurs portables ;
- *Menksoft*[10] (蒙科立) est une entreprise clé pour la numérisation du mongol qui se concentre sur la méthode de saisie, les polices, l'encodage et la bureautique en écriture classique (Ao et al., 2011).

---

5 *Google Traduction*, disponible sur : `https://translate.google.fr/`

6 *NiuTrans*, disponible sur : `https://niutrans.com/`

7 *PoliLingua*, disponible sur : `https://www.polilingua.com/`

8 *Oyun*, disponible sur : `https://www.nmgoyun.com`

9 *Yijunyun*, disponible sur : `https://www.mengguyu.cn/`

Les défis auxquels est confrontée la TA du mongol peuvent être attribués aux facteurs suivants : manque de ressources (*open source* en particulier), encodage incohérent (du mongol classique), création difficile des corpus parallèles à grande échelle, recherche insuffisante sur la langue mongole (Wu, 2017). Notre travail entend justement pallier le manque de recherche et revitaliser la langue mongole en la reliant à une langue principale, le français.

4 Expérimentation

Nous avons choisi l'outil *OpenNMT* (Klein et al., 2020) comme cadre de développement (*framework*) en raison de sa gratuité et du fait qu'il soit implémenté à base de *Transformer*. Nous avons décidé de travailler sur le corpus parallèle français-mongol (cyrillique) au format .tmx disponible sur le site *OPUS* (Tiedemann, 2012) dans le projet *MultiCCAligned v1.1*[11]. Nous avons préparé le corpus aux exigences d'*OpenNMT*. Nous remarquons que le corpus a été construit automatiquement et qu'aucune correction manuelle n'a été effectuée. Après avoir survolé le contenu du corpus, nous nous sommes rendu compte de sa qualité médiocre : il existe beaucoup de phrases incomplètes ; certaines entrées ne sont pas normalisées et ni même alignées. Une correction humaine serait pourtant bien trop longue à réaliser, nous avons donc conservé le corpus en l'état pour ce premier travail.

Le but de notre expérimentation est donc de construire notre propre système de traduction français-mongol à partir de ces ressources et d'outils *open source*, avant d'évaluer sa performance avec des métriques automatiques et humaines pour estimer ses forces et ses faiblesses par rapport à des moteurs commerciaux. Dans notre expérimentation, nous voudrions également mesurer l'influence des tailles relatives des ensembles d'entraînement, de validation et de test sur la qualité des résultats mesurés. Pour répondre à cette question, nous paramétrons deux rapports de division (abrégés « rd » ci-après) pour diviser respectivement ces deux corpus monolingues en « train », « val » et « test » (TABLE 1). Nous discuterons des résultats obtenus en section 5.

Les trois schémas (FIGURES 2, 3 et 4) illustrent en détail l'expérimentation, qui s'est déroulée en réalité étape par étape. Nous avons obtenu le résultat d'une évaluation avant d'entreprendre la suivante. Les résultats seront réunis à la fin de cette section et seront discutés en section 5.

Rapport de division	train	val	test
6 : 2 : 2	172 892	57 725	57 725
7 : 1 : 2	201 754	28 861	57 725

TABLE 1 : Jeu de données (en paires de phrases)

L'entraînement a été fait avec la configuration de base[12] sur un serveur distant fourni par notre laboratoire. Nous obtenons quatre modèles de traduction décrits dans la TABLE 2, à partir desquels sont générées les traductions. Ces prédictions sont ensuite fournies à deux évaluations automatiques (FIGURE 2 ; FIGURE 3) basées sur l'évaluateur BLEU de *Tilde*[13] et une évaluation humaine originale que nous avons nommée « *IFF* » (FIGURE 4). L'évaluation automatique (1) porte sur la totalité du

[10] *Menksoft*, disponible sur : `http://www.menksoft.com/`

[11] *MultiCCAligned v1.1*, disponible sur : `https://opus.nlpl.eu/MultiCCAligned-v1.1.php`

[12] Configuration de base, disponible sur : `https://opennmt.net/OpenNMT-py/examples/Translation.html`

[13] *Interactive BLEU Score Evaluator*, disponible sur : `https://www.letsmt.eu/Bleu.aspx`

test set tandis que l'évaluation automatique (2) et l'évaluation humaine portent sur une sélection manuelle des 50 phrases complètes. Cette sélection joue un rôle d'assurance qualité eu égard à la qualité médiocre du corpus d'*OPUS*, mais a aussi l'avantage d'être appropriée pour une évaluation humaine étant donné sa taille réduite.

Dénomination	Rapport de division	Nombre d'itérations
m1_rd1_5000	6 : 2 : 2	5 000
m2_rd1_50000	6 : 2 : 2	50 000
m3_rd2_50000	7 : 1 : 2	50 000
m4_rd2_100000	7 : 1 : 2	100 000

TABLE 2 : Modèles de traduction sauvegardés

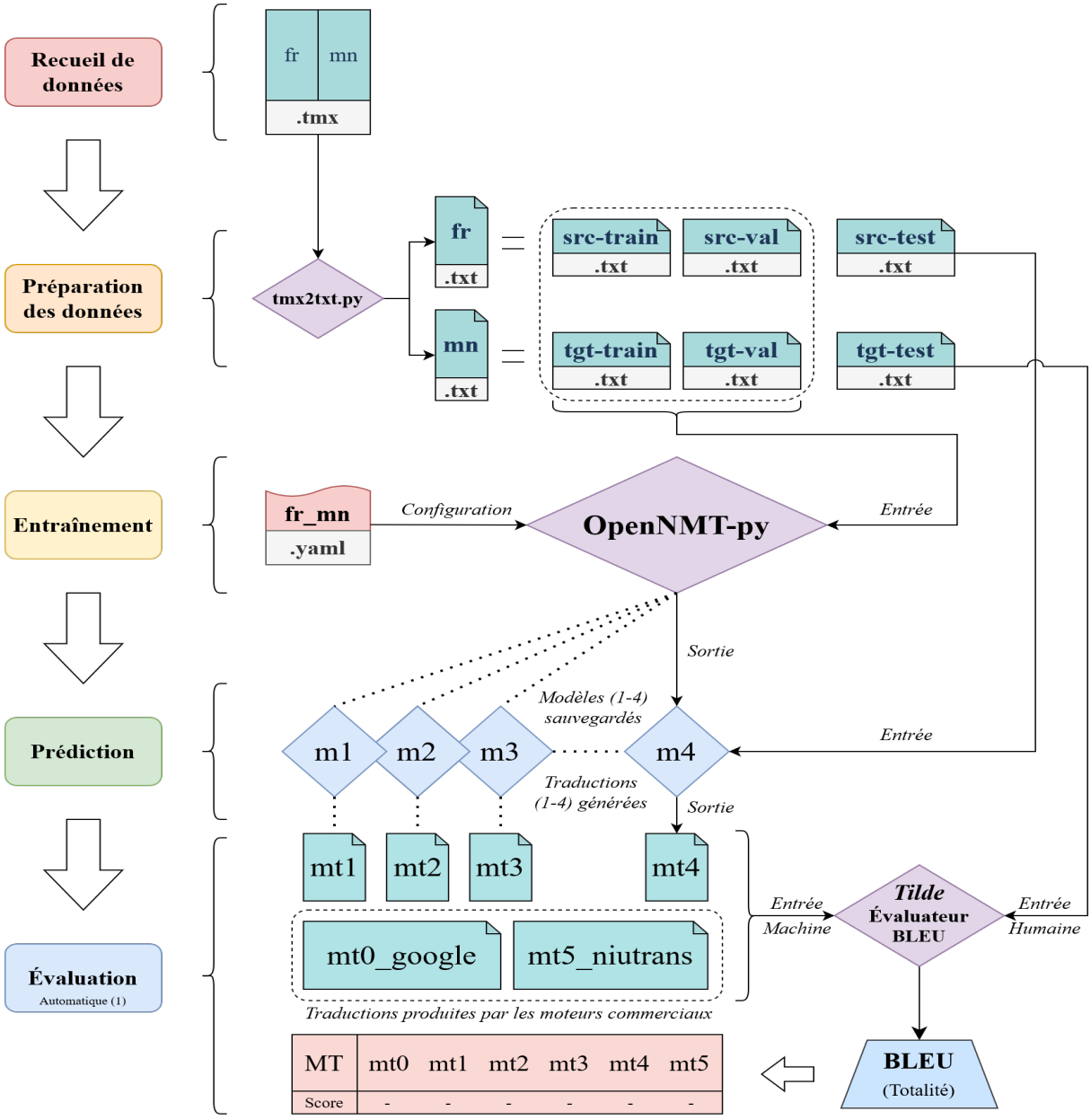

FIGURE 2 : Schéma général de l'expérimentation

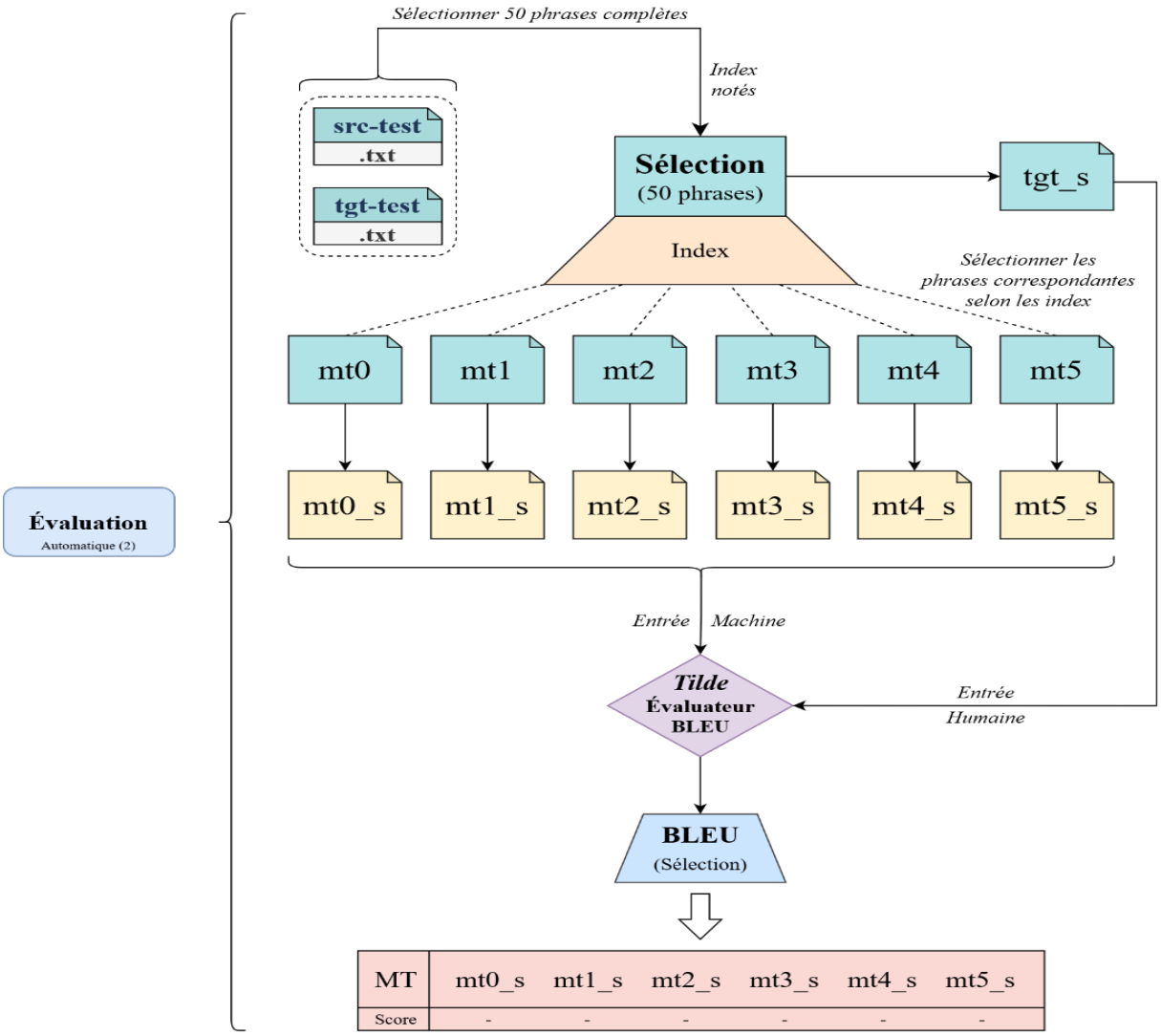

FIGURE 3 : Schéma de l'évaluation automatique (2) pour la sélection du *test set*

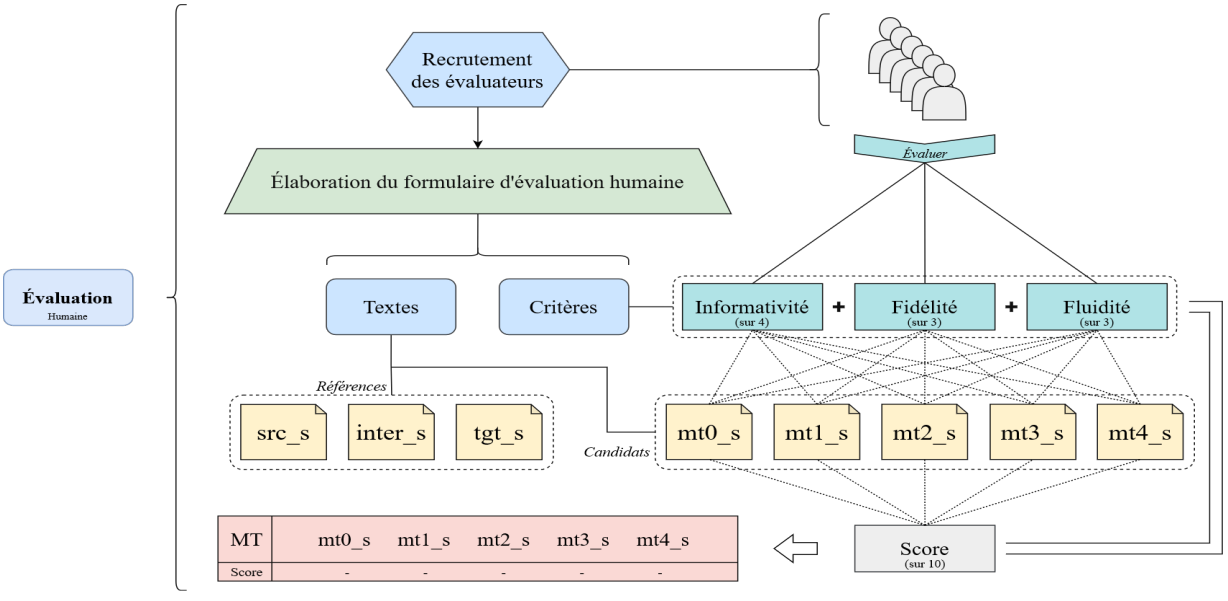

FIGURE 4 : Schéma de l'évaluation humaine « *IFF* » pour la sélection du *test set*

Pour l'évaluation humaine, nous avons recruté six locuteurs natifs (bilingues français-mongol ou bilingues chinois-mongol). Il est à noter que les quatre évaluateurs ont appris l'écriture cyrillique de manière tardive dans leurs parcours scolaires ou professionnels, et qu'ils utilisent plutôt l'écriture traditionnelle. Ce point est pris en considération quand nous mettrons en commun leurs retours. Afin d'aider les non-francophones à comprendre le contenu du texte de source, nous ajoutons la traduction automatique en chinois comme une version intermédiaire.

Ensuite, nous avons élaboré le formulaire d'évaluation (TABLE 3) à partir de la sélection. Le formulaire se compose de deux parties : textes et critères. Notre méthode d'évaluation est appelée « *IFF* », qui est un sigle de « Informativité », « Fidélité » et « Fluidité », trois attributs que nous proposons pour évaluer une traduction. Cette définition est inspirée par d'autres campagnes d'évaluation où l'on utilise certaines critère similaires, comme ce que décrivent Carroll (1966), White et al. (1994) et Miller and Vanni (2005).

评测 Évaluation / 文本代号 Nom des textes	文本 5 Texte 5	信息度 Informativité	忠实度 Fidélité	流畅度 Fluidité	总分 Score
src_s	Tous les transferts et téléchargements sont cryptés à l'aide du cryptage SSL 256 bits. Ainsi, les données de vos documents Excel et PDF ne seront pas susceptibles d'être soumises à un accès non autorisé.				
inter_s	所有传输和下载均使用 256 位 SSL 加密进行加密。这将防止您的 Excel 和 PDF 文档中的数据受到未经授权的访问。				
tgt_s	Бүх байршуулалт болон таталтыг 256 битийн SSL кодчилол ашиглан шифрлэдэг. Үүнийг хийснээр Excel болон PDF баримтаас өгөгдөл нь зөвшөөрөлгүй хандалтанд өртөмтгий биш болно.				
mt0_s	Бүх дамжуулалт, татан авалтыг 256 битийн SSL шифрлэлт ашиглан шифрлэдэг. Энэ нь таны Excel болон PDF баримт бичиг байгаа өгөгдлийг зөвшөөрөлгүй хандахаас урьдчилан сэргийлэх болно.	4	2	3	9
mt1_s	Бүх байршуулалт болон таталтыг 256 битийн SSL кодчилол ашиглан шифрлэдэг. Үүнийг хийснээр PDF болон <unk> өртөмтгий биш болно.	2	1	1	4
mt2_s	Бүх байршуулалт болон таталтыг 256 битийн SSL кодчилол ашиглан шифрлэдэг. Үүнийг хийснээр PDF болон PPT <unk> өгөгдөл нь зөвшөөрөлгүй <unk> өртөмтгий биш болно.	3	2	2	7
mt3_s	Бүх байршуулалт болон таталтыг 256 битийн SSL кодчилол ашиглан шифрлэдэг. Үүнийг хийснээр PDF болон <unk> өгөгдөл нь зөвшөөрөлгүй <unk> өртөмтгий биш болно.	3	2	2	7
mt4_s	Бүх байршуулалт болон таталтыг 256 битийн SSL кодчилол ашиглан шифрлэдэг. Үүнийг хийснээр PDF болон <unk> өгөгдөл нь зөвшөөрөлгүй <unk> өртөмтгий биш болно.	3	2	2	7

TABLE 3 : Exemple d'un groupe du formulaire d'évaluation « *IFF* »

Il y a au total 50 groupes de textes dans le formulaire. Chaque groupe contient huit éléments :

1. **src_s**[14] (texte source en français)
2. **inter_s** (texte intermédiaire en chinois)
3. **tgt_s** (texte cible en mongol cyrillique)
4. **mt0_s** (traduction de *Google* en mongol cyrillique)
5. **mt1_s** (traduction N°1 en mongol cyrillique)
6. **mt2_s** (traduction N°2 en mongol cyrillique)
7. **mt3_s** (traduction N°3 en mongol cyrillique)
8. **mt4_s** (traduction N°4 en mongol cyrillique)

Les trois premiers éléments constituent la référence et ne participent pas à la notation tandis que les cinq derniers éléments sont les candidats et seront notés selon leurs attributs : « Informativité »,

[14] « s » pour « sélection »

« Fidélité » et « Fluidité ». Ces trois derniers et la grille de notation constituent notre stratégie d'évaluation (TABLE 4).

Score	Critère		
	Attributs		
	Informativité —— Quantité d'informations de la référence que contient la traduction	Fidélité —— Degré de similarité du sens de la traduction à celui de la référence	Fluidité —— Fluidité de la traduction
0	La traduction ne contient **aucune information** contenue dans la référence.	Le sens de la traduction **n'a rien à voir** avec la référence.	La traduction **n'a pas de syntaxe** et elle **n'est pas du tout fluide**.
1	La traduction contient **peu d'informations** dans la référence.	Le sens de la traduction est **quelque peu lié** à la référence.	La traduction **a la syntaxe**, mais elle **n'est pas très fluide**.
2	La traduction contient **certaines informations** contenues dans la référence.	Le sens de la traduction est **à peu près le même** que la référence.	La traduction **est conforme à la syntaxe** et elle est **en grande partie fluide**.
3	La traduction contient **la plupart des informations** contenues dans la référence.	Le sens de la traduction est **exactement le même** que la référence.	La traduction est **conforme à la syntaxe** et elle est **parfaitement fluide**.
4	La traduction contient **toutes les informations** contenues dans la référence.		

TABLE 4 : Stratégie d'évaluation *IFF*

Notre stratégie d'évaluation, dérivée du MOS (*Mean Opinion Score*) (Viswanathan & Viswanathan, 2005), est comme suit : le score complet est de 10. À noter que la totalité des points n'est pas la même pour ces trois attributs : 4 pour « Informativité », 3 pour « Fidélité », 3 pour « Fluidité ». Étant donné la qualité du corpus, nous avons l'intuition des performances médiocres de nos modèles entraînés, et nous considérons donc que pour nos modèles, c'est la quantité des informations contenues dans la traduction qui importe plus ; d'où le privilège de l'informativité avec un point de plus.

L'évaluation humaine s'avère être une tâche très coûteuse, pendant laquelle se sont produits quelques imprévus :

– Nous n'avons malheureusement pas pu utiliser toutes les évaluations humaines. En effet, nous avons dû éliminer les scores donnés par deux évaluateurs car leurs scores font preuve d'un optimisme extrême et d'une fausseté évidente peut-être faute d'une bonne maîtrise de l'écriture cyrillique.
– L'ajout de « mt5 » a été effectué à la fin de notre expérimentation et après que l'évaluation humaine avait été terminée. Avant, nous n'avions pas connaissance de l'existence de *NiuTrans* donc ne savions pas que ce moteur de traduction prenait en compte la rare paire de langues français-mongol. En conséquence, nous n'avons pas eu le temps de l'ajouter pour l'évaluation humaine, mais que pour l'évaluation automatique.

En ce qui concerne les évaluations automatiques, il convient de faire référence aux scores BLEU cumulatifs à 1-4 gram lors de la description de la performance d'un système de génération de texte (Brownlee, 2017). Les scores BLEU se présentent sous forme de pourcentage. Afin que les scores de l'évaluation humaine soient aussi clairs que les scores BLEU, nous les avons transformés pour chaque attribut (4-informativité, 3-fidélité, 3-fluidité) également en pourcentage.

Les FIGURES 5, 6 et 7 montrent les résultats de nos trois évaluations.

Même si l'évaluation automatique est essentielle pour l'entraînement et le développement (Nakhlé, 2021), pour obtenir une appréciation finale sur la performance d'un modèle, l'évaluation humaine est toujours considérée comme déterminante dans la communauté scientifique (Bojar et al., 2016).

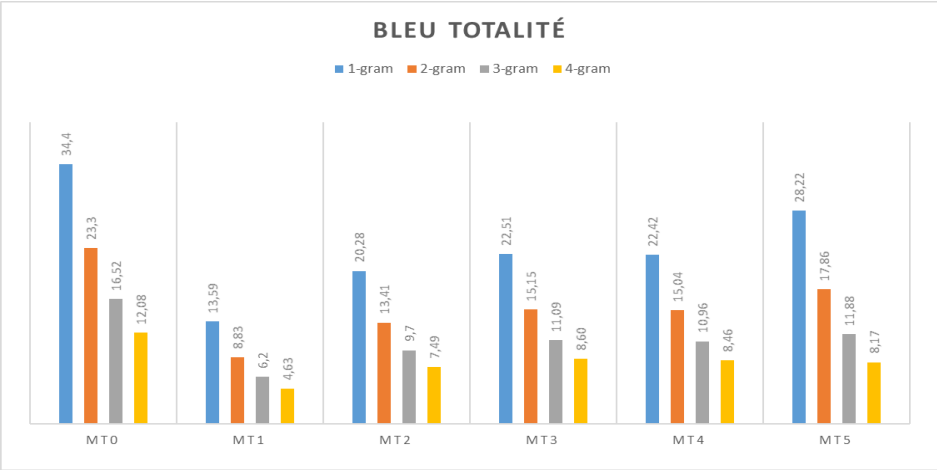

FIGURE 5 : Comparaison des scores BLEU pour la totalité du *test set*

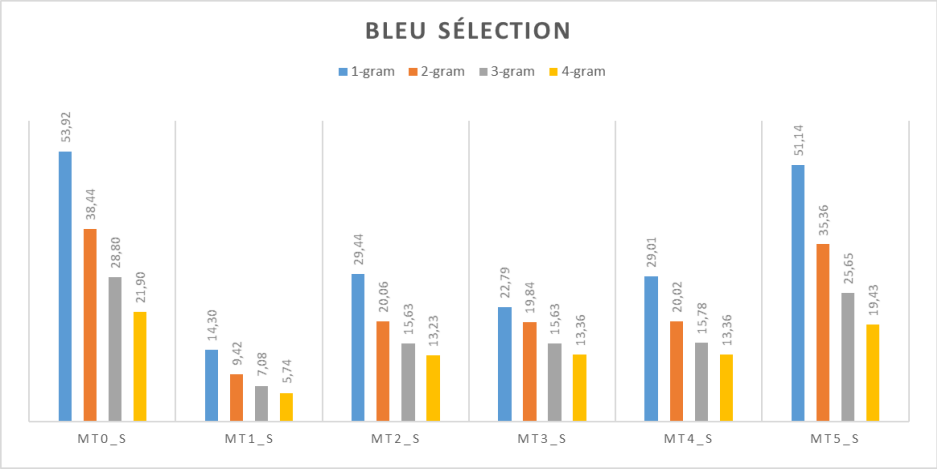

FIGURE 6 : Comparaison des scores BLEU pour la sélection du *test set*

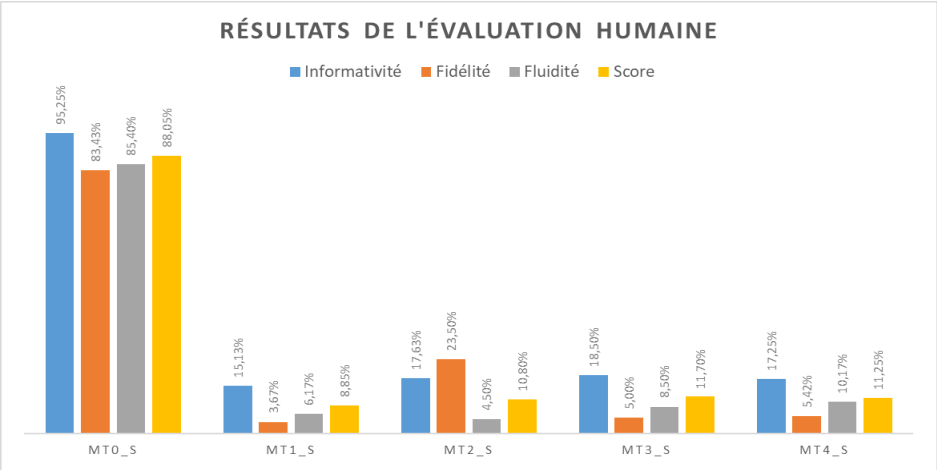

FIGURE 7 : Comparaison des scores *IFF* pour la sélection du *test set*

5 Discussion

Afin d'obtenir une appréciation la plus exhaustive possible sur les performances de nos modèles, nous avons effectué trois évaluations, deux automatiques et une humaine. Ceci suit en fait un ordre de priorités : nous effectuons d'abord une évaluation automatique avec la métrique de notre choix, le score BLEU, qui peut donner des retours tout de suite sans investissement humain ; si les résultats sont satisfaisants, nous concevons ensuite une méthode d'évaluation humaine pour savoir comment les locuteurs natifs trouvent les traductions produites par nos modèles par rapport à celles produites par les moteurs de traduction à usage commercial, ce qui prend beaucoup de temps et requiert un grand investissement humain.

L'évaluation automatique (1) en FIGURE 5, même si les scores restent bas, donne des résultats encourageants car nous constatons une petite progression prometteuse entre « m1 » et « m4 », et les scores de « m3 » et de « m4 » ne sont pas très loin de *Google* et, supérieurs à l'autre moteur de TA payant, *NiuTrans*. Notez que « m1 » ne sert qu'à tester le bon fonctionnement du *framework*.

Ensuite, nous sommes passés à l'évaluation humaine, dans laquelle nous ne pouvions inclure qu'une petite partie du corpus étant donné son caractère chronophage. C'est pourquoi nous avons ensuite préparé une sélection. Lors de la création manuelle de la sélection, nous avons réalisé que la qualité du corpus est médiocre (phrases incomplètes, entrées non normalisées ounon alignées. Aussi avons-nous effectué une évaluation automatique (BLEU) et humaine (*IFF*) sur une sélection de 50 phrases complètes, dont la qualité est assurée après une vérification humaine.

Dans l'évaluation automatique (1) en FIGURE 5, bien que les scores de nos certains modèles soient proches de *Google*, tous les scores BLEU-4 se situent autour de 10. Selon l'interprétation du score BLEU proposée par *Google*[15], les traductions produites par nos modèles sont toutes les « traductions presque inutiles », ce qui est possiblement dû à la mauvaise qualité du corpus utilisé lors de l'entraînement. Dans l'évaluation automatique (2) en FIGURE 6, l'ensemble des scores a beaucoup augmenté tandis que les écarts des scores se sont élargis : les scores de deux moteurs à usage commercial sont devenus assez proches, autour de 20, qui indique que « l'idée générale apparaît clairement, mais le texte comporte de nombreuses erreurs grammaticales » ; les scores de nos modèles utiles (« mt1 » excepté) sont autour de 13 qui signifie « l'idée générale est difficilement compréhensible ».

À partir des résultats de ces deux évaluations automatiques, nous pouvons donc affirmer que l'amélioration de la qualité du *test set* (de la totalité à la sélection) fait ressortir l'écart des performances entre les moteurs de traduction à usage commercial et nos modèles. De plus, si l'on regarde les différences entre nos modèles entraînés, nous pourrions dire que l'augmentation du nombre d'itérations semble à l'origine améliorer considérablement la performance (« m1 et m2 »), mais cela n'a pas changé grand-chose après 50 000 étapes (« m3 et m4 ») ; et que l'augmentation du train set (rapports de division, 6 : 2 : 2 contre 7 : 1 : 2) peut améliorer légèrement la performance.

Pour ce qui est de l'évaluation humaine, nous avons comparé nos modèles avec le moteur de *Google*, mais n'avons pas eu le temps d'inclure « mt5_s » de *NiuTrans* comme expliqué ci-dessus. Mais étant donné les résultats de l'évaluation automatique (2) (« m0 » est proche de « m5 »), nous pourrions supposer que pour la sélection d'une bonne qualité, les résultats de l'évaluation humaine de *NiuTrans* seront probablement similaires à ceux de *Google*.

[15] Interprétation du score BLEU de *Google*, disponible sur : `https://cloud.google.com/translate/automl/docs/evaluate#bleu`

Selon les résultats *IFF*, nous constatons que *Google Traduction* surpasse tous nos modèles avec une supériorité écrasante. Les jugements humains sont sans appel, nos modèles ne rivalisent pas avec celui de *Google*, ce qui n'est pas surprenant mais est encore plus visible lors de cette évaluation. Cela confirme également l'intérêt de l'évaluation humaine, bien plus sévère que les scores BLEU. Notons que dans le formulaire d'évaluation humaine, il arrive que nos modèles (surtout « m3 » et « m4 ») omettent de traduire certaines entrées, ce qui a eu des effets négatifs sur les scores *IFF*.

Somme toute, les résultats de l'évaluation humaine sur la sélection du *test set* sont frustrants, mais ne sont pas surprenants vu que nous nous comparons à un système mis au point par une grande entreprise technologique disposant de moyens considérables. Mais nous pourrions tout de même constater que :

1. plus le nombre d'itérations est grand pour l'entraînement, plus la fluidité est bonne (« m1 » excepté en tant que modèle réduit) ;
2. l'augmentation du *train set* (différence entre « m2 » et « m3 ») n'a amélioré qu'un peu le score final et la fluidité, mais a grandement dégradé la fidélité, peut-être à cause de la réduction de la taille du jeu de validation ;
3. malgré un nombre d'étapes doublé (100 000 contre 50 000), « m4 » ne parvient pas à surpasser « m3 » de manière significative, et perd même des points en termes d'informativité et de score final ;
4. le nombre d'itérations ne semble pas influer l'informativité.

Comme évoqué à la fin de la section 4, « l'évaluation humaine est toujours considérée comme déterminante dans la communauté scientifique ». Malgré les résultats passables dans deux évaluations automatiques, nous reconnaissons, d'après l'évaluation humaine, que nos quatre modèles entraînés laissent beaucoup à désirer.

Nous réalisons avec ces expérimentations qu'un corpus de bonne qualité et volumineux est crucial pour construire un bon système de NMT français-mongol.

6 Conclusion et perspectives

Dans cet article, nous avons décrit l'historique de la traduction automatique, présenté la langue mongole et défini un état de l'art concernant la traduction automatique pour la paire de langues peu dotée, français-mongol. Faute de recherches connexes, nous avons entamé une nouvelle étape dans le domaine de TA en développant notre propre système (quatre modèles) de traduction automatique neuronale français-mongol.

Ce développement s'est avéré être coûteux en temps et en moyens. D'abord, nous avons trouvé des données sur la paire français-mongol qu'il a fallu préparer. Après l'entraînement, nous avons ensuite effectué deux évaluations avec la métrique automatique populaire, BLEU. Mais nous ne nous en sommes pas contentés et avons en outre proposé « *IFF* », une métrique originale pour l'évaluation humaine en vue de mieux comprendre les forces et les faiblesses de notre système, ce qui a requis un grand investissement du temps et des ressources humaines. Notre travail n'échappe pas à certaines limites, que nous résumons ici en trois points :

1. **Ressources** : Il y a très peu de ressources disponibles sur cette paire de langues, et celle que nous avons trouvée, pour ce que nous en avons vu, a une qualité médiocre. Cela a eu un effet négatif critique sur l'apprentissage et la performance de notre système.

2. **Entraînement** : Nous n'avons pas assez étudié le paramétrage et n'avons utilisé que la configuration de base pour tous les entraînements sans chercher à tester plusieurs configurations ; nous ne savons donc pas si la configuration de base est problématique.
3. **Évaluation humaine** : Nous n'avons pas pu recruter suffisamment d'évaluateurs et nous n'avons pas non plus pu les former. Ils n'avaient pas tous des connaissances nécessaires pour comprendre parfaitement le contenu des textes dans le formulaire. Certains de leurs retours font preuve d'un grand désaccord inter-annotateur, c'est pourquoi nous avons finalement adopté uniquement quatre retours sur six.

Nous souhaitons poursuivre notre travail en essayant différentes configurations, même plusieurs *frameworks* pour trouver le meilleur paramétrage possible. Ensuite, pour l'évaluation humaine, nous pourrions recruter plus tôt les évaluateurs pour qu'ils soient plus nombreux et puissent être formés sur les techniques d'évaluation ainsi que sur le contenu du corpus. Finalement, c'est particulièrement dans les données qu'il faut mettre l'effort pour aller plus loin dans le domaine de NMT, puisqu'elle est orientée données. Nous envisageons donc, pour nos travaux futurs, de développer un corpus parallèle français-mongol (cyrillique et classique) d'une meilleure qualité.

Inévitablement imparfait, notre travail entend revitaliser la langue mongole et revaloriser la paire de langues peu dotée, français-mongol. Il s'agit là d'une tâche ambitieuse et difficile mais dont les retombées sont telles qu'elles ne font que renforcer notre motivation, d'autant plus que les méthodes que nous souhaitons explorer pourraient bénéficier à d'autres langues minoritaires.

Remerciements

Merci à toutes celles et ceux qui ont contribué à l'encadrement du travail et à la rédaction de cet article : M. Damien NOUVEL, M. Qingyu JIANG, Mme Ilaine WANG, Mme Charlotte MARCHINA.

Mercis particuliers aux participants de l'évaluation humaine, sans qui le travail d'évaluation n'aurait pu être aussi approfondi : Mme Urantuya, Mme Sainjargal Anya, Mme Surigage, Mme Sovin, Mme Hairina et M. Ananda.

Références

Ao, M., Xiong, Z., & Hu, H. (2011). 基于蒙科立输入法的蒙古语同形异码词研究. 第十一届全国人机语音通讯学术会议 (NCMMSC2011) 论文集,

Bao, M. (2016). Construction of Mongolian Language Media Texts Corpus. *内蒙古师范大学学报: 哲学社会科学版*, *45*(4), 70-74.

Bojar, O., Federmann, C., Haddow, B., Koehn, P., Post, M., & Specia, L. (2016). Ten years of WMT evaluation campaigns: Lessons learnt. Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem,

Brownlee, J. (2017, le 19 décembre 2019). *A Gentle Introduction to Calculating the BLEU Score for Text in Python*. https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

Cao, Y. (2020). *Research on Chinese-Mongolian Neural Machine Translation Based on Monolingual Corpora* University of Science and Technology of China].

Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mech. Transl. Comput. Linguistics*, *9*(3-4), 55-66.

CWMT. (2017). CWMT 2017 MACHINE TRANSLATION EVALUATION PARTICIPATING SITE AGREEMENT. In.

FAN, W., HOU, H., WANG, H., WU, J., & LI, J. (2018). Mongolian-Chinese Neural Machine Translation with Priori Information. *Journal of Chinese Information Processing*, 06.

Fei, D., Yuan, L., & Quan, C. (2019). Analysis of the construction of minority corpus oriented to information processing. *无线互联科技*, *19*.

Gaunt, J. (2004). *Modern Mongolian: A course-book*. Routledge.

Hou, H., & Liu, Q. (2007). 基于实例的汉蒙机器翻译. *中文信息学报*, *21*(4), 65-72.

Hua, S. (1997). 实现 500 万词级《现代蒙古语文数据库》的主要措施.

Jaimai, P., & Chimeddorj, O. (2008). Corpus building for Mongolian language. Proceedings of the 6th Workshop on Asian Language Resources,

Janhunen, J. (2005). *The Mongolic Languages*. Taylor & Francis e-Library.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., & Rush, A. M. (2020). OpenNMT: Neural Machine Translation Toolkit: 2020 Edition.

Liu, W. (2018). *Research on Mongolian-Chinese Machine Translation Based on LSTM Neural Network* Inner Mongolia University of Technology].

Miller, K. J., & Vanni, M. (2005). *Inter-rater agreement measures and the refinement of metrics in the PLATO MT evaluation paradigm*.

Nakhlé, M. (2021). *Évaluation globale d'un système de traduction automatique de documents structurés dans le domaine financier* Université de Paris Nanterre].

Ochir, & Serguleng, W. (2003). The Design of an English-M ongolian Machine Translation System. *内蒙古大学学报: 自然科学版*, *34*(5), 582-587.

Shen, Z. (2017). *基于注意力神经网络的蒙汉机器翻译系统的研究* 内蒙古大学].

Su, C. (2014). *基于层次短语模型的蒙—汉统计机器翻译研究* 内蒙古大学].

Tiedemann, J. (2012, may). Parallel Data, Tools and Interfaces in OPUS.*Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey.

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language*, *19*(1), 55-83.

Wang, H. (2018). *多粒度蒙古文汉文神经网络机器翻译研究* 内蒙古大学].

Wang, Y., Su, Y., Zhao, Y., Sun, X., & Ren, Q. (2020). Mongolian-Chinese Neural Machine Translation Model Based on Parameter Transfer. *Computer Applications and Software*, *37*(9), 81-87. https://doi.org/10.3969/j.issn.1000-386x.2020.09.014

White, J. S., O'Connell, T. A., & O'Mara, F. E. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of the First Conference of the Association for Machine Translation in the Americas,

Wu, J. (2017). 多方法融合蒙汉机器翻译与译文重排序研究. *内蒙古大学*, 1-z.

Xiao, T., & Zhu, J. (2021). *Machine Translation: Foundations and Models*. https://opensource.niutrans.com/mtbook/homepage.html

Yang, M., Hu, X., Xiong, H., Wang, J., Jiaermuhamaiti, Y., He, Z., Luo, W., & Huang, S. (2019). Ccmt 2019 machine translation evaluation report. China Conference on Machine Translation,

Zhang, G. (2009). *The Experimental Study and Realization of Mongolian-Chinese Alignment corpora* Inner Mongolia Normal University].

Крылов, С. А. (2017). Монгольские аналитические конструкции в количественном аспекте. *Oriental Studies*(5 (33)).