UFR de mathématique et d'informatique
Université de Strasbourg

MASTER D'INFORMATIQUE
PARCOURS SIL
SCIENCE ET INGÉNIERIE DU LOGICIEL

Travail d'Étude et de Recherche

Iman IRAJ DOOST
iman.iraj-doost@etu.unistra.fr

TRADUCTION ET ÉVALUATION ASSISTÉE PAR L'IA

13 avril 2023

TER encadré par
Bérenger BRAMAS
bbramas@unistra.fr

Table des matières

Chapitre 1

Introduction

Traitement automatique des langues (NLP en anglais) utilisant l'intelligence artificielle et l'apprentissage automatique se développe rapidement aujourd'hui, ce qui signifie que de nouvelles méthodes d'évaluation et de révision des traductions utilisant l'IA sont possibles. Dans ce travail, l'évaluation des traductions à l'aide de l'IA sera discutée et testée. Ces IA peuvent être des assistants aux traductions et aux révisions de traduction.

Ce travail utilise un ensemble de données textuelles en français et leurs traductions en anglais réalisées par des traducteurs humains.

Dans ce travail, une forme de prétraitement des données et d'évaluation de la traduction à l'aide d'un score appelé "Review Score" sera discutée. La qualité des traducteurs IA existants tels que le traducteur Google et Deepl sera comparée, puis deux modèles intelligents artificiels seront réalisés.

L'un basé sur les données numériques obtenues à l'étape "Review Score" et l'autre basé sur les données textuelles. À la fin, les modèles formés seront comparés et le potentiel d'une IA en tant qu'assistant traducteur et évaluateur est déterminé. en termes simples, l'idée est d'évaluer une traduction et de la remplacer par une meilleure traduction réalisée à l'aide d'une IA si cela est nécessaire.

Ce travail utilise les données textuelles du projet 7-Shapes [1] avec la permission de l'entreprise. 7-Shapes est une entrerpise à Angoulême qui forme les gens sur 'Lean management' [4] en utilisant un projet ludique et interactif. Les données analysées dans ce travail sont les données textuelles du projet (ex. utilisées dans les dialogues, l'interface utilisateur, etc.) et leur traduction dans différentes langues (anglais, français, chinois, etc.).

1. Le site web de l'entreprise 7-Shapes : https://www.7-shapes.com/

Chapitre 2

Le contexte

2.1 Le travail attendu

Ce travail consiste à analyser et à évaluer les données textuelles du projet "7-Shapes" et à appliquer des algorithmes d'apprentissage automatique pour mettre en œuvre et améliorer le flux de traduction automatique en utilisant une IA existante de traduction automatique à réseau neuronal.

Dans un premier temps, une combinaison de méthodes est utilisée pour évaluer la traduction des textes effectuée par l'IA. Dans ce cas, l'IA Deepl [1] est utilisée pour traduire les textes. Le texte de référence du projet est le français et les traductions effectuées par des traducteurs humains sont l'anglais et le chinois. Dans ce travail, seul la langue anglaise est utilisée.

Une note de révision appelée "Review Score" comprise entre 0 et 1 est calculée à l'aide de cette méthode, où 0 signifie qu'il n'est pas nécessaire de réviser la traduction éffectuée et 1 que la traduction doit être révisée. Cela permettra de filtrer les textes qui n'ont pas besoin d'être analysés, comme les textes où la référence et la traduction sont exactement les mêmes ou les textes qui ne contiennent que des chiffres, etc. Cette partie est une étape de prétraitement pour le processus d'apprentissage automatique.

Dans un deuxieme temps, pour les textes ayant obtenu une note élevée, les textes de référence et les textes traduits sont comparés. Le meilleur entre les deux candidats est choisi sur la base de l'apprentissage automatique.

Le résultat de ce processus est une traduction de haute précision basée sur le contexte (dans ce cas, le Lean Management qui contient des mots spéciaux et des acronymes dans le contexte) sans avoir besoin d'un traducteur humain (sauf pour les données d'entraînement).

2.1.1 État de l'art

La traduction automatique neuronale [5] (Neural machine translation en anglais) a récemment dépassé la traduction automatique statistique [9] (Statistical machine translation en anglais) et cette méthode est maintenant utilisée dans de nombreux nouveaux systèmes de traduction IA tels que Deepl, Bing translation, Google Translation, etc. Cependant, dans divers contextes, ces modèles ne peuvent pas être aussi précis qu'on pourrait l'espérer.

Il existe de nombreux algorithmes pour évaluer les traductions ; des algorithmes qui évaluent les traductions sur la base des mots traduits sans tenir compte de leur signification et d'autres

1. Le site web de Deepl : https://www.deepl.com/

qui utilisent l'apprentissage automatique pour comprendre le contexte et étiqueter le texte afin de les évaluer en leur attribuant un score. Un exemple est **les algorithmes de 'Distance d'édition' ou 'Edit distance' en anglais**.

La plupart de ces algorithmes prennent en compte la distance de chaque caractère dans un mot et calculent la différence entre deux mots sur la base de leur différence de caractères. **La distance de Levenshtein** [7] est une distance, au sens mathématique du terme, donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

Contrairement à la distance de Levenstein, l'algorithme de score **BLEU** [11] (bilingual evaluation understudy) fonctionne de manière plus intelligente et peut comparer deux phrases et calculer un score entre 0 et 1 où 1 signifie un chevauchement parfait entre la phrase traduite et la phrase de référence. Le score BLEU est principalement utilisé pour comparer un texte traduit par une machine avec une traduction faite par un expert humain. Cette méthode ne permet pas d'étiqueter le texte et ne peut pas détecter le sens ou le contexte, mais avec une tokénisation et une **racinisation** [3] appropriés des mots, la précision de cet algorithme est assez intéressante.

Une autre méthode d'analyse syntaxique des textes consiste à utiliser une approche vectorielle. Cette méthode est massivement utilisée dans l'apprentissage automatique textuel et elle utilise des matrices pour comparer les poids de tous les mots d'un texte. [10]

Les méthodes utilisées dans ce travail consistent en une analyse syntaxique et une évaluation à l'aide de ces méthodes.

2.1.2 Contributions personnelles

Dans ce travail, le potentiel d'un assistant IA dans l'évaluation de la traduction est montré. Les modèles IA formés utilisant des données numériques et textuelles montrent qu'avec suffisamment de données d'entraînement, l'ensemble du processus de traduction et l'évaluation peut être effectué automatiquement. Ce processus peut se faire de manière incrémentale pour obtenir un meilleur résultat.

L'évaluation de la nécessité de réviser une traduction se fait à l'aide d'un score appelé "Review Score" qui utilise un ensemble de variables abordées dans la section 2.1.1 telles que le score BLEU et la distance de Levenstein.

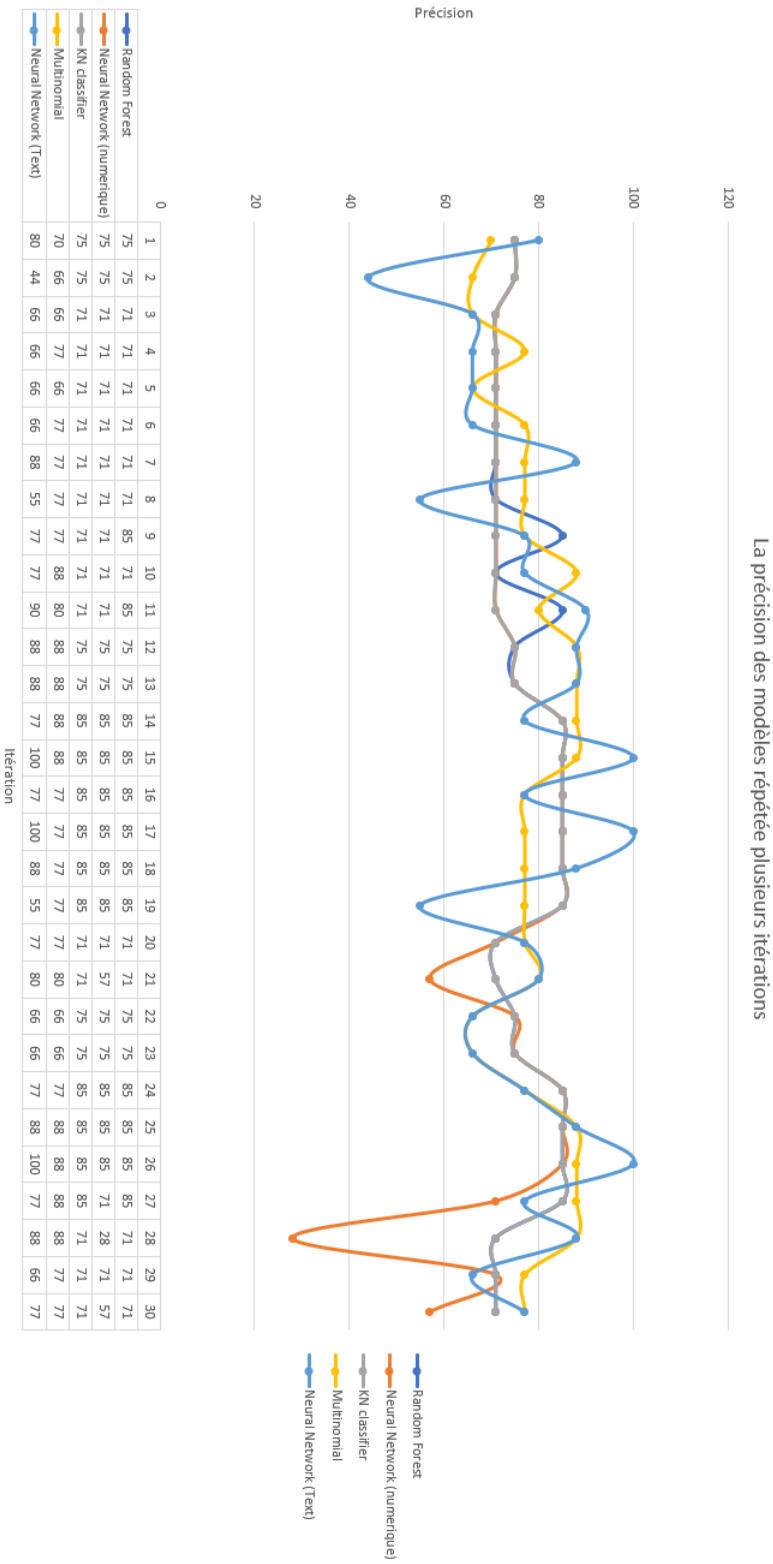
Ensuite, deux modèles d'IA sont entraînés à l'aide des variables obtenues dans la section précédente pour effectuer automatiquement la révision de la traduction sans intervention humaine. Le résultat est un assistant IA qui peut évaluer et réviser les traductions et peut apprendre progressivement pour obtenir des résultats de traduction plus précis.

2.1.3 Les résultats

Les résultats finaux montrent que le modèle Multinomial textuel obtient une précision de 78,43 % et produit de bons résultats lors de l'examen d'un ensemble de données de test. Le modèle Random Forest numerique obtient également de bons résultats, avec une précision de 77,4 %. Les deux modèles montrent une certaine diminution de la précision lorsque plus de cas de 1 sont ajoutés aux données de prédiction (cas où ils doivent choisir la traduction effectuée par Deepl et non la traduction originale). La précision est obtenue en calculant une moyenne des précisions après avoir fait une validation croisée (Cross-Validation en anglais) K-Fold avec une répétition de 30 fois. Les détailles de ce calcul est montré dans la figure 2.1.

Il y avait de rares fois où la précision était d'environ 28 %, qui n'a pas tellement impacté la

moyenne (l'itération numéro 28 pour le réseau de neuronal numérique dans la figure 2.1). Cela montre qu'un ensemble de données plus grand pour l'entraînement sera nécessaire pour obtenir des résultats plus précis.



Itération	Random Forest	Neural Network (numerique)	KN classifier	Multinomial	Neural Network (Text)
1	75	75	75	70	80
2	75	75	75	66	44
3	71	71	71	66	66
4	71	71	71	77	66
5	71	71	71	66	66
6	71	71	71	77	66
7	71	71	71	77	88
8	71	71	71	77	55
9	85	71	71	77	77
10	71	71	71	88	77
11	85	71	71	80	90
12	75	75	75	88	88
13	75	75	75	88	88
14	85	85	85	88	77
15	85	85	85	88	100
16	85	85	85	77	77
17	85	85	85	77	100
18	85	85	85	77	88
19	85	85	85	77	55
20	71	71	71	77	77
21	71	71	57	80	80
22	75	75	75	66	66
23	75	75	75	66	66
24	85	85	85	77	77
25	85	85	85	88	88
26	85	85	85	88	100
27	85	85	71	88	77
28	71	71	71	88	88
29	71	71	71	77	66
30	71	71	57	77	77

FIGURE 2.1 – La précision des modèles répétéé 30 fois

Chapitre 3

L'évaluation des traductions

3.1 Présentation

Cette étape consiste à évaluer automatiquement la traduction effectuée par Deepl par rapport à la traduction de référence effectuée par un traducteur humain.

Pour mieux démontrer cette méthode, un échantillon des données textuelles existantes est présenté dans la figure 3.1.

Ligne	Réference français	Traduction humain	Traduction Deepl
1	Lorsqu'on fait un "5 Pourquoi", que faut-il éviter ou ne PAS faire ?	When you do a "5 why", what should be avoided or not doing ?	When doing a "5 Why", what should you avoid or NOT do ?
2	Semaine 3 : Lean Six Sigma	Week 3 : Lean Six Sigma	Week 3 : Lean Six Sigma
3	Considérant deux équipes, 8 heures par jour, 30 minutes de pause par équipe, 220 jours travaillés par an. Quel est le takt time si la demande moyenne mensuelle est de 12000 pièces ?	Considering two teams, 8 hours a day, 30 minutes of break per team, 220 days worked per year. What is the takt time if the average monthly demand is 12,000 pieces ?	Considering two shifts, 8 hours per day, 30 minutes break per shift, 220 days worked per year. What is the Takt Time if the average monthly demand is 12000 pieces ?
4	Sous-traiter	Outsourcing	Subcontracting

TABLE 3.1 – Quelques lignes de texte du projet

L'idée principale de l'évaluation est la suivante :

- de voir si la traduction de Deepl est suffisamment précise

- de mesurer la possibilité que Deepl fasse un meilleur travail dans certains cas qu'un traducteur humain.

à noter que les traductions ne sont pas faites par des traducteurs experts et qu'il est donc possible que les traductions ne soient pas exactes à 100%. Cela peut créer quelques difficultés pour les données d'entraînement dans l'apprentissage automatique. C'est pourquoi le prétraitement et l'analyse du texte avant l'apprentissage automatique sont importants.

3.2 L'analyse des traductions existantes

Les traductions anglaises existantes (on va les appeler "les traductions existantes" ou "les traductions par humains" ou "les traductions originales") ne sont pas forcément correctes.

Par exemple, à la ligne 3 du tableau 3.1, la traduction de Deepl est plus précise que la traduction originale. Ceci est confirmé par les experts de Lean. Des exemples comme celui-ci montrent l'idée qu'un "glossaire" ou un "dictionnaire" de mots du Lean est nécessaire pour que le traducteur IA soit capable de traduire avec précision le texte donné.

Un tel glossaire est préparé et utilisé pour Deepl dans les traductions. Un exemple est montré dans la figure 3.2

Ligne	Mot français	Mot anglais
1	AMDEC	FMEA
2	Bac à sable	Sandbox
3	Flux tendu	Just-in-time
4	NVAE	NNVA

TABLE 3.2 – Quelques exemples du glossaire

Deepl AI est parfois capable de comprendre le contexte du texte en l'étiquetant. Ainsi, lorsqu'on lui donne un long texte, les résultats sont parfois plus satisfaisants que les traductions de mots isolés. [1]

Outre les traductions elles-mêmes, un autre problème des traductions existantes est la nécessité d'un prétraitement tel que la suppression des espaces vides, la correction des fautes d'orthographe, la conservation du format des données, etc. Ce traitement est également effectué à l'aide d'outils ou par des experts Lean.

Ces traductions peuvent contenir un seul mot ou de longues phrases. Elles comprennent parfois des chiffres, des formules ou des liens. Dans certains cas (comme les chiffres purs), le texte ne doit pas être traduit du tout et dans d'autres cas, comme lorsqu'il contient des liens, une évaluation humaine sera nécessaire parce que le lien doit être remplacé complètement pour d'autres langues.

3.3 Déterminer la nécessité d'une révision avec le "Review Score"

Pour gérer tout cela et déterminer les traductions qui nécessitent une révision, un score appelé "Review Score" (note d'évaluation ou score de révision) est calculé entre 0 et 1, où 1 signifie qu'une révision est absolument nécessaire et 0 signifie qu'aucune révision n'est nécessaire. La formule pour calculer le review score est la suivante :

$$ReviewScore = \Sigma_{i=1}^{n}(weight_i \times variable_i)$$

où la variable de l'indice i est un critère d'évaluation comme le score BLEU, le fait que le texte contienne des formules ou des liens ou non, etc. Le poids de l'indice i est le degré d'importance

1. Ceci est indiqué dans la documentation de Deepl : "Le moteur de traduction DeepL prend en compte le contexte plus large d'un texte ou d'un document source lors de la traduction - en particulier, le contexte des phrases qui sont proches les unes des autres. En général, l'inclusion d'un contexte plus large dans un texte ou un document source peut se traduire par une traduction de meilleure qualité sur DeepL." https://www.deepl.com/docs-api/general/working-with-context/

ou d'impact de cette variable sur la note d'évaluation. N est le nombre de variables/poids existant pour l'évaluation.

À la fin de ce calcul, une normalisation du score de "Review Score" est effectuée pour normaliser les scores entre 0 et 1.

$$\forall ReviewScore, NormalizedReviewScore_{reviewscore} = (ReviewScore - min) \div (max - min)$$

Les variables et leurs poids correspondants sont indiqués dans le tableau 3.3.

Ligne	Variable	Poids
1	Contient une formule	2
2	Score BLEU	21
3	Score Levenstein	5
4	La source est vide	42
5	Contient seulement les chiffres	-10
6	Nombre de mots *	0.01
8	Contient des crochets	0.2
7	Contient formattage	0
9	Score TER	0
10	Score TRI	0
11	Score LCS	0

TABLE 3.3 – Les variables et leurs poids pour le "Review Score"

* La variable Nombre de mots à la ligne 6, a un impact inverse sur le résultat, donc dans la formule, elle est inversée à 1 / NombreDeMots.

Un poids négatif signifie que l'impact de la variable est inversé en fonction de la variable (diminue le review score). Les valeurs des poids sont choisies par essai et erreur [2] (Trial and Error en anglais). L'impact le plus important est l'erreur de la source, par exemple, il arrive que la traduction de la source soit vide. Dans ce cas, le review score doit être proche de 1, et la pondération est donc élevée (42). en cas de score BLEU de 1, le poids de la variable est multiplié par -2 parce qu'un score BLEU de 1 signifie que les deux traductions sont identiques et que le Review Score doit donc être proche de 0, ce qui signifie qu'aucune révision n'est nécessaire.

Un poids de 0 signifie que la variable n'a aucun impact sur le Review Score, mais il est calculé pour l'apprentissage automatique dans la section 4.3.

Les variables booléennes telles que "Contient une formule" ont une valeur de 0 ou 1 tandis que les variables telles que le score BLEU sont des variables flottantes et peuvent être comprises entre 0 et 1.

La variable "Nombre de mots" a un impact inverse sur le résultat. Plus il y a de mots dans une même traduction, plus elle est précise. Cette logique repose sur le fait que Deepl peut comprendre le contexte au fur et à mesure que les phrases s'allongent. Cependant, une plus grande variance a été observée lors de la traduction d'un seul mot à l'aide de Deepl, c'est pourquoi le 1/nombreDeMots a été choisi comme formule pour être plus sensible aux mots uniques.

Le tableau 3.4 présente certaines données ainsi que le Review Score calculé.

Dans le tableau 3.4, seul le score BLEU est indiqué car il a l'impact le plus important parmi les autres variables. Le score d'examen normalisé est également calculé pour chaque ligne. Comme affiché dans le tableau 3.4, les lignes ayant un score BLEU plus élevé ont un score de Review

Ligne	Réference français	Traduction humain	Traduction Deepl	Score BLEU	Review Score
1	Lorsqu'on fait un "5 Pourquoi", que faut-il éviter ou ne PAS faire ?	When you do a "5 why", what should be avoided or not doing ?	When doing a "5 Why", what should you avoid or NOT do ?	0.6153	0.2452
2	Semaine 3 : Lean Six Sigma	Week 3 : Lean Six Sigma	Week 3 : Lean Six Sigma	1	0.1680
3	Considérant deux équipes, 8 heures par jour, 30 minutes de pause par équipe, 220 jours travaillés par an. Quel est le takt time si la demande moyenne mensuelle est de 12000 pièces ?	Considering two teams, 8 hours a day, 30 minutes of break per team, 220 days worked per year. What is the takt time if the average monthly demand is 12,000 pieces ?	Considering two shifts, 8 hours per day, 30 minutes break per shift, 220 days worked per year. What is the Takt Time if the average monthly demand is 12000 pieces ?	0.8387	0.4668
4	Sous-traiter	Outsourcing	Subcontracting	0	0.9031

TABLE 3.4 – Quelques lignes de texte du projet avec le Review Score

moins élevé. Bien entendu, d'autres variables ont également un impact sur les modifications du Review Score, mais elles ne sont pas affichées ici. Un score BLEU de 0, comme celui de la ligne 4, produit un Review Score élevé qui indique que cette ligne devrait être révisée et qu'il pourrait y avoir une erreur.

Avec ces étapes, les lignes avec des Review Score élevés sont trouvées et peuvent être révisées. Après avoir testé différentes valeurs, un score de révision supérieur à 0,8 semble être un bon début pour voir les écarts importants entre les traductions et les textes existants. Cependant, ces lignes sont très nombreuses et une révision humaine de tous ces textes prendra du temps. C'est pourquoi l'apprentissage automatique sera utilisé dans la suite du processus.

Chapitre 4

Machine Learning

4.1 Présentation

Cette étape se compose de deux parties : la première consiste à comparer les IA de traduction existantes pour voir s'il est possible d'obtenir une meilleure traduction ; la seconde consiste à utiliser l'apprentissage automatique pour évaluer les traductions et comparer le texte existant et la traduction afin de décider laquelle conserver.

4.2 Une comparaison entre les IA de traduction

Avant d'utiliser Deepl comme traducteur par défaut, quelques évaluations et comparaisons ont été effectuées. Ces évaluations sont détaillées dans les sections suivantes. Une centaine de lignes de texte pour chaque IA ont été testées et examinées par des experts humains afin de déterminer la qualité de la traduction.

4.2.1 Google AutoML

Google propose plusieurs services de traduction. Le service AutoML ou automatique machine learning peut évaluer les traductions et calculer un score BLEU sur la base d'un ensemble de textes donnés.

Une série de paires de phrases a été donnée à Google AutoML pour évaluer le score BLEU. Cet ensemble de données se compose du texte source en français et des textes traduits en anglais. Google AutoML a ensuite traduit ces paires et les a comparées à la traduction de référence pour calculer le score BLEU.

Cette méthode a permis d'obtenir un score de 53,2 sur 100. Le modèle a ensuite été entraîné à l'aide du dictionnaire Lean (un gloassaire) et a été réévalué, ce qui a permis d'augmenter le score BLEU d'environ 3 %. Le résultat amélioré est de 56,62 sur 100. Cela confirme le fait que si l'IA connaît le contexte, nous obtiendrons de meilleurs résultats en traduction. Le résultat est montré dans 4.1. Google NMT ou Google neural machine translation est le système utilisé par Google Traduction.

A noter que selon Google, un score BLEU compris entre 50 et 60 est considéré comme une "traduction de très haute qualité" [1]

1. Voir https://cloud.google.com/translate/automl/docs/evaluate?hl=fr

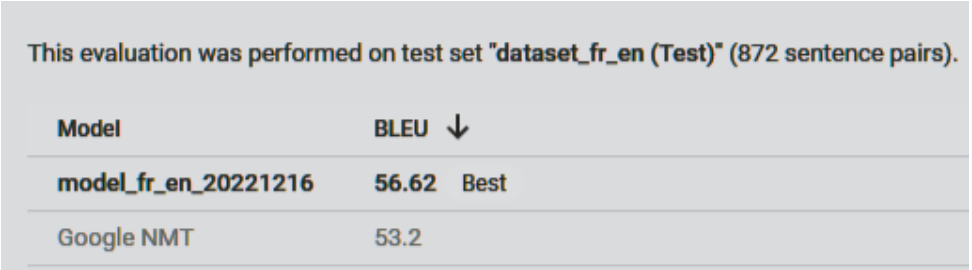


FIGURE 4.1 – Résultat de Google AutoML avec le glossaire

4.2.2 OpenAI

À ce jour, OpenAI ne fournit pas de modèle spécifique pour la traduction, mais ses différents modèles d'IA GPT-3, tels que Davinci [2], sont capables de traduire du texte. Il s'agit des mêmes modèles que ceux utilisés dans le Chat GPT [3], largement répandu, et qui sont capables de comprendre le contexte et de répondre aux questions de l'utilisateur.

Cependant, l'essai de ce modèle pour la traduction automatique n'a pas donné les meilleurs résultats. Pour utiliser ce modèle, un "prompt" doit être envoyé au serveur où le modèle analyse en utilisant les modèles NLP et répond à la demande avec la réponse appropriée. De nombreux types de "prompt" de formats différents ont été utilisés pour obtenir les meilleurs résultats, mais les capacités de personnalisation de l'API OpenAI sont insuffisantes, ce qui rend difficile l'obtention de la bonne réponse. Les réponses étaient parfois "trop créatives" et l'IA a ajouté des explications à la traduction ou a parfois simplement renvoyé le "prompt" de l'utilisateur comme réponse. Dans les cas où la traduction a été renvoyée, la qualité des traductions était meilleure que celle de Google translate.

Son vaste potentiel pourrait être utilisé pour des tâches plus compliquées telles que la rédaction du glossaire. Le glossaire "Lean" n'était pas complet et certains mots n'étaient pas traduits, ce qui pourrait être automatisé à l'aide d'Open AI. Un petit ensemble de données d'environ 10 mots du glossaire en français a été donné à Open AI et il lui a été demandé de donner l'équivalent dans le contexte du "Lean management" en anglais. De nombreux résultats étaient corrects, mais le même problème persistait dans certains cas.

4.2.3 Deepl

Les résultats obtenus avec l'IA Deepl sont les meilleurs parmi les autres. Bien qu'elles ne disposent pas du système d'analyse de Google AutoML, les traductions utilisant le glossaire personnalisé étaient beaucoup plus claires et plus proches des traductions originales (et parfois meilleures).

Les traductions effectuées par OpenAI et Deepl étaient assez proches l'une de l'autre, mais compte tenu de la nature quelque peu aléatoire d'OpenAI et de la précision de Deepl, le gagnant était clair. Dans tous les cas, même avec le glossaire, la précision de la traduction Google était nettement inférieure à celle des autres modèles.

2. Voir https://platform.openai.com/docs/models/gpt-3
3. Voir https://openai.com/blog/chatgpt

4.3 Utilisation de machine learning pour déterminer la meilleur traduction

L'idée générale de cette section est d'effectuer automatiquement la révision de la traduction obtenue dans la section 3.3 sans l'aide d'un traducteur humain. Deux types différents d'IA sont entraînés ; l'un avec le texte et l'autre avec des chiffres tels que le score BLEU ou le Review Score. Ces IA sont des IA de classification binaire qui donnent un résultat de 0 ou 1. Dans ce cas, 0 signifie que la traduction originale en anglais doit être conservée et que le texte traduit par Deepl n'est pas aussi précis que l'original, et 1 signifie que Deepl a fait un meilleur travail et que la nouvelle traduction doit être remplacée.

Pour entraîner l'IA, environ 400 lignes de texte ont été examinées par un expert humain qui devait choisir entre la traduction anglaise existante et la traduction réalisée par Deepl. Quelques exemples de lignes sont présentés dans le tableau 4.1.

Ligne	Réference français	Traduction humain	Traduction Deepl	Remplacer avec la trad Deepl ?
1	Quelle est l'affirmation fausse ?	What is the false affirmation ?	Which statement is false ?	1
2	Accidents	Accidents	Collisions	0
3	Une formation croisée	Cross-training	Cross-training	0
4	Une démarche Lean s'appuie notamment sur quel principe ?	A Lean approach relies on which principle ?	What is the main principle behind a Lean approach ?	0

TABLE 4.1 – Quelques lignes de texte révisés par un humain expert

Dans certains cas, comme à la ligne 1, la traduction Deepl est considérée comme meilleure que la traduction originale. Dans les cas où la traduction originale et la traduction Deepl sont identiques, la traduction originale est conservée (ce qui donne un 0 dans ce cas). Il convient de noter que ces lignes font partie des textes qui ont un Review Score élevé et peuvent éventuellement présenter des problèmes même dans la traduction anglaise originale (Comme ne pas avoir de traduction du tout, ce qui, dans ce cas, donne le résultat 1).

Le code des projets sont sur le répository Gitlab [8].

4.3.1 apprentissage automatique avec des données de type numérique

pour chaque ligne de texte, une variable de Review Score est calculée. Indépendamment de son poids, cette variable numérique peut être utilisée comme critère pour l'apprentissage automatique. Ce type d'IA ne prend pas de texte en entrée comme données d'apprentissage, mais utilise les variables calculées dans le tableau 3.3 pour s'entraîner et prédire si une traduction doit être remplacée ou non.

Pour maintenir un bon équilibre entre les données de formation et de validation, 80 % des données ont été utilisées pour l'entraînement et 20 % pour la validation et le test. Une IA a été écrite en langage Python avec l'aide de la bibliothèque SKLearn. [4].

4. Voir https://scikit-learn.org/

Le modèle met en œuvre 3 types d'apprentissage automatique en utilisant les algorithmes de Random Forest, de réseau neuronal et de KN Classifier. En testant les trois algorithmes avec des données de prédiction choisies au hasard (jamais vues auparavant par l'IA), le modèle de Random Forest s'est avéré meilleur que les autres. Les résultats sont présentés dans le tableau 4.3.

Modèle	**Précision**	**Prédiction correctes des données jamais vues**
Random Forest	77,4 %	9 sur 12
Neural Network	73,63 %	9 sur 12
KN classifier	76,46 %	7 sur 12

TABLE 4.2 – Résultats des modèles d'apprentissage pour des données numériques

Cependant, le modèle semble donner des résultats variés lorsque les données d'entraînement sont modifiées (sélection aléatoire de 80 % des données des 400 lignes). Cela montre que les données d'entraînement ne sont pas suffisantes pour former un modèle très précis, mais dans la pratique, après avoir prédit les données réelles et vérifié les résultats, le modèle a bien réussi à prédire quelle traduction était bonne et quelle traduction ne l'était pas.

Le modèle a prédit 10584 lignes des textes du projet dans lesquelles 8968 étaient considérées comme 0 (ce qui signifie que la traduction originale en anglais devrait être conservée) et 1616 étaient considérées comme 1 (ce qui signifie que la nouvelle traduction par Deepl devrait les remplacer). Environ 100 lignées ont été choisies dans chaque catégorie (0 et 1) et ont été vérifiées par des experts humains. Le modèle avait une précision d'environ 72 % sur la base des examens humains. Ce fait doit cependant être gardé à l'esprit que dans certains cas, les relecteurs humains ne savaient pas quelle traduction choisir.

4.3.2 apprentissage automatique avec des textes

Dans ce type de modèle, la langue source française a été utilisée comme modèle d'entrée. Ce modèle utilise les idées de vectorisation de texte discutées ici [10] et ici [1] pour créer une matrice de mots avec leur présence (0 ou 1) dans chaque texte. Ce modèle utilise d'abord la racinisation [3] pour obtenir la racine des mots et supprime des mots tels que "le" et "les".

Les modèles d'IA utilisés dans ce type d'apprentissage automatique sont les algorithmes de réseau Multinomial [6] et Neural. Le même principe que la section 4.3.1 est appliqué à ces modèles. 80 % des données sont utilisées pour la formation et les 20 % restants pour la validation et les tests. Une précision de 78,43 % est obtenue dans le cas Multinomial et 77 % pour le réseau neuronal. Dans la pratique également, le modèle Multinomial a de meilleurs résultats que le réseau de neurons.

Le modèle Multinomial a prédit 10584 lignes des textes du projet dans lesquelles 10231 étaient considérées comme 0 (ce qui signifie que la traduction originale en anglais devrait être conservée) et 353 étaient considérées comme 1 (ce qui signifie que la nouvelle traduction par Deepl devrait les remplacer)

Modèle	**Précision**	**Prédiction correctes des données jamais vues**
Multinomial	78,43 %	28 sur 31
Neural Network	77 %	22 sur 31

TABLE 4.3 – Résultats des modèles d'apprentissage pour des textes français

La même méthode a été appliquée aux textes originaux en anglais, les résultats n'étaient pas aussi bons que ceux obtenus en utilisant la source française mais une précision d'environ 69 % a été obtenue ce qui montre le potentiel de cette méthode également s'il y a plus de données d'entraînements.

4.3.3 Récapitulatif et résultat

En général, les modèles textuels et numériques ont montré un grand potentiel et, lorsqu'ils sont appliqués aux données réelles, ils ont tous deux obtenu de bons résultats. Le modèle textuel a cependant eu des résultats plus précis que l'autre. Une comparaison entre les modèles est montré dans la figure 4.2.

La différence entre deux modèles en nombre de 1 obtenus, montre que dans certains cas le premier modèle a mal révisé le texte (Cela signifie qu'une analyse plus précise du texte lui-même est requise plutôt que des seules variables numériques). Cependant, dans certains cas, comme mentionné précédemment, même un relecteur humain dirait que le texte traduit et le texte original pourraient être utilisés. Même si l'on considère cette variance, le modèle textuel est préféré par les examinateurs humains.

Dans certains cas où la prédiction était de 0 (AI a choisi la traduction originale en anglais), la traduction Deepl a été préférée par le relecteur humain, mais cela ne signifiait pas que la traduction originale était erronée (Cela peut être considéré comme un faux négatif). Alors que la plupart des erreurs les plus importantes étaient dans les cas de 1 où la traduction originale a été remplacée par une traduction pas si bonne par Deepl (cela peut être considéré comme un faux positif).

Pour avoir une détermination précise des résultats obtenus, toutes les prédictions faites par les deux types d'IA doivent être vérifiées. Le fait qu'il y ait plus de 100 000 lignes de texte rend cette tâche difficile, donc environ 100 lignes ont été choisies au hasard et ont été vérifiées pour voir si les prédictions étaient exactes.

Environ 75 % des données ont été prédites correctement par les deux modèles, mais plus les résultats de 1 ont été ajoutés à la vérification, moins la précision de la prédiction devient. Dans tous les cas, le modèle textuel multinomial a montré un meilleur potentiel que le modèle numérique car plus de résultats ont été ajoutés au processus de vérification, le modèle multinomial a diminué moins rapidement que le modèle de réseau neuronal (Mutlinomial est resté environ 5 % plus élevé que le réseau neuronal après avoir ajouté quelques textes aux données de vérification).

Cela dit, en ayant plus des données d'entraînement, les résultats obtenus peuvent être différents et les modèles peuvent avoir un meilleur résultat.

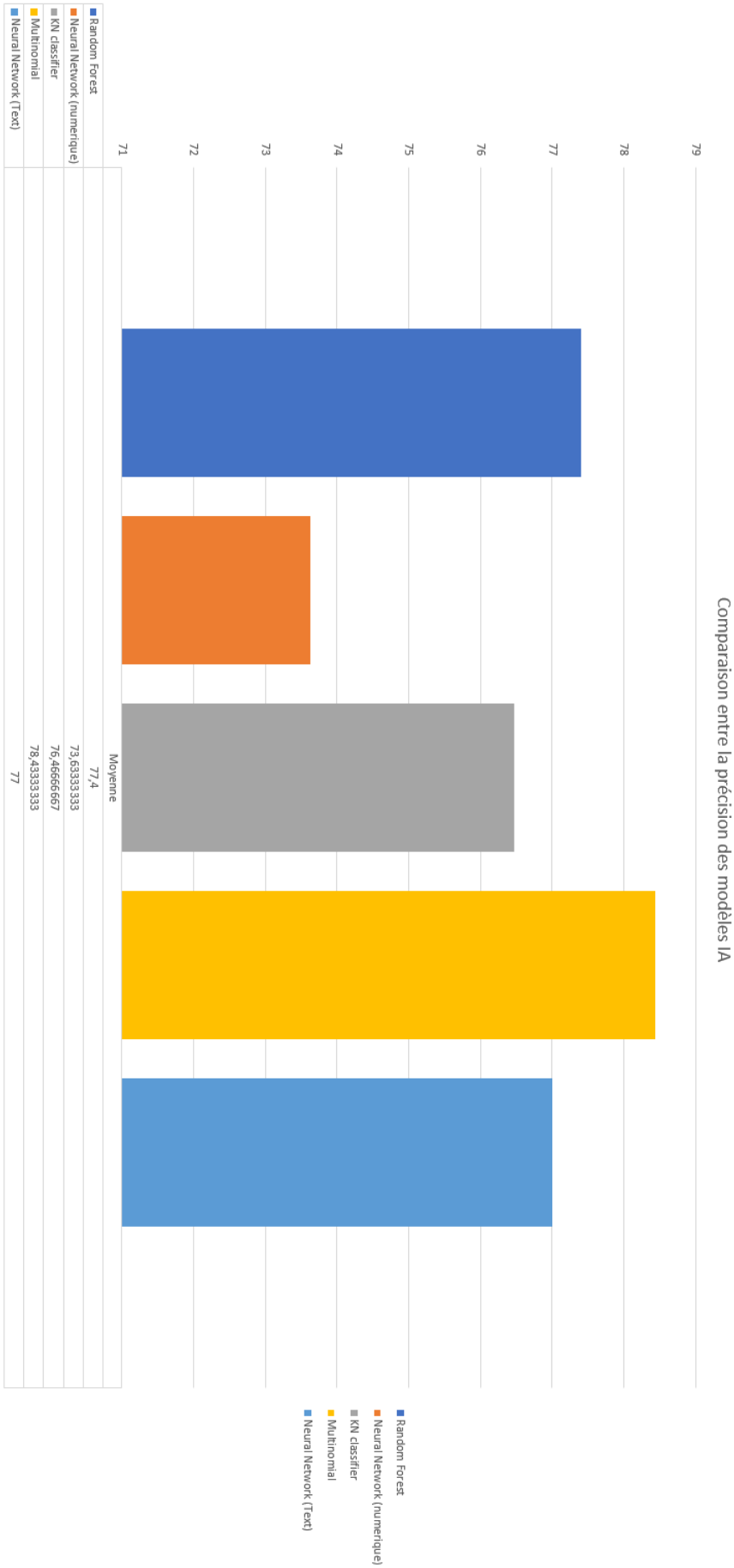


FIGURE 4.2 – Comparaison entre la précision des modèles

Chapitre 5

Conclusion

Traitement automatique des langues (NLP en anglais) utilisant l'intelligence artificielle et l'apprentissage automatique se développe rapidement aujourd'hui, ce qui signifie que de nouvelles méthodes d'évaluation et de révision des traductions utilisant l'IA sont possibles. Ce travail a discuté de l'évaluation des traductions existantes sur la base d'une traduction source et du calcul d'un score appelé "Review Score" afin de l'utiliser dans l'apprentissage automatique pour choisir la meilleure traduction.

Cet article a discuté de deux modèles d'IA basés sur des données numériques et textuelles obtenues via le processus de calcul "Review Score" et les a comparés pour voir lequel fonctionne le mieux sur les données textuelles. Ces modèles d'IA ont ensuite été utilisés pour prédire si une traduction obtenue par une IA (Deepl dans ce cas) est meilleure qu'une traduction existante ou non.

Les résultats étaient intéressants et ont montré le potentiel d'utiliser un modèle textuel multinomial basé sur des textes sources français ou d'utiliser un modèle de réseau neuronal numérique pour prédire quelle traduction est meilleure que l'autre. Cependant, pour obtenir des résultats et des comparaisons plus précis, un ensemble de données d'entraînement plus grands est préférable. Les précisions de chaque modèle est montré dans la figure 4.2.

5.0.1 Prochaines étapes de la recherche

Pour avancer dans la recherche, il existe de nombreuses possibilités d'essais et d'erreurs et de combinaisons qui peuvent être faites. A titre d'exemple pour l'évaluation, en plus du score BLEU, d'autres types de variables peuvent être utilisées. Au lieu de comparer les chaînes de texte à l'aide du score BLEU, un système de notation plus intelligent tel que COMET [1] peut être utilisé. Il utilise un système d'étiquetage et, contrairement au score BLEU, est capable de comprendre le sens du texte.

Une autre possibilité consiste à utiliser le service de traduction d'Amazon [2] au lieu de Deepl qui a montré un grand potentiel ces dernières années.

1. Voir https://unbabel.github.io/COMET/html/index.html
2. Voir https://aws.amazon.com/translate/

Annexe A

Bibliographie

[1] Count vectorizer. https://kavita-ganesan.com/how-to-use-countvectorizer/#.ZDMCInZBxD9. Accessed : 09-04-2023.

[2] Méthode essai-erreur. https://fr.wikipedia.org/wiki/MÃľthode_essai-erreur. Accessed : 09-04-2023.

[3] Racinisation. https://fr.wikipedia.org/wiki/Racinisation. Accessed : 04-03-2023.

[4] What is lean management. https://www.manutan.com/blog/fr/lexique/le-lean-management-definition-et-outils. Accessed : 06-03-2023.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[6] Denzil G Fiebig, M Keane, Jordan Louviere, and Nada Wasi. The generalized multinomial logit model. *Marketing Science*, 2007.

[7] Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods : An improved approach, 2011.

[8] IRAJ DOOST Iman. Binary classification machine learning. https://git.unistra.fr/irajdoost/ter_m1s2_2023, 2023.

[9] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

[10] Elsa Negre. Comparaison de textes : quelques approches... working paper or preprint, April 2013.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.