UNIVERSITÉ DU QUÉBEC À MONTRÉAL

RECONNAISSANCE D'ENTITÉS NOMMÉES INUKTITUT POUR LA TRADUCTION AUTOMATIQUE NEURONALE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

SOUMIA KASDI

JUILLET 2023

REMERCIEMENTS

Tout d'abord, j'adresse mes remerciements à ma directrice de recherche, la professeure Fatiha Sadat et mon superviseur Tan Le, pour leur excellent encadrement, leurs précieux conseils et leur soutien tout au long de ma maîtrise.

Je tiens également à remercier toute l'équipe du laboratoire, particulièrement mes collègues et amis Antoine Cadotte, Habiba Chakour et Soheila Ansari pour leur aide et leurs encouragements.

J'adresse un remerciement spécial à mes chers parents, mes soeurs : Nachida, Lamia et Rym, mon frère Amine, mes beaux-frères, ma belle-soeur, mes nièces et mes neveux.

Enfin je tiens à remercier mes amies Thiziri Boukhalfa et spécialement Hanaa Benyerbah d'avoir toujours été présentes à mes côtés.

DÉDICACE

À mama et à la mémoire de mon cher papa "Mahmoud Kasdi"

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

ACE	Génération automatique de définitions (Automated Concatenation of Embeddings)
CBOW	Continuous Bag-of-Words
CMV	Vote contextuel à la majorité (Contextual Majority Voting)
CRF	Champs aléatoires conditionnels (Conditional Random Fields)
EI	Extraction d'information
EMLO	Embeddings from Language Model
EN	Entité nommée
GPT	Generative Pre-training
HMM	Modèle de Markov caché (hidden Markov model)
LSA	Analyse sémantique latente (Latent Semantic Analysis)
LSTM	Réseaux de neurones récurrents à mémoire court terme et long terme(Long Short-Term Memory)
MLM	Modèle de langue masqué (Masked language model)
REN	Reconnaissance des entités nommées
RNR	Réseaux de neurones récurrents (Recurrent Neural Networks)
TALN	Traitement automatique du langage naturel
TAN	Traduction automatique neuronale
TEA	Taux d'erreurs d'alignement

RÉSUMÉ

La reconnaissance des entités nommées (REN) est une étape cruciale pour garantir une bonne qualité des performances de plusieurs applications et outils du traitement automatique du langage naturel (TALN), notamment la traduction automatique et la recherche d'information.

Dans ce projet, nous visons à construire un système REN dans la langue inuktitut, l'une des principales langues inuites du Canada. Étant une langue peu dotée, pauvre en ressources linguistiques et en données labellisées, ceci représente un défi majeur pour le développement de notre système. De ce fait, en considérant que les données labellisées sont disponibles principalement en anglais, nous proposons un modèle capable de détecter les entités nommées dans la langue inuktitut en faisant le transfert des caractéristiques linguistiques de l'anglais vers l'inuktitut. Nous proposons deux modèles, le premier basé sur des règles et le deuxième basé sur les plongements des mots bilingues. Par la suite, nous évaluons les deux approches et nous fairons une étude comparative entre les deux méthodes d'extraction des EN.

Les résultats d'évaluation montrent une amélioration des performances de la traduction automatique neuronale par rapport à notre modèle de base.

Mots-clés : Traitement automatique des langues naturelles, reconnaissance des entités nommées, plongements des mots, alignement des mots, inuktitut.

CHAPITRE I

INTRODUCTION

Ces dernières années, l'intelligence artificielle, la science de reconstitution des actions et des raisonnements intelligents à l'aide de moyens artificiels, a connu un énorme succès dans la recherche scientifique, notamment, dans le domaine du traitement automatique du langage naturel (TALN).

Le TALN est un domaine multidisciplinaire qui traite la linguistique et vise à développer des techniques qui permettent aux systèmes informatiques de comprendre et de manipuler les langages naturels (Joseph *et al.*, 2016).

Dans ce projet de recherche, on s'intéresse à la reconnaissance d'entités nommées. La REN est une tâche consiste à détecter et classer les noms des catégories spécifiées selon des types sémantiques prédéfinis, tels que les noms de personnes, le lieu, l'organisation ainsi que les expressions numériques, notamment avec la devise, les dates et les pourcentages (Yadav et Bethard, 2019). De plus, La REN compte parmi les tâches fondamentales du TALN : elle est utilisée dans l'extraction d'informations, la traduction automatique et les systèmes de dialogues, par exemple, la réponse aux questions (Li *et al.*, 2022).Cependant, la réussite de ces méthodes du TALN dépend fortement de la quantité de données annotées, disponibles, qui sont rares et difficiles à obtenir pour certaines langues. Les premières

langues cibles des méthodes de REN sont des langues riches en ressources linguistiques (par exemple, l'anglais, l'espagnol, le chinois, etc.) ce qui permet d'atteindre des résultats intéressants. Néanmoins, à cause de l'indisponibilité des données annotées, il est plus difficile d'appliquer ces méthodes sur des langues pauvres en ressources linguistiques, appelées langues peu dotées, telles que l'inuktitut, l'une des principales langues autochtones du Canada.

Selon l'UNESCO, 75% des langues autochtones sont menacées de disparition, et la perte de langue se produit actuellement à un rythme accéléré en raison de la mondialisation. De ce fait, la revitalisation des langues en danger est devenue une cible primordiale pour le maintien de la diversité culturelle sur notre planète. Dans cette lignée, (ACL, 2022) invite les chercheurs à soumettre des ressources et des articles qui se concentrent sur le *" rôle des technologies de la parole et du langage dans le maintien de l'utilisation de la langue "* (Bird, 2020). D'autre part, Assemblée générale des Nations unies proclame la période entre 2022 et 2032 la décennie internationale des langues autochtones (IDIL2022-2032, 2022). Elle déclare aussi que :

" *En 2019, l'Assemblée générale des Nations Unies a adopté une résolution proclamant la période 2022-2032 Décennie internationale des langues autochtones, sur la base d'une recommandation de l'Instance permanente sur les questions autochtones. La proclamation de l'IDIL2022-2032 est un résultat clé de l'Année internationale des langues autochtones 2019 (IYIL2019).*"

Le traitement automatique du langage naturel admet un rôle majeur et contribue énormément à la revitalisation des langues autochtones en danger à travers les applications qui favorisent la compréhension de ces langues (Zhang *et al.*, 2020).

1.1 Problématique

Dans la majorité des applications du traitement automatique du langage naturel, les expressions telles que les noms des personnes, les noms des organisations et les noms de lien, sont importantes. De ce fait, la reconnaissance des entités nommées est une tâche fondamentale du TALN, dédiée à identifier ces expressions et à les classer.

Étant donné que la compréhension du langage compte parmi les objectifs du TALN, il s'avère nécessaire d'appliquer les méthodes du TALN sur toutes les langues, riches en ressources linguistiques, telles que l'anglais et le français. Il est également important de l'appliquer sur les langues moins riches en ressources telles que l'inuktitut. Une langue à faibles ressources, aussi appelée langue peu dotée, est une langue mal informatisée, qui n'admet peu ou pas de corpus annotés, de dictionnaires de noms et de bons analyseurs morphologiques, des ressources pour l'application des méthodes REN, et ainsi présente un défi majeur dans le domaine du TALN.

L'apprentissage automatique, étant une approche utilisée pour la reconnaissance d'entités nommées la rend applicable à différentes langues. De plus, l'apprentissage automatique utilise des algorithmes supervisés basés sur un grand volume de données annotées. Ce dernier nécessite l'intervention humaine pour l'annotation des corpus, ce qui s'avère un processus long et coûteux.

Dans le travail de (Hatami *et al.*, 2021), ils proposent une approche basée sur la projection translingue. Cette approche consiste à faire un transfert des annotations depuis une langue, riche en ressources linguistiques, considérée comme langue

source, vers une langue cible peu dotée, et ce, grâce à un alignement de mots dans des corpus de textes parallèles. Par conséquent, les résultats des annotations obtenus sont utilisés pour un entraînement supervisé dans la langue cible pauvre en ressources linguistiques. Cependant, les performances des approches basées sur la projection des annotations linguistiques dépendent fortement de l'efficacité des algorithmes d'alignement des mots.

D'autre part, de nombreux travaux se sont intéressés aux plongements des mots bilingues (Chandar *et al.*, 2014; Luong *et al.*, 2015), estimant qu'il est possible d'apprendre les représentations des mots de concepts similaires s'ils sont très proches dans l'espace partagé. Pourtant, cette méthode nécessite des dictionnaires bilingues dont la traduction dépend fortement de leur qualité.

L'inuktitut est une langue polysynthétique ou agglutinante plus riche lorsqu'elle est comparée à l'anglais. La langue inuktitut se distingue des autres langues par la complexité élevée de ses mots, qui se manifeste de différentes manières : l'incorporation de noms qui consiste à former un mot à partir d'une combinaison d'un verbe et de son sujet ; l'incorporation de verbe qui forme un mot en le combinant à un autre verbe ; les adjectifs et les adverbes peuvent aussi faire partie d'un long mot. Aussi, contrairement à la majorité des langues, les marqueurs grammaticaux dans la langue inuktitut ne se présentent pas sous forme de petits mots, mais sont plutôt formés de mots plus longs. L'exemple cité dans (Compton, 2021) est :

" Illujjjuaraalumuulaursimannginamalittauq " qui signifie en français, *"mais aussi, parce que je ne suis jamais allé à la très grande loge"*.

Les caractéristiques de la langue inuktitut et la rareté des données disponibles représentent un défi majeur pour l'application des méthodes du TALN, notamment, la reconnaissance des entités nommées.

1.2 Objectifs et contributions

Notre principal objectif dans projet est de relever les défis de la langue inuktitut et de détecter les EN pour cette langue à travers les contributions suivantes :

- Explorer le domaine de la REN pour la langue inuktitut. À notre connaissance, les travaux sur cette tâche en liaison avec les langues autochtones comme l'inuktitut, sont presque inexistants. De ce fait, notre étude sera la première à être réalisée pour cette tâche. L'enjeu de cette tâche est de pouvoir aligner les mots pour la langue inuktitut dont la morphologie est complexe (polysynthétique).

-Effectuer une étude comparative entre deux méthodes : (i) une projection basée sur les règles en utilisant un analyseur morphologique et les outils d'alignement des mots, (ii) les plongements des mots bilingues en utilisant la similarité sémantique dans un espace vectoriel bilingue. Le défi pour cette méthode est de transférer les connaissances de la langue anglaise vers la langue inuktitut, deux langues différentes quant à la morphologie et la richesse.

- Construire un corpus annoté en langue inuktitut pour la tâche de détection des entités nommées. Cette contribution facilitera la mise en place de travaux futurs pour divers sous-domaines du TALN, à savoir l'extraction d'informations, la traduction automatique neuronale et les agents conversationnels (Chatbots). Aussi, ce travail contribuera à la préservation et revitalisation de la langue inuktitut ainsi que d'autres langues autochtones.

- Bien que la traduction des EN soit une tâche très importante pour le TALN, les systèmes de TAN n'offrent pas toujours la traduction correcte des EN. C'est pour cela que nous tentons d'améliorer la traduction automatique avec la reconnaissance des EN dans la langue source et leurs traductions vers la langue cible.

1.3 Organisation de l'ouvrage

Dans ce premier chapitre, nous avons présenté le contexte général de notre travail, les enjeux et les défis à relever ainsi que les contributions de nos études. Le reste du rapport de mémoire est organisé comme suit :

Le chapitre 2 présente en détail les concepts de bases impliqués dans l'étude de notre projet et est organisé en 6 sections. La langue inuktitut, son système d'écriture, ses caractéristiques et ses spécificités sont expliqués dans la première section. La deuxième section explique la notion de reconnaissance des entités nommées. La troisième section décrit l'alignement des mots. La quatrième section, quant à elle, présente les prolongements des mots (en anglais *word embeddings*). La cinquième section explique la projection d'annotation pour les REN. Et enfin, la dernière section introduit différentes approches d'apprentissage automatique.

Le chapitre 3 est un état de l'art qui fait un survol des travaux liés à notre problématique. Ce chapitre présente, dans un premier temps, les travaux effectués dans le domaine de la reconnaissance des entités nommées pour les langues peu dotées et les modèles d'alignement. Il décrit, par la suite, les différentes approches utilisées pour la projection d'annotation. Pour finir, il présente les recherches les plus pertinentes dans le domaine de la traduction automatique en utilisant des dictionnaires bilingues des EN.

Le chapitre 4 décrit notre méthodologie et présente les deux approches proposées. La première section détaille une méthode basée sur les règles et spécifie les différents concepts utilisés dans cette approche. La deuxième section présente la méthode basée sur les plongements des mots bilingues. La troisième section discute des entités nommées pour la traduction automatique.

Les résultats d'évaluation de nos méthodes proposées sont énumérés et expliqués

dans le chapitre 5. Une étude comparative des deux approches est présentée après la définition des données et des métriques utilisées pour l'évaluation de nos modèles.

Enfin, la conclusion générale résume le travail présenté dans ce mémoire et les principales contributions de notre travail. Les perspectives et travaux futurs sont aussi discutés dans cette partie.

CHAPITRE II

NOTIONS PRÉLIMINAIRES

Dans ce chapitre, nous élaborons les notions de bases fondamentales liées à notre problématique. Nous commençons d'abord par introduire le savoir autochtone, la recherche autochtone, la langue, la culture et l'identité. Ensuite, nous définissions la langue inuktitut et son écriture. Enfin, nous définissons le système de reconnaissances d'entités nommées ainsi que ses outils et notions nécessaires à sa réalisation, et ce d'une manière monolingue et translingue.

2.1 Les savoirs autochtones

Les peuples autochtones sont les premiers peuples d'Amérique du Nord. Trois groupes de peuples autochtones reconnus par la Constitution canadienne (gouvernement du Canada, 2022b), sont :

— **les membres des Premières Nations :** (gouvernement du Canada, 2021c) rapporte que 630 collectivités des Premières Nations ont été recensées, soit plus de 50 Nations et 50 langues différentes.

— **les Inuits :** le peuple autochtone de l'Arctique. Ils vivent dans 53 communautés des régions du Nord du Canada dans l'Inuit Nunangat, dont quatre régions : les Territoires du Nord-Ouest et Yukon (l'Inuvialuit) ; le Nord du

Québec (le Nunavik) ; le Nunatsiavut ; et le Nunavut (gouvernement du Canada, 2021a).

— **les Métis :** (gouvernement du Canada, 2021b) rapporte que Statistique Canada ont recensé, en 2016, 587 545 Canadiens déclarés Métis.

Ces trois différents groupes de peuples autochtones ont chacun ses propres langues, sa propre histoire et sa propre culture. Cette dernière, prend en compte l'importance et la force du système de connaissance et du savoir des peuples autochtones.

2.1.1 La recherche autochtone

Au Canada, le Conseil de Recherches en Sciences Humaines (CRSH) est l'organisme subventionnaire fédéral qui soutient et aide à la recherche et la formation universitaires dans le domaine des sciences humaines (gouvernement du Canada, 2022a). Le CRSH s'est engagé à soutenir et à encourager la recherche autochtone et la définit comme suit :

" Recherche réalisée dans n'importe quel domaine ou discipline qui est menée « par et avec » des communautés, des sociétés ou des personnes des Premières Nations, des peuples inuit ou métis ou d'autres nations autochtones et qui les concerne et repose sur leur sagesse, leurs cultures, leurs expériences ou leurs systèmes de connaissances exprimés dans des formes dynamiques, passées et actuelles. La recherche autochtone peut englober les dimensions intellectuelles, physiques, émotionnelles et (ou) spirituelles du savoir de manière à créer des liens créatifs entre les personnes, les endroits et l'environnement naturel. " (gouvernement du Canada, 2022a)

En 2020 de nouvelles orientations pour la recherche autochtone ont été mises en place (comité de coordination de la recherche au Canada, 2020), dans le but

d'aider des Métis et des Inuits à mener leurs propres recherches et à établir des partenariats avec l'ensemble du milieu de la recherche (gouvernement du Canada Conseil de recherches en sciences humaines, 2022). Comme la langue représente un volet essentiel du savoir autochtone, elle fait partie aussi des disciplines importantes de la recherche autochtone au Canada. Dans la section suivante, nous présentons les différentes langues autochtones.

2.1.2 Langue, culture et identité

La langue est un moyen de communication entre différentes personnes et facilite l'interaction et le contact avec l'environnement, mais elle est également un élément principal et important de toute culture. Chaque langue est adaptée aux particularités de sa culture, il est dans ce cas difficile, voire impossible, pour une traduction de transmettre la signification exacte et entière de la langue originale. Les langues autochtones au Canada ont changé et ont évolué avec le temps et au fil des générations. Comme toutes les langues, elles portent des valeurs littéraires, culturelles, traditionnelles, mais aussi historiques. Une des particularités des langues autochtones du Canada est que, pour certaines, elles ne sont pas parlées ailleurs dans le monde et sont propres au Canada (Statistique Canada, 2017). De ce fait, ces langues ont besoin être préservées car elles représentent une des richesses linguistiques et donc culturelles du Canada.

Il est mentionné dans (Statistique Canada, 2017), que le recensement de 2016 a enregistré plus de 70 langues autochtones réparties en 12 familles linguistiques énumérées dans le tableau 2.1.

Les langues inuites sont considérées comme la deuxième famille linguistique autochtone ayant le plus grand nombre de locuteurs après les langues algonquiennes. La langue la plus utilisée dans cette famille linguistique est l'inuktitut, principale-

Familles linguistiques autochtones	**Population**	**Principales concentrations provinciales et territoriales**
Langues algonquiennes	175 825	Manitoba (21,7 %), Québec (21,2 %), Ontario (17,2 %), Alberta (16,7 %), Saskatchewan (16,0 %)
Langues inuites	42 065	Nunavut (64,1 %), Québec (29,4 %)
Langues athabascanes	23 455	Saskatchewan (38,7 %), Territoires du Nord-Ouest (22,9 %), Colombie-Britannique (18,4 %)
Langues salishennes	5 620	Colombie-Britannique (98,8 %)
Langues siouennes	5 400	Alberta (74,9 %), Manitoba (14,2 %)
Langues iroquoiennes	2 715	Ontario (68,9 %), Québec (26,9 %)
Langues tsimshennes	2 695	Colombie-Britannique (98,1 %)
Langues wakashanes	1 445	Colombie-Britannique (98,6 %)
Le mitchif	1 170	Saskatchewan (41,9 %), Manitoba (17,5 %)
L'haïda	445	Colombie-Britannique (98,9 %)
Le tlingit	255	Yukon (76,5 %), Colombie-Britannique (21,6 %)
Le kutenai	170	Colombie-Britannique (100,0 %)

TABLEAU 2.1 – Population d'identité autochtone selon la famille linguistique. (Source : (Statistique Canada, 2017))

ment parlée au Nunavut et au Québec. Dans notre travail de recherche, on s'intéresse particulièrement à cette langue, riche et en même temps complexe, présentée dans la section suivante.

2.1.3 La langue inuktitut

Les langues autochtones du Canada sont considérées comme des langues en danger. Selon l'UNESCO, 75% de ces langues sont en voie de disparition et ne sont parlées que par des aînés (Gouvernement du Canada, 2019). L'inuktitut est l'un des quatre principaux ensembles de dialectes des langues inuites du Canada, qui s'étend de l'Alaska au Groenland. Principalement parlée au Nunavut et au Québec, elle est aussi parlée dans les zones de Terre-Neuve-et-Labrador ainsi que dans les Territoires du Nord-Ouest. En 2016, le recensement a comptabilisé 39 770 locuteurs, avec 65% qui vivaient au Nunavut et $30,8\%$ au Québec. La figure 2.1 montre la carte géographique de la langue inuite.

La préservation des langues inuites un défi, car ce sont des langues qui ne sont pas parlées ailleurs dans le monde et leur transmission aux générations futures n'est pas facile. En effet, Statistiques Canada rapportent qu'en 2006, 21.4% de la population autochtone a déclaré être dans la capacité d'entretenir une conversation dans une langue autochtone. Néanmoins, ce pourcentage a diminué à $15,6\%$ en 2016.

Le tableau 2.2 présente le nombre de locuteurs de la langue inuktitut dans chacune des quatre régions, la population inuite totale de chaque région et la proportion de la population qui déclare parler l'inuktitut.

FIGURE 2.1 – Carte de la langue inuite au Canada. (Source : (Compton, 2021))

Région	Population inuite	Nombre de locuteurs approximatif	Pourcentage de résidents qui parlent l'inuktut
Région désignée des Inuvialuit	3 110	685	22%
Nunavut	30 140	26 840	89%
Nunavik	11 800	11 705	99%
Nunatsiavut	2 285	490	21%
Inuit Nunangat (Total)	47 335	39 715	84%

TABLEAU 2.2 – Population inuite selon la connaissance de la langue inuktitut. (Source : (Argetsinger, 2017))

2.1.3.1 Écriture de la langue inuktitut

L'inuktitut étant une langue de tradition orale, elle a développé un système d'écriture. En effet, (Drapeau, 2011) rapporte que *"Les partisans de l'alphabétisation considèrent que les langues minoritaires non écrites ou ne possédant pas de système d'écriture standardisé courent un plus grand risque de disparaître à plus ou moins long terme que celles qui jouissent d'une tradition littéraire bien établie."*

L'inuktitut s'écrit avec un système syllabique : Elle possède aussi une orthographe basée sur l'alphabet romain et l'orientation d'écriture des phrases se fait, comme pour le français ou l'anglais, de gauche à droite.

Le système d'écriture syllabique représente les syllabes sous différents symboles, comme la combinaison de voyelle et de consonne *pa* de l'alphabet romain. Cette écriture est représentée dans le système d'écriture inuktitut par le symbole (Compton, 2021). Ce système représente les consonnes par des symboles et les voyelles u, i et a, par l'orientation (tournée ou inversée) de ses symboles. Une syllabe peut avoir le son initial de g, j, k, l, m, n, p, q, r, s, t, v, ng, ł ou aucune consonne. Si on prend l'exemple du caractère syllabique $\subset$, représentant le *ta* de l'alphabet romain, en inversant le symbole, $\supset$ représente le son *tu* et finalement $\cap$ c'est le *ti*. Dans la langue inuktitut les voyelles peuvent être courtes ou longues et la distinction se fait par un point au-dessus du caractère syllabique pour les voyelles longues. Une consonne suivie d'une autre consonne ou écrite seule à la fin d'un mot est représentée par une version petite de son symbole. La figure 2.2, montre un tableau qui résume le syllabaire inuktitut.

FIGURE 2.2 – Écriture syllabique inuktitut. (Source : www.espace-inuit.org)

Le syllabaire inuktitut connaît des différences entre les dialectes. Cela est dû au fait que certains sons existent dans un dialecte et pas dans l'autre. Cette caractéristique est retrouvée aussi dans le système d'épellation ou l'orthographe de l'alphabet romain de la langue inuktitut, ces différences sont représentées par des symboles additionnels. L'orthographe de la langue inuktitut, basée sur les lettres de l'alphabet romain, vise à être plus fidèle aux prononciations et aux spécificités de la langue afin d'être standardisée et rendue plus systématique (Compton, 2021).

2.1.3.2 Caractéristiques et défis de la langue inuktitut

La langue inuktitut possède une grammaire particulière et des compositions de mots assez complexes qui la différencient des autres langues. Exemple [1] :

Tusaatsiarunnanngittualuujunga, qui signifie ***Je n'entends pas très bien***

Peut être décomposé comme suit :

La racine **Tusaa**- (entendre) suivie de 5 suffixes :

tsiag- (bien), -**junnag**- (être capable), -**nngit**- (négation), -**tualuu**- (beaucoup), -**junga** (marque de la première personne du singulier et du présent).

Phonétique : La langue inuktitut possède seulement 3 voyelles et 14 consonnes, un ensemble phonétique assez restreint comparé aux autres langues comme le français par exemple avec un total de 26 lettres, 6 voyelles et 20 consonnes.

Ergativité : La langue inuktitut adopte un système ergatif-absolutif dans la manière de former les participants dans une phrase (Compton, 2021). Cela signifie

1. `https://www.mustgo.com/worldlanguages/inuit/` (consulté en décembre 2022)

que le sujet est formé d'une manière particulière au lieu de garder le même sujet lors d'une proposition transitive.

Polysynthèse : Les langues polysynthétiques sont parlées partout dans le monde et sont richement représentées parmi les peuples autochtones du Nord et du Sud américain (Bird, 2009). Pour les langues autochtones, la structure des mots est longue et complexe. Cette complexité réside dans différentes caractéristiques :

- Des mots sont formés à partir d'une combinaison de verbe et de son objet, appelée l'incorporation nominale ;
- Des mots sont formés en combinant deux verbes ;
- Un mot de structure longue peut contenir des adverbes ou des adjectifs ;
- Les marqueurs grammaticaux sont inclus dans des mots complexes, formant des mots de taille plus grande(Compton, 2021).

2.1.3.3 Toponymie inuite

La toponymie est l'étude des noms propres qui désignent un lieu (les toponymes). Cette discipline vise à rechercher la signification des toponymes, ainsi que leur étymologie, leur ancienneté, leur évolution et leurs rapports avec la langue. Les toponymes inuits sont qualifiés dans (Collignon, 2002) par une description du territoire, c. -à-d. que les noms des lieux inuits sont adaptés aux caractéristiques de la nature du milieu arctique et décrivent ses différents paysages et reliefs. Selon (Collignon, 2002), pour les inuits, les toponymes sont une assistance indispensable au déplacement et sont considérés comme des cartes au voyageur occidental.

Au Québec, a partir de 1977, la Commission de toponymie a établi un programme d'officialisation des noms géographiques d'origine autochtone, en initiant un inventaire de recherche de conservation de milliers de toponymes autochtones (gouvernement du Québec, 2022). Cette collecte de toponymie autochtone est diffusée

sur le Web et est actuellement accessible dans la banque de noms de lieux du Québec (Gouvernement du Québec, 2012).

Les langues autochtones du Canada font partie des 7 000 langues les plus menacées au monde (Bird, 2009). Dans le but de permettre la diversité linguistique et le multilinguisme dans le monde, l'UNESCO encourage la recherche et le développement des technologies pour les langues autochtones (Unesco, 2019). Ce qui contribuera aux utilisateurs d'avoir accès à l'information et aux connaissances dans les langues maîtrisées. La REN est une tâche qui est considérée comme une technique de pré-traitement nécessaire pour la réalisation de multiples applications liées au TALN. De ce fait, la REN ouvrira des portes pour la revitalisation et la préservation de la langue et la culture autochtone.

2.2 Reconnaissance des entités nommées

La reconnaissance d'entités nommées est le processus d'identification des entités telles que le nom d'une personne, le nom d'une organisation, le lieu, les mesures et les quantités et les modèles de chaîne comme les adresses e-mail, les numéros de téléphone ou les adresses IP (Yadav et Bethard, 2019). Ces entités nommées sont importantes dans la majorité des applications de traitement du langage naturel. Par conséquent, les systèmes de reconnaissance des entités nommées sont généralement considérés comme une première étape pour la recherche d'information, la modélisation de sujets et la réponse aux questions (Yadav et Bethard, 2019).

La figure 2.3 montre un exemple d'un système REN, qui prend une séquence de mots S en entrée et retourne 2 entités nommées détectées dans cette phrase, en sortie. En effet, les données textuelles contiennent beaucoup d'ambiguïté. Par exemple, le mot *Sydney* peut faire référence à un lieu et à un prénom de personne.

FIGURE 2.3 – Exemple d'un modèle de système REN.

Plusieurs outils de reconnaissance d'entités nommées, sont disponibles et facilement accessibles en ligne. Le tableau 2.3 classifie ses outils proposés par le milieu académique et le domaine de l'industrie (Li *et al.*, 2018).

Systèmes REN	URL
StanfordCoreNLP	https ://stanfordnlp.github.io/CoreNLP/
OSU Twitter NLP	https ://github.com/aritter/twitter$_nlp$
Illinois NLP	http ://cogcomp.org/page/software/
NeuroNER	http ://neuroner.com/
NERsuite	http ://nersuite.nlplab.org/
Polyglot	https ://polyglot.readthedocs.io
Gimli	http ://bioinformatics.ua.pt/gimli
spaCy	https ://spacy.io/api/entityrecognizer
NLTK	https ://www.nltk.org
OpenNLP	https ://opennlp.apache.org/
LingPipe	http ://alias-i.com/lingpipe-3.9.3/
AllenNLP	https ://demo.allennlp.org/
IBM Watson	https ://natural-language-understandingdemo.ng.bluemix.net
FG-NER	https ://fgner.alt.ai/extractor/
Intellexer	http ://demo.intellexer.com/
Repustate	https ://repustate.com/named-entityrecognition-api-demo
AYLIEN	https ://developer.aylien.com/text-api-demo
Dandelion API	https ://dandelion.eu/semantic-text/entityextraction-demo
displaCy	https ://explosion.ai/demos/displacy-ent
ParallelDots	https ://www.paralleldots.com/namedentity-recognition
TextRazor	https ://www.textrazor.com/named $_entity_recognition$

TABLEAU 2.3 – Quelques outils REN disponibles en ligne, en date de : Août 2022.

2.3 Apprentissage automatique pour la reconnaissance d'entités nommées

Les modèles de reconnaissance d'entités nommées basés sur l'apprentissage automatique, ayant l'avantage de découvrir les caractéristiques cachées liées aux textes, ont réussi à atteindre de bonnes performances par rapport aux systèmes de REN basés sur les méthodes classiques. Dans ce qui suit, nous décrivons les modèles d'apprentissage automatique les plus utilisés pour la tâche de la REN.

2.3.1 Réseaux de neurones récurrents

Un réseau de neurones récurrents (RNR) est une classe de réseaux de neurones artificiels dans laquelle les connexions entre les nœuds forment un graphe le long d'une séquence temporelle (Du *et al.*, 2019). Les RNR traitent des séquences d'entrées de longueur variable en utilisant leur état de mémoire. Ainsi, ces modèles sont bien adaptés aux tâches du TALN, telles que la classification et la génération de texte, l'étiquetage de séquence ainsi que la reconnaissance des entités nommées, où les phrases en entrée sont de longueurs variables. La figure 2.4 présente un exemple d'un réseau de neurones récurrents A qui traite une séquence de taille variable, ce modèle prend en entrée une séquence x_t et produit en sortie h_t .

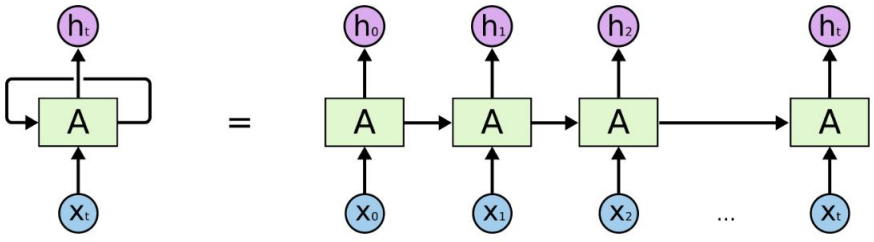

FIGURE 2.4 – Réseau de neurones récurrents. (Source : (Du *et al.*, 2019))

L'un des attraits des RNN est l'idée qu'ils pourraient être en mesure de relier les informations précédentes à la tâche actuelle. Cependant, appliquant l'algorithme de rétropropagation du gradient pour l'apprentissage des paramètres, ces modèles souffrent de l'effet du *problème du gradient de fuite* (*vanishing gradient*) lorsque la valeur du gradient prend une valeur petite et d *gradient explosif* (*exploding gradient problem*) dans le cas contraire (Bengio *et al.*, 1994).

2.3.2 Réseaux de neurones récurrents avec mémoire à court long terme

Introduit par (Hochreiter et Schmidhuber, 1997), les modèles LSTM (en anglais *Long short-term memory*) ont été appliqués avec succès à diverses prédictions de séquences et à la tâche d'étiquetage de séquences. Il a été démontré que les modèles LSTM sont plus performants que les RNN pour l'apprentissage de langages sans contexte et sensibles au contexte et pour contourner *le problème du gradient explosif* (Gers et Schmidhuber, 2001). En effet, ces modèles permettent la résolution du problème de dépendance à long terme. Lors du traitement des éléments, LSTM retient l'information pertinente pour la tâche à travers le temps, ce processus étant appelé la cellule mémoire (en anglais *The memory cell*).

Le modèle LSTM contient des blocs de mémoire, dans la couche cachée récurrente, qui contiennent à leur tour des cellules mémoires, avec des autoconnexions, qui enregistrent l'état temporel du réseau en plus et qui sont maintenus grâce au principe de portes (en anglais *gates*) pour contrôler le flux d'informations. Chaque bloc mémoire contient une porte d'entrée qui contrôle le flux d'activations d'entrée dans la cellule mémoire, une porte de sortie contrôle le flux de sortie de la cellule activations dans le reste du réseau et une porte d'oubli qui met à l'échelle l'état interne de la cellule avant de l'ajouter comme entrée de la cellule via la connexion autorécurrente de la cellule, oubliant ou réinitialisant ainsi la mémoire de la cellule

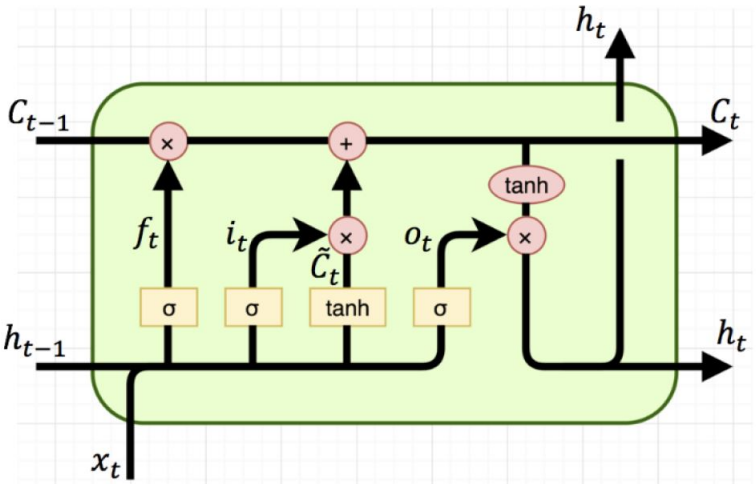

FIGURE 2.5 – Cellule LSTM. (Source : (Olah, 2015))

(Sak *et al.*, 2014). La figure 2.5 montre l'architecture d'une cellule LSTM.

Un réseau LSTM calcule, de manière itérative de $t = 1$ à T, les activations des unités du réseau comme suit :

$$
\begin{aligned}
\mathrm{i}_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i),\\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f),\\
c_t &= f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c),\\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_{t-1} + b_o),\\
m_t &= o_t \odot h(c_t),\\
y_t &= \emptyset(W_{ym}m_t + b_y),
\end{aligned}
$$

où, les indices i, f, o, et c représentent, respectivement, la porte d'entrée, la porte d'oubli, la porte de sortie et le vecteur d'activation de cellule, tous de même taille que le vecteur d'activation de sortie de cellule m, g et h sont les fonctions d'activation d'entrée et de sortie de cellule, les termes W désignent les matrices de poids, les b représentent les vecteurs de biais, σ est la fonction sigmoïde logistique, l'opérateur $\odot$ représente le produit (élément par élément) des vecteurs et $\emptyset$ est la

fonction d'activation de la sortie du réseau (Sak *et al.*, 2014).

2.3.3 Transformeurs

Un transformeur, principalement utilisé dans les domaines du TALN, est un modèle d'apprentissage profond qui adopte le mécanisme de l'auto-attention, en pondérant différemment l'importance de chaque partie des données d'entrée. Conçus pour traiter des données d'entrée séquentielles, les transformeurs sont bien adaptés aux tâches de traitement de langage naturel. À la différence des RNR, les modèles transformeurs traitent l'intégralité de l'entrée en une seule fois. Ils sont basés sur le mécanisme d'attention qui fournit un contexte pour toute position dans la séquence d'entrée, c.-à-d. permettre au modèle de tirer parti de l'état à n'importe quel point précédent, le long de la séquence (Graves, 2013).

Ce modèle a une structure encodeur-décodeur, où l'encodeur associe une séquence d'entrée de représentations de symboles $(x_1, x_2, ..., x_n)$ à une séquence de représentations continues $(z_1, z_2, ..., z_n)$ et en sortie une séquence de symboles $(y_1, y_2, ..., y_n)$ un élément à la fois. À chaque étape, lors d'une nouvelle génération de symboles, le modèle prend en entrée supplémentaire les symboles générés précédemment. Le modèle est auto-régressif (Graves, 2013).

La figure 2.6 illustre l'architecture du transformeur, qui utilise une auto-attention empilée et ponctuelle ainsi que des couches connectées pour l'encodeur et le décodeur.

Encodeur : composé de $N = 6$ couches identiques empilées, chacune a deux sous-couches. La première sous-couche est un mécanisme d'auto-attention à plusieurs têtes, la deuxième est un réseau de rétroaction entièrement connecté en fonction de

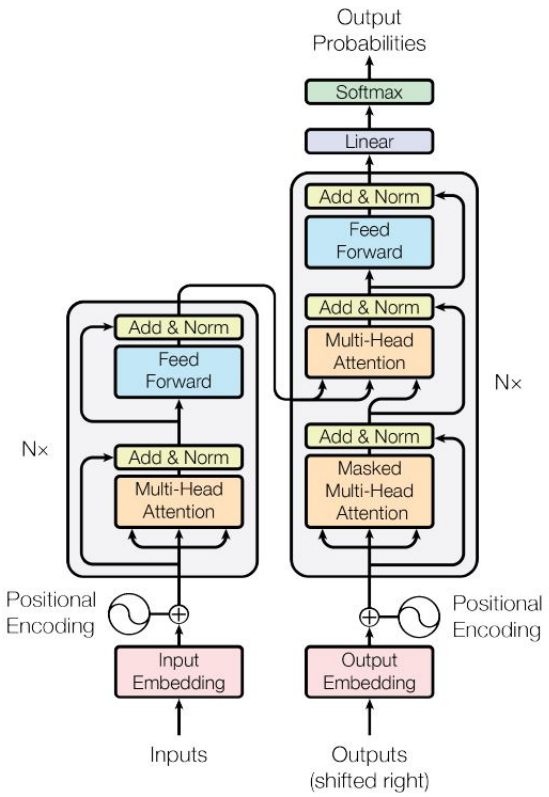

FIGURE 2.6 – Modèle d'architecture du transformeur. (Source : (Vaswani *et al.*, 2017))

la position. $SousCouche(x)$, étant la fonction implémentée par la sous-couche, la sortie de cette dernière est $NormDeCouche(x + SousCouche(x))$ (Vaswani *et al.*, 2017). Dans ce modèle, les sorties de toutes les sous-couches sont de dimension $d = 512$, afin de faciliter les connexions résiduelles.

Décodeur : également composé d'un empilement de $N = 6$ couches, en plus des deux sous-couches, le décodeur insère une troisième sous-couche qui effectue l'attention à plusieurs têtes sur la sortie de la pile d'encodeurs. Aussi, la sous-couche de l'auto-attention est modifiée pour empêcher les positions d'occuper des positions ultérieures. De ce fait, les prédictions pour la position i ne peuvent dépendre que des sorties connues aux positions inférieures à i.

La fonction d'attention consiste à lier une requête et un ensemble de paires clé-valeur à une sortie. Cette dernière est calculée comme une somme pondérée des valeurs et le poids de chaque valeur est calculé par une fonction de compatibilité des requêtes avec la clé correspondante. La requête, les clés, les valeurs et la sortie sont toutes modélisées sous forme de vecteurs.

À l'ère de l'apprentissage automatique moderne, l'apprentissage profond reste la technique la plus utilisée pour le développement des outils. Par ailleurs, plusieurs tâches de TALN sont basées sur les outils d'alignement des mots tels que la projection d'annotation pour étendre les ensembles de données ainsi que l'amélioration de la traduction.

2.4 Alignement des mots

L'alignement des mots est le processus d'identification automatique de traduction ou la correspondance entre les mots dans deux langues différentes. Ce processus est généralement représenté comme un ensemble de liens entre les phrases

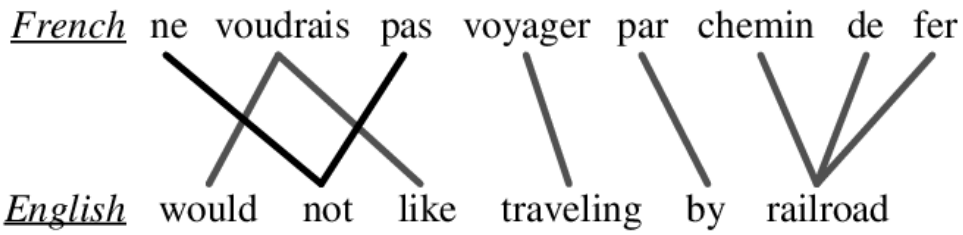

FIGURE 2.7 – Exemple d'alignement des mots. (Source : (Du *et al.*, 2019))

et les mots d'un segment d'une langue source et une langue cible. Un Alignement de mots peut être complet ou incomplet, il est dit complet lorsque tous les éléments des deux segments sont reliés à des éléments correspondants dans l'autre langue. Dans le cas contraire, l'alignement est incomplet. L'alignement des mots représente une étape essentielle pour la traduction automatique statistique ainsi que pour la traduction automatique neuronale (Dufter *et al.*, 2018). Il est utilisé dans de nombreuses tâches telles que la projection d'annotation (Huck *et al.*, 2019) et la création des plongements de mots multilingues (Dufter *et al.*, 2018). La figure 2.7 montre un exemple d'alignement entre deux phrases français/anglais. L'application du processus d'alignement des mots passe par trois grandes étapes (Tiedemann, 2003) :

- *La modélisation* : Trouver un modèle approprié pour l'alignement des textes en langues source et cible ;
- *Estimation des paramètres* : estimer les paramètres du modèle choisi ;
- *Récupération de l'alignement* : trouver l'alignement optimal des phrases et des mots, selon le modèle et les paramètres choisis, pour une traduction donnée.

Dans ce qui suit, nous décrivons les modèles d'alignement des mots.

2.4.1 Modèles d'alignement des mots

2.4.1.1 Les modèles IBM

Les modèles IBM sont des modèles d'alignement complexe utilisés dans la traduction automatique statistique (Moore, 2004).

— **Modèle IBM1** : probabilités d'alignement lexical. Ce modèle est faible en termes de réorganisation. En effet, il traite tous les types de réorganisation comme également possibles. Or, une séquence de mots qui se suivent, dans une langue donnée, aurait un ordre différent dans une autre langue (Moore, 2004).

— **Modèle IBM2** : positions absolues. Ce modèle résout le problème du modèle 1. Il utilise une distribution de probabilité d'alignement, notée a, pour modéliser la traduction d'un mot d'une langue source en position p_s en un mot en position p_c de langue cible. Cette probabilité est formulée comme suit :

$$a(p_s \vee p_c, l_s, l_c)$$

où l_s et l_c représentent, respectivement, la longueur de la phrase en entrée et la longueur de la phrase après la traduction (Dyer *et al.*, 2013).

— **Modèle IBM3** : généralement la traduction d'un mot produit un seul mot en sortie. Cependant dans certains cas un mot peut être supprimé ou traduit en plusieurs mots. Ce problème a été résolu dans le modèle 3 qui prend en charge les insertions.

— **Modèle IBM4** : positions relatives. Dans ce modèle chaque mot dépend du mot précédemment aligné et des classes des mots environnants.

— **Modèle IBM5** : corrige les lacunes en améliorant le modèle d'alignement avec plus de paramètres d'entraînement. Ce modèle garantit qu'aucun mot

ne peut être aligné sur la même position.

2.4.1.2 Modèle de Markov caché

Le modèle de Markov caché à été utilisé dans (Vogel *et al.*, 1996b) afin de développer une approche comportant des probabilités de traduction lexicale et un alignement relatif. Ce modèle introduit l'association $j \rightarrow a_j$, qui assigne un mot f_j en position j à un mot e_i en position $i = a_j$. Dans ce cas, le modèle mathématique devrait essayer de capturer la forte dépendance de a_j sur le précédent alignement. De ce fait, la probabilité d'alignement a_j pour la position j, devrait dépendre de l'alignement précédent a_{j-1}. En considérant I la longueur de la phrase de la langue cible, la probabilité d'alignement est formulée comme suit (Vogel *et al.*, 1996b) :

$$P(a_j|a_{j-1}, I)$$

Cette probabilité peut être reformulée en introduisant les alignements "cachés" (Jelinek, 1976) $a_1^j = a_1, ..., a_j, ..., a_J$, pour une paire de phrases $[f_1^J, e_1^I]$:

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} Pr(f_J, a_J|f_1^{j-1}, a_1^{j-1}, e_1^I).$$

2.4.2 Évaluation d'alignement de mots

Le taux d'erreurs d'alignement (TEA, en anglais *Alignement Error rate*) est une métrique qui mesure la qualité de la méthode d'alignement dont la performance est évaluée sur un ensemble de données de test préalablement établi par des annotateurs humains. Ces derniers marquent deux types d'alignements : un alignement sûr (S) et un alignement possible (P) pour les alignements ambigus. De

ce fait, l'alignement référence peut contenir des relations un mot-à-un mot, un mot-à-plusieurs et plusieurs-à-un mot (Deng et Byrne, 2008). Enfin, la qualité d'alignement (TEA) est mesurée comme suit :

$$TEA = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}. \tag{2.1}$$

La section suivante décrit une autre approche sur laquelle plusieurs modèles de projection d'annotation se basent actuellement.

2.5 Les plongements des mots

Le plongement des mots (en anglais *word embedding*) est une approche de représentation des mots et des documents, considérée comme l'une des principales percées de l'apprentissage profond sur les problèmes complexes du traitement automatique du langage naturel (Sigl, 2020). C'est une méthode d'apprentissage qui consiste à représenter les mots sous forme vectorielle, c.-à-d. chaque mot est représenté par un vecteur de nombre réel, afin de permettre à des mots ayant la même signification d'avoir une représentation similaire. En d'autres termes, dans un espace multidimensionnel avec un nombre de dimensions fixe, les mots ayant des sens sémantiques proches sont placés dans un espace proche et regroupés ensemble. La figure 2.8 montre un exemple de la représentation vectorielle de mots dans un espace à 3 dimensions. Les vecteurs des mots "*One*" et "*many*", sont plus proches l'un de l'autre que le vecteur du mot "example" car les deux mots sont reliés sémantiquement.

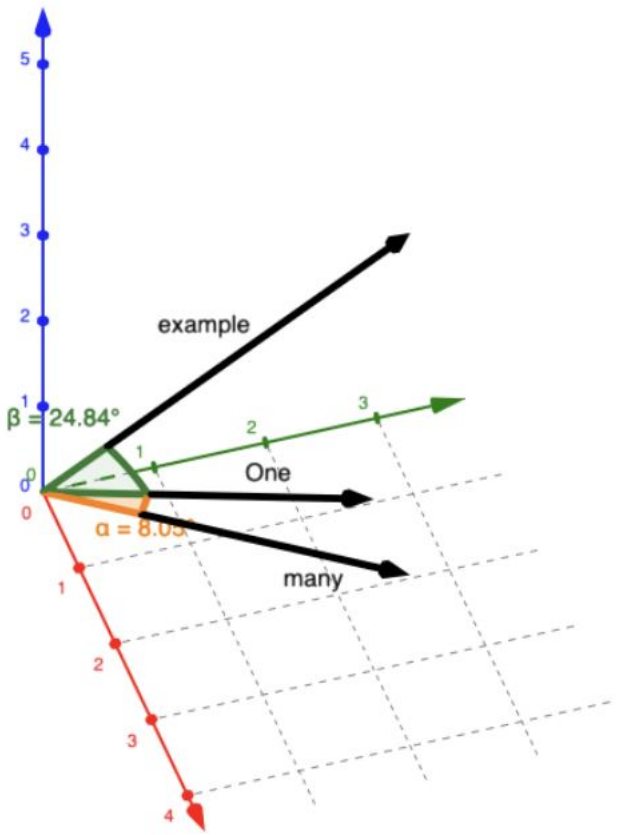

FIGURE 2.8 – Exemple de *word embedding*. (Source : (Sigl, 2020))

2.5.1 Méthodes de plongement des mots

À partir d'un corpus de texte, les méthodes de plongement des mots apprennent une représentation vectorielle à valeur réelle pour un vocabulaire de taille préalablement fixée. Les avancées de l'apprentissage non supervisé du word embedding permettent de manipuler des données massives non étiquetées en utilisant l'entraînement des réseaux de neurones (Chiu et Nichols, 2015). Parmi les méthodes du word embedding :

Word2vec :Basée sur une architecture simple, Word2vec est une approche statistique qui permet d'apprendre efficacement un mot autonome intégré, à partir d'un corpus de texte. Développée par (Mikolov *et al.*, 2013a), deux différents modèles d'apprentissage ont été proposés, le modèle *Continuous Bag-of-Words* (CBOW) et le modèle *Continuous Skip-Gram*. Le modèle CBOW apprend l'intégration en prédisant le mot cible en fonction de son contexte. Le modèle *Skip-Gram* apprend en prédisant les mots environnants étant donné le mot cible, au lieu de prédire le mot cible lui-même. La figure 2.9 montre l'architecture de l'approche *Word2vec* (modèles CBOW et *Skip-Gram*).

GloVe : Développé par (Pennington *et al.*, 2014), *Global Vectors for Word Representation* est un algorithme qui apprend efficacement les vecteurs de mots. Étant une extension de la méthode *word2vec*, l'approche *GloVe* utilise, conjointement, l'apprentissage basé sur le contexte local de la méthode *word2vec* et les statistiques globales des techniques de factorisation matricielle comme *l'analyse sémantique latente* (LSA). En utilisant des statistiques sur l'ensemble du corpus, *GloVe* construit une matrice explicite de contexte de mot donnant comme résultat un modèle d'apprentissage pour un meilleur plongement des mots.

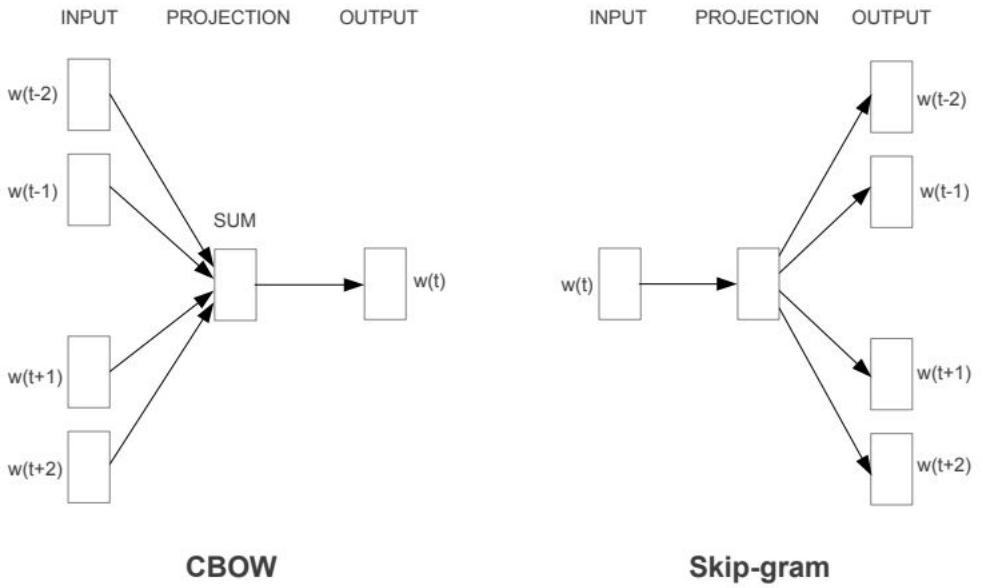

FIGURE 2.9 – Architecture des modèles CBOW et *Skip-Gram*. (Source : (Mikolov *et al.*, 2013a))

Les méthodes de plongement des mots présentées sont statiques. La représentation de chaque mot est déterminée après un entraînement. Le problème rencontré dans ces approches est la polysémie, c'est le cas où la sémantique d'un mot varie selon le contexte. Par exemple, le mot "*souris*" est connu comme le nom d'un mammifère rongeur, mais dans les rapports sur la technologie, il fait plus souvent référence à un dispositif de pointage pour ordinateur. Pour affronter ce problème et obtenir un plongement de mots contextualisé, les recherches récentes visent à inclure l'apprentissage des Word Embedding dans les modèles de langage neuronal (Jiao et Zhang, 2021). Ces approches permettent d'ajuster les vecteurs de représentation des mots selon de contexte d'entrée. La figure 2.10 illustre trois modèles de plongement des mots basés sur cette approche.

EMLO : Le modèle *Embeddings from Language Model* (Peters *et al.*, 2018), est

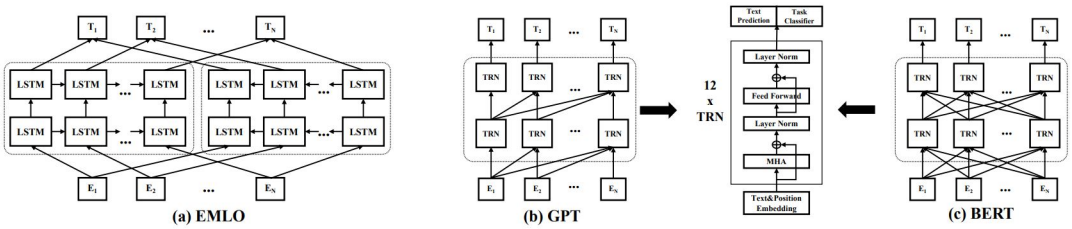

FIGURE 2.10 – Modèle EMLO, GPT et BERT. (Source : (Jiao et Zhang, 2021))

constitué d'un réseau *Long short-term memory* (LSTM) bidirectionnel ayant en entrée des mots représentés par des Word Embeddings statiques, préentraînés ou initialisés aléatoirement. *EMLO* utilise la prédiction des mots cibles, des phrases données, dans le but d'optimiser les réseaux LSTM. Ainsi, après l'entraînement, les états cachés du LSTM bidirectionnel peuvent être utilisés comme la représentation contextualisée des mots (Jiao et Zhang, 2021).

GPT : Étant très similaire à l'approche *EMLO*, le modèle *Generative Pre-training* est un préentrainement basé sur un modèle de langue. *GPT* utilise les couches de codage d'un Transformeur, un très puissant réseau de neurones avec mécanisme d'attention (Jiao et Zhang, 2021).

BERT : Le modèle BERT utilise également un réseau de type Transformeur à 12 couches et comprend une modélisation bidirectionnelle similaire à *EMLO*. Aussi, de nouvelles techniques sont considérées dans ce modèle, telles que, le plongement des positions (*position embedding*) qui est combiné au plongement statique des mots et saisi comme entrée du modèle. Ce modèle vise à améliorer la capacité de compréhension linguistique des réseaux avec un préentraînement non supervisé à grande échelle (Jiao et Zhang, 2021).

2.6 Conclusion

Dans ce chapitre, nous avons défini les concepts essentiels à la réalisation de notre projet consistant à la reconnaissance des entités nommées pour la langue inuktitut.

Le chapitre suivant présentera un état de l'art sur les travaux liés à notre projet de recherche.

CHAPITRE III

ÉTAT DE L'ART

Comme déjà mentionné dans le chapitre précédent, la REN est la tâche appartenant au domaine d'extraction d'information qui consiste à identifier et classifier des éléments d'informations textuels en des catégories prédéfinies appelées entités nommées (Yadav et Bethard, 2019). Autrement dit, la REN sert comme base pour de nombreux autres domaines cruciaux en TALN tels que les systèmes des réponses aux questions (Mollá *et al.*, 2006), la recherche d'information (Guo *et al.*, 2009), (Petkova et Croft, 2007), la compréhension du texte (Cheng et Erk, 2019), (Zhang *et al.*, 2019) et la traduction automatique (Babych et Hartley, 2003).De ce fait, la REN est une méthode qui peut traiter les scénarios pour les langues peu dotées dans l'extraction d'informations et la compréhension du langage naturel.

Les sections suivantes représentent les travaux liés aux langues autochtones ainsi que les travaux liés à la REN pour les langues riches en ressources et pour les langues peu dotées.

3.1 Langues autochtones

Les langues autochtones sont un moyen de transmettre la culture, le savoir et les visions autochtones à travers le monde. Cependant, ces langues sont menacées pour

des raisons historiques et sociales. La préservation des langues autochtones est primordiale pour assurer la protection de l'identité culturelle et les connaissances traditionnelles des peuples autochtones (Battiste, 2013). De plus, pour la plupart de ces langues, les données textuelles numérisées sont rares, voire inexistantes pour être utilisées comme données d'entraînement. Aussi, ces données proviennent généralement d'un domaine restreint (Littell *et al.*, 2018). Par conséquent, les systèmes entraînés sur ces données se généralisent difficilement avec d'autres ensembles de données. De ce fait, le conseil national de recherches du Canada a lancé le projet sur les technologies pour les langues autochtones canadiennes en collaboration avec les communautés autochtones afin de créer un logiciel qui vise à préserver et à étendre l'utilisation des langues autochtones (Kuhn *et al.*, 2020). Ce qui a donné naissance à plusieurs sous-projets, notamment la publication d'un corpus de phrases dans la langue inuktut alignées sur des phrases de l'anglais, l'étude d'un système de traduction automatique entraînés sur ce même corpus, la construction d'un outil de conjugaison de verbes pour les langues polysynthétiques hautement flexionnelles et un logiciel pour la prédiction de texte pour les langues autochtones (Kuhn *et al.*, 2020).

En plus des applications de TALN, la collecte des données est une tâche cruciale pour les langues en danger car elle joue un rôle très important dans la préservation et le développement de la diversité linguistique. Beaucoup de recherches se sont concentrées sur la construction des corpus pour plusieurs langues autochtones ; (Chiruzzo *et al.*, 2020) pour la langue guarani du Paraguay, (Duan *et al.*, 2020) pour la langue mapundungun du Chili, (Sierra Martínez *et al.*, 2020) pour les langues autochtones du Mexique et (Bustamante *et al.*, 2020) pour les langues autochtones du Pérou. D'autres ressources sont essentielles au développement des systèmes de TALN pour les langues autochtones, notamment les dictionnaires bilingues. (Hunt *et al.*, 2019) ont introduit un dictionnaire bilingue yupik-anglais.

le yupik est un groupe de langues de l'Alaska. Tandis que (Gutierrez-Vasques et Mijangos, 2017) se concentrent sur la tâche d'extraction automatique des dictionnaires bilingues pour la paire de langues espagnol-nahualt, langue autochtone du Mexique. (Collignon, 2004) se concentre sur la collection de toponymes inuits. Cette attention particulière a mené à la construction des bases de données des nom de lieu. Aussi, ce travail a mené au remplacement des noms des lieux anglais ou français par des noms inuits.

Les langues autochtones étant des langues polysynthétiques, elle ont généralement besoin d'analyse morphologique. Cette tâche, qui consiste à diviser les mots en leurs morphèmes constitutifs, a été largement étudiée dans le domaine du TALN. (Mager *et al.*, 2020) explorent deux modèles basés sur l'apprentissage profond entraînés sur l'allemand et l'indonésien, deux langues morphologiquement complexes comme les langues autochtones du Mexique (populca et tebehua). Le premier modèle proposé est un modèle générateur de pointeur LSTM (Sharma *et al.*, 2018) et le deuxième est un transducteur neuronal formé avec l'apprentissage par imitation (Makarov et Clematide, 2018). Cependant, les modèles neuronaux s'avèrent être performants sur cette tâche dans le cas où une grande quantité de données est disponible pour l'entraînement (Ruzsics et Samardžić, 2017).

(Chen et Schwartz, 2018) introduisent un analyseur morphologique de la langue yupik, ce dernier représente une adaptation des règles grammaticales et morphologiques de cette langue. (Le et Sadat, 2021) adaptent une approche non-supervisée pour la segmentation morphologique basée sur la grammaire adaptative pour l'inuinnaqtun, langue inuite du Canada.

La traduction automatique est l'une des tâches linguistiques fondamentales dans le domaine de la recherche. Il est couramment demandé par les peuples autochtones d'avoir de bons résultats de traduction automatique pour leurs langues. Par

conséquent, plusieurs recherches se sont intéressées à la traduction automatique pour les langues autochtones dans le but de sauver ces langues menacées ou en voie de disparition. (Zhang *et al.*, 2020) introduisent un corpus parallèle cherokee-anglais, pour faciliter la recherche de traduction automatique entre cherokee et anglais. (Mager *et al.*, 2018) étudient les traductions de trois langues polysynthétiques à faibles ressources soient nahuatl, wixarika et yorem nokki vers espagnol et inversement. (Feldman et Coto-Solano, 2020) mettent en oeuvre un modèle de traduction neuronal pour la langue chibchan (bribri), une langue autochtone de la Colombie. Quant à l'inuktitut, (Le et Sadat, 2020) étudient en premier lieu la segmentation des mots en utilisant un ensemble de caractéristiques et en exploitant les plongements pré-entraînés basés sur les mots et les (bi-)caractères. Ensuite, ils intègrent cette étape de pré-traitement dans un système de traduction automatique neuronale.

3.2 Modèles d'alignement

(Lin et Cherry, 2003) ont introduit cinq modèles d'alignement de mots appelés IBM [1..5] dédiés à la tâche de traduction automatique. Étant donné des paires de phrases, ces modèles fournissent la probabilité qu'un mot soit la traduction d'un autre mot dans l'autre langue. (Vogel *et al.*, 1996b) un modèle basé sur le modèle de Markov caché (HMM). (Lin et Cherry, 2003) (Vogel *et al.*, 1996b) sont des modèles statistiques qui s'appuient sur de grands corpus et qui utilisent des techniques non supervisées pour établir des alignements au niveau des mots. Dans le même contexte, GIZA++ (Och, 2003) est basée sur le modèle génératif ou les paramètres sont estimés en utilisant l'algorithme espérance-maximisation (EM), ce qui permet à ce modèle d'extraire, à partir des corpus parallèles, un dictionnaire bilingue sans l'utilisation de données annotées. Fast_align (Dyer *et al.*, 2013) est une variation du modèle IBM2 qui permet un alignement mot-à-mot

efficace et plus rapide. (Wang et Lepage, 2016) trouvent que Fast_align n'est pas aussi performant que GIZA++ quand il est appliqué à des langues avec ordre de mots distincts. Dans leur travail, ils réalignent les mots en utilisant une approche d'alignement de sous-phrases hiérarchique étant donné la matrice d'alignement des phrases. Cependant, ces techniques ne prennent pas en compte le contexte des mots et les mots à basse fréquence (Ho et Yvon, 2020).

Les approches neuronales permettent de résoudre ces problèmes grâce à leur avantage d'extraction des caractéristiques (Young *et al.*, 2017). (Chen *et al.*, 2020b) ont proposé un modèle d'alignement directionnel similaire au modèle IBM basé sur les RNR. L'historique d'alignement de ce dernier est représenté par des couches cachées connectées de manière récurrente. De plus, ils effectuent un apprentissage non supervisé en utilisant l'estimation par contraste de bruit. (Chen *et al.*, 2020a) introduisent MASK-ALign, un modèle d'alignement auto-supervisé qui saisit le contexte complet de la langue cible. Ce modèle masque les mots dans la phrase de la langue cible et prédit en s'appuyant sur les mots restants en langue source et cible. (Chen *et al.*, 2020b) montrent que les poids du mécanisme d'attention saisissent l'alignement des mots et présentent deux méthodes inductives d'alignement de mots (SHIFT-ATT et SHIFT-AET). Le principe de ces deux méthodes est d'induire des alignements lors de l'étape où le mot cible à aligner est dans l'entrée du décodeur. SHIFT-ATT induit l'alignement à partir des poids d'attention d'un transformeur et ne nécessite pas des paramètres de mise à jour. SHIFT-AET extrait les alignements à partir d'un module d'alignement secondaire qui est intégré dans un transformeur. (Zenkel *et al.*, 2020) présentent une approche neuronale qui utilise un modèle entraîné pour la tâche de traduction automatique supervisée pour une tâche d'alignement de mots non-supervisée. (Sabet *et al.*,

2020) proposent une méthode (SimAlign) qui ne nécessite pas de corpus parallèles. Leur idée clé consiste à exploiter les plongements de mots multilingues créés à partir des données monolingues exclusivement sans avoir recours à des données parallèles ou à des dictionnaires.

Tandis que les méthodes neuronales performent mieux que les méthodes statistiques quant aux langues riches en ressources, il est difficile de connaître leur niveau de performances en ce qui concerne les langues peu dotées, car les modèles neuronaux sont entraînés sur de larges données d'apprentissage. Dans le but d'améliorer l'alignement pour les langues peu dotées, (Tang *et al.*, 2018) introduisent un modèle de traduction automatique statistique qui apprend les corrélations bilingues qui est bénéfique à la traduction automatique pour les langues peu dotées, et ce, en utilisant la structure sémantique de la langue cible.(Pourdamghani *et al.*, 2018) proposent une approche basée sur les liens communs entre les paires de mots sémantiquement similaires. Ils construisent d'abord un dictionnaire bilingue en utilisant le modèle IBM puis l'ajoutent au corpus parallèle. (Marchisio *et al.*, 2021) introduisent le modèle GIZA++ avec plongements améliorés, qui prend en considération l'avantage de l'espace des plongements de mots monolingues de la langue source et la langue cible uniquement. (Asgari *et al.*, 2020) proposent un modèle d'alignement basé sur l'échantillonnage de sous-mots des unités de texte. L'hypothèse de cette méthode considère que l'agrégation des différentes granularités de texte pour certaines langues peut être bénéfique pour l'alignement des mots. Les défis des langues peu dotées se manifestent dans la quantité des données disponibles, (Conneau *et al.*, 2017) montrent la possibilité de construire un dictionnaire bilingue entre deux langues sans utiliser de corpus parallèles, en alignant l'espace de plongements des mots monolingues de manière non-supervisée.

3.3 Reconnaissance d'entités nommées

La première tâche de REN a été organisée par Grishman et Sundheim (1996) lors de la sixième conférence MUC-6. Depuis, les systèmes de REN ont été largement étudiés et développés. Les premiers systèmes de REN étaient basés sur des règles, des lexiques, des caractéristiques orthographiques et des ontologies. Cette approche n'utilise pas de données annotées, car elles reposent sur des règles élaborées à la main. Parmi les systèmes de reconnaissance d'entités nommées les plus connus qui sont principalement basés sur des règles syntaxiques et sémantiques, FASTUS (Appelt *et al.*, 1995), LaSIE-II (Humphreys *et al.*, 1998), NetOwl (Krupka et Hausman, 1998), Facile (Black *et al.*, 1998) et SAR (Aone *et al.*, 1998). Dans le domaine biomédical, le système ProMiner est proposé dans (Hanisch *et al.*, 2005). Ce système utilise un dictionnaire de synonymes prétraité dans le but d'identifier les occurrences potentielles de noms dans le texte biomédical et associer des identifiants de bases de données de protéines et de gènes aux correspondances détectées. (Kim et Woodland, 2000) ont proposé un système de reconnaissance d'entité nommée à base de règle qui utilise l'approche d'inférence de la règle de Brill pour la saisie vocale. Ce système génère automatiquement des règles basées sur l'étiquetage grammatical du tagueur Brill. Les systèmes basés sur des règles fonctionnent très bien lorsque le lexique est complet. Cependant, en raison des règles spécifiques à chaque domaine et à chaque langage, la précision est généralement élevée, mais le rappel est souvent faible. Ainsi, ces systèmes ne peuvent pas être appliqués dans d'autres domaines ou d'autres langues (Li *et al.*, 2022). Aussi, la réussite des systèmes REN basés sur les règles a besoin d'experts du domaine pour construire et maintenir les ressources du savoir.

Ces systèmes ont été suivis par des systèmes REN basés sur l'ingénierie des fonctionnalités et apprentissage automatique (Yadav et Bethard, 2019). (Nadeau *et al.*, 2006) ont proposé un système qui combine l'extraction d'entités et la désambiguïsation basée sur des heuristiques simples et efficaces, dans le but de créer un répertoire géographique et résoudre l'ambiguïté d'entité nommée. (Zhang et Elhadad, 2013) ont proposé une approche non supervisée pour extraire des entités nommées d'un texte biomédical. Le modèle proposé recourt à des terminologies, à des statistiques de corpus telles que la fréquence de document inverse, vecteurs de contexte et une connaissance syntaxique superficielle comme la segmentation des phrases nominales. Toujours dans le domaine biomédical, des travaux plus récents comme (Sachan *et al.*, 2018) entraînent un modèle de langage bidirectionnel (BiLM) formé de manière non supervisée en utilisant uniquement les données non étiquetées et transfère ses poids pour "pré-entraîner" un modèle de REN avec la même architecture que le BiLM. Les auteurs de cette technique ont montré qu'un tel pré-entraînement des poids du modèle REN est une bonne stratégie d'initialisation, car il conduit à des améliorations substantielles des scores F1 pour quatre ensembles de données référence. De plus, pour obtenir un score F1 particulier, le modèle pré-entraîné nécessite moins de données entraînées par rapport à un modèle initialisé aléatoirement. Ainsi, un modèle pré-entraîné converge également plus rapidement lors de la mise au point du modèle. (Bari *et al.*, 2020) proposent un modèle NER translingue non supervisé capable de transférer des connaissances NER d'une langue à une autre de manière totalement non supervisée, sans s'appuyer sur un dictionnaire bilingue ou des données parallèles, et ce grâce à l'apprentissage automatique contradictoire au niveau des mots peaufiné avec l'augmentation des caractéristiques. Des expériences sur cinq langues différentes ont démontré l'efficacité de cette approche.

Le modèle BERT, quant à lui, est introduit dans (Devlin *et al.*, 2019). Ce mo-

dèle est conçu pour pré-entraîner des représentations bidirectionnelles profondes à partir de texte non étiqueté en conditionnant conjointement les deux contextes, gauche et droit, dans toutes les couches. En conséquence, le modèle BERT pré-entraîné peut être affiné avec une seule couche de sortie supplémentaire pour créer des modèles de pointe pour un large éventail de tâches, notamment REN. (Luoma et Pyysalo, 2020) présentent une étude systématique en explorant l'utilisation d'ajout de contexte pour la reconnaissance d'entité nommée à l'aide de modèles BERT en cinq langues. Les auteurs rapportent que l'ajout de contexte sous forme de phrases supplémentaires à l'entrée BERT augmente systématiquement les performances REN. En effet, il permet d'étudier les prédictions des phrases dans différents contextes. Ils ont proposé une méthode simple, le vote contextuel à la majorité (en anglais *Contextual Majority Voting*, CMV), pour combiner ces différentes prédictions et améliorer les performances du REN. (Yamada *et al.*, 2020) proposent LUKE, un modèle de pré-traitement de représentation des mots et des entités contextualisées en se basant sur les transformeurs. Ce modèle traite les mots et les entités dans un texte donné en tant que tokens indépendants, et en produit des représentations contextualisées. LUKE est entraîné à l'aide d'une nouvelle tâche de pré-traitement basée sur le modèle de langue masqué (MLM) de BERT. Cette tâche consiste à prédire aléatoirement les mots et les entités masquées dans un grand corpus d'entités annotées récupérées de Wikipédia. Le modèle proposé atteint des résultats empiriques impressionnants sur un large éventail de têches REN. (Sapci *et al.*, 2021) se sont concentrés sur l'apprentissage actif pour la REN. Pour ce faire, ils ont proposé un modèle qui utilise une stratégie de clustering semi-supervisée des plongements des mots de BERT, afin d'identifier les tokens positifs. Les auteurs ont également proposé une approche de normalisation basée sur les données pour pénaliser les phrases trop longues ou trop courtes. Les expériences ont été faites sur trois ensembles de données de domaines différents et ont révélé que les approches proposées réduisent le nombre de jetons annotés

tout en obtenant des performances de prédiction meilleures ou comparables aux méthodes conventionnelles.

(Wang *et al.*, 2021) propose *Automated Concatenation of Embeddings* (ACE) pour automatiser le processus de recherche de meilleures concaténations des plongements des mots pour les tâches de prédiction structurées. Pour ce faire, un contrôleur échantillonne alternativement une concaténation d'incorporations, selon sa conviction actuelle de l'efficacité des types d'incorporation individuels dans la considération pour une tâche, et met à jour la croyance basée sur une récompense. Les auteurs ont utilisé des stratégies d'apprentissage par renforcement pour optimiser les paramètres du contrôleur et calculer la récompense en fonction de la précision d'un modèle de tâche, qui est alimenté avec la concaténation échantillonnée comme entrée et formé sur un ensemble de données de tâche. Les résultats empiriques de (Wang *et al.*, 2021), testé sur 6 tâches et 21 ensembles de données, montrent que l'approche proposée surpasse les bases de référence solides et atteint des performances inégalées avec des plongements des mots *fine-tunés* dans toutes les évaluations. Au lieu de considérer uniquement les représentations au niveau du mot, (Tran *et al.*, 2017; Kuru *et al.*, 2016) incorporent les représentations des mots basées sur les caractères apprises d'un modèle neuronal. Ils estiment que, en plus de gérer les mots hors vocabulaire, les plongements des mots au niveau de caractère sont un moyen utile pour l'exploitation des sous-mots tels que les préfixes et les suffixes.

3.4 Reconnaissance d'entités nommées pour les langues peu dotées

Le transfert de modèle de TALN tels que la reconnaissance d'entités nommées à partir des langues riches en ressources qui n'admettent pas de données annotées est une réalisation attrayante. Par ailleurs, le succès des méthodes monolingues

(**?**Collobert *et al.*, 2011; Huang *et al.*, 2015) revient aux grandes quantités de données annotées disponibles. Certaines méthodes se basent sur des ensembles de règles d'identification pour chaque type d'entité en se basant sur le contexte et les caractéristiques morphologiques (Fong *et al.*, 2011).

En général, les méthodes qui utilisent les corpus parallèles afin de projeter l'annotation entre les langues font recours aux outils d'alignement. (Hatami *et al.*, 2021) utilisent l'outil Fast_align pour extraire les correspondances des mots. Ensuite, ils appliquent deux heuristiques pour obtenir des alignements dans les deux directions pour les données parallèles anglais-portugais brésilien, cette dernière étant une langue à faibles ressources. (Ehrmann *et al.*, 2011) proposent une méthode de projection de REN en utilisant des corpus multi-parallèles. Leur but était d'annoter des corpus en plusieurs langues (français, espagnol, allemand). Dans leur travail, ils utilisent le modèle IBM pour extraire les alignements mot-à-mot et donc aligner les entités qui représentent un groupe de mots. (Stengel-Eskin *et al.*, 2019) introduisent un modèle d'alignement basé sur une architecture encodeur-décodeur, qu'ils intègrent dans un modèle de traduction automatique basé sur les transformeurs. Ils évaluent les performances de leur système sur la projection de donnée REN de l'anglais vers le chinois et surpassent le modèle basé sur Fast_align en termes de F-mesure.

D'autres méthodes font recours à la traduction automatique afin de projeter l'annotation entre les langues. (Tiedemann *et al.*, 2014) visent à écarter l'annotation bruyante quant à la langue source d'un corpus parallèle. Pour cela, ils s'appuient sur l'annotation manuelle à travers la banque d'arbre UD (Universal Dependencies) combinée à la traduction automatique. Cette combinaison a permis d'entraîner un analyseur entièrement lexicalisé. D'autre part, (Mayhew *et al.*, 2017) effectuent la traduction mot-à-mot ou phrase-à-phrase en utilisant

des lexiques pour traduire les données annotées disponibles en langues riches en ressources.(Jain *et al.*, 2019) proposent un système qui s'améliore à travers trois méthodes de projection d'entités : (a) exploiter les systèmes de traduction automatique à deux reprises : d'abord, la traduction des phrases. Ensuite, la traduction des entités ; (b) effectuer les correspondances des entités en se basant sur la similarité orthographique et phonétique ; et (c) identifier les correspondances en se basant sur les statistiques distributionnelles tirées de l'ensemble des données parallèles. Leur approche réalise des améliorations sur la tâche de détection des entités nommées translingue et obtient des scores F1 de pointe pour la langue arménienne. Aussi, (Azmat *et al.*, 2020) introduisent une méthode de transfert d'annotation (EN) basée également sur la TAN. Leur approche consiste à pré-entraîner un système de TAN, à partir d'un corpus parallèle ouïghour-chinois. Ensuite, les informations des frontières qui marquent les EN sont ajoutées aux phrases de la langue source pour ré-entraîner le modèle préalablement entraîné afin que ce dernier puisse apprendre à aligner les EN. Les résultats montrent que leur système obtient une amélioration considérable par rapport au modèle de base en termes de F-mesure.

Dans la traduction automatique, les entités nommées provoquent souvent des échecs de traduction. Les modèles de représentation contextuelle pré-entraînés tels que (Peters *et al.*, 2018; Devlin *et al.*, 2019) ont réalisé de grandes avancées sur de nombreuses tâches de TALN. Or, il est possible d'apprendre le mappage lexical à travers les plongements des mots bilingues, et ce, en utilisant des dictionnaires bilingues dans le but de projeter deux espaces de plongements des mots monolingues dans un seul espace cohérent (Mikolov *et al.*, 2013b; Artetxe *et al.*, 2016), ou d'une manière non-supervisée en utilisant l'entraînement contradictoire ou les chaînes de caractères identiques (Artetxe *et al.*, 2017; Conneau *et al.*, 2017).

Dans les années passées, de nombreuses approches ont bénéficié des espaces de plongements partagés pour de nombreuses applications trans-lingues, notamment le transfert des parties du discours (en anglais *POS Tagging*) (Zhang *et al.*, 2016; Fang et Cohn, 2017) et le transfert des EN. (Bharadwaj *et al.*, 2016) introduisent un modèle neuronal basé sur le mécanisme d'attention qui utilise la représentation des caractères pour le langage phonologique universel. Ils démontrent que cette dernière facilite le transfert translingue des EN de l'anglais vers l'ewondo, langue peu dotée du Cameroun. (Ni *et al.*, 2017) proposent deux approches faiblement supervisées pour les *REN* multilingues sans annotation humaine dans une langue cible. La première approche consiste à créer automatiquement des données REN étiquetées pour une langue cible à travers la projection sur les données parallèles en développant un schéma heuristique. La deuxième méthode consiste à projeter les représentations des mots (*Word Embeddings*) d'une langue source à une langue cible pour que le système conçu pour la langue source puisse être appliqué à la langue cible sans ré-entraînement. (Xie *et al.*, 2018) proposent un système qui, en premier lieu, entraîne les plongements des mots monolingues, projette les deux espaces de plongements des mots des deux langues dans le même espace, traduit chaque mot dans la langue source en trouvant le voisin le plus proche, utilise les traductions pour traduire les entités nommées. À travers cette méthode, ils obtiennent des résultats de pointe pour la langue ouïghour. De leur côté (Adelani *et al.*, 2020) considèrent que l'incorporation des plongements des mots représente un élément clé pour la REN. En premier lieu, ils utilisent une méthode basée sur les règles pour identifier les EN en plus des listes d'entités obtenues à partir des dictionnaires. Enfin, ils utilisent une technique d'élimination de bruit basée sur la méthode de (Hedderich et Klakow, 2018) afin de nettoyer les corpus annotés automatiquement pas la méthode basée sur les règles. Les performances obtenues montrent que leur méthode est réussie pour deux langues autochtones de l'Afrique, le haoussa et le yorouba.

Parmi les méthodes qui traitent les langues à faibles ressources en utilisant les modèles basés sur l'apprentissage machine, (Yohannes et Amagasa, 2022) introduisent TigRoBERTa qui a été entrainé sur des corpus en Tigrinya, une langue sémitique éthiopienne. Ensuite ils effectuent un *fine-tuning* sur des tâches en aval telles que la REN. (Feng *et al.*, 2018) pour leur part étudient les connaissances translingues afin d'enrichir les représentations sémantiques des langues pauvres en ressources (espagnol, néerlandais). Ils développent d'abord une architecture basée sur les réseaux de neurones pour améliorer les représentations des mots à travers le transfert des connaissances à partir d'une langue riche en ressources à l'aide des dictionnaires bilingues. Ensuite, ils considèrent les caractéristiques de distributions des entités au niveau du mot comme une connaissance indépendante de la langue externe et les incorporent dans l'architecture neuronale. Les expérimentations montrent que l'intégration des représentations sémantiques a mené une amélioration considérable.

Quant à (Schweter et Baiter, 2019), ils utilisent l'apprentissage par transfert de l'allemand vers l'allemand historique. Dans leur travail, ils montrent que l'utilisation des Bi-LSTM avec une couche supérieure CRF surpasse les modèles basés exclusivement sur les CRF.

3.5 Conclusion

Dans ce chapitre, nous avons présenté les différents travaux présents dans la littérature liés à notre problématique qui consiste à détecter les EN dans les langues peu dotées. Dans le chapitre suivant, nous décrivons de notre méthodologie, dans laquelle nous proposons deux approches pour étendre le corpus annoté en anglais vers l'inuktitut, la première étant basée sur les règles et la deuxième sur les plongements des mots bilingues.

CHAPITRE IV

MÉTHODOLOGIE

L'apprentissage automatique est beaucoup utilisé dans la tâche REN. Les modèles de REN basés sur l'apprentissage profond parviennent à atteindre de bonnes performances, ceci est dû à sa facilité d'adaptation à différentes langues riches en ressources. Cependant, cette solution ne s'adapte pas aux traitements des langues peu dotées, car les algorithmes supervisés utilisés dans cette approche sont basés sur des données annotées nécessaires à l'apprentissage. De ce fait, une solution prometteuse pour les langues peu dotées, sans ressources annotées, est la reconnaissance d'entités nommées à partir des langages riches en ressources à l'aide de modèles de transfert non supervisés. Toutefois, le principal défi de cette méthode est l'association des éléments lexicaux entre les langues. En effet, cela est dû aux différences des mots et l'ordre des mots à travers les langues. Pour faire face à ce problème, des méthodes utilisent les corpus parallèles pour transférer l'annotation à travers l'alignement des mots (Ehrmann *et al.*, 2011; Wang et Manning, 2013; Ni *et al.*, 2017). D'autre part, l'association lexical est fait à travers les plongements des mots bilingues en utilisant des dictionnaires de mots bilingues (Mikolov *et al.*, 2013b; Xie *et al.*, 2018) .

Dans ce mémoire, notre objectif est de construire un modèle capable de détecter les entités nommées dans la langue inuktitut. Étant donné la rareté de données la-

bellisées pour la langue inuktitut et la disponibilité de ces dernières dans l'anglais, l'idée principale de notre approche est de transférer les caractéristiques linguistiques de l'anglais vers l'inuktitut.Le défi majeur de cette méthode est la différence morphologique, entre les deux langues qui rend la projection difficile. La figure 4.1 présente le pipeline qui décrit l'entrée, le modèle et la sortie. De plus, la langue inuktitut, étant une langue autochtone, elle est liée à un savoir et à une culture autochtone et doit être traitée en tenant compte de ces spécificités culturelles.

FIGURE 4.1 – Description du pipeline proposé.

Dans ce chapitre, nous présenterons nos deux modèles proposés. Nous commençons par la première approche. Cette dernière consiste à transférer l'annotation de l'anglais vers l'inuktitut en combinant des règles en utilisant un analyseur morphologique avec l'alignement des mots. Ensuite, nous proposons une deuxième approche basée sur les plongements des mots bilingues en utilisant un dictionnaire bilingue (anglais-inuktitut) que nous avons construit.

4.1 Approche basée sur les règles

Parmi les méthodes de détection des entités nommées pour les langues peu dotées, l'utilisation des outils d'alignement des mots pour des corpus parallèles dont le corpus concernant la langue riche en ressources a été annoté. Dans cette approche, nous accompagnons l'alignement des mots d'une base de règles morphologiques en

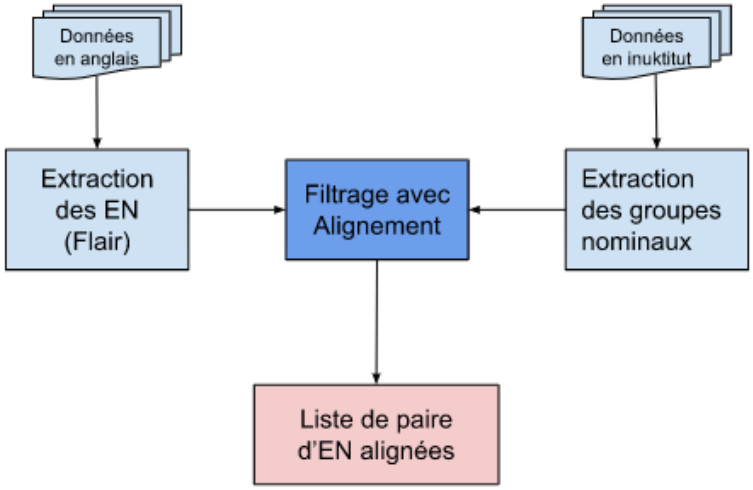

FIGURE 4.2 – Méthode basée sur les règles.

utilisant l'outil UQAILAUT [1]. La méthodologie suivie dans cette approche consiste à :

- Extraire les entités nommées du corpus anglais en utilisant le modèle Flair Embeddings.
- Effectuer une analyse morphologique des phrases inuktitut.
- Identifier les groupes nominaux du texte inuktitut en utilisant l'outil UQAILAUT.
- Filtrer les groupes nominaux qui ne représentent pas des EN, en utilisant l'alignement de mots.
- Construire un dictionnaire d'EN bilingues (anglais-inuktitut).
- Construire une base de connaissance dans la langue autochtone (inuktitut), ce qui facilite la réalisation des tâches de TALN en aval et à la préservation de la culture autochtone.

La figure 4.2 illustre le pipeline de notre méthode basée sur les règles.

1. `https://www.inuktitutcomputing.ca/index.php` (consulté en janvier 2023)

Afin de détecter les entités nommées dans une langue peu dotée, ces dernières doivent être d'abord identifiées dans une langue riche en ressources (Ehrmann *et al.*, 2011).

4.1.1 Détection des entités nommées (Flair Embeddings)

Dans notre projet, nous utilisons le modèle Flair Embeddings, qui est un modèle de langue basé sur les caractères entraînés pour produire un type de plongements des mots appelés *plongements des chaînes contextuelles*. Ces plongements des propriétés consistent en (a) l'entraînement sans notion explicite de mots. Le texte est modélisé fondamentalement comme séquence de caractères, et (b) la contextualisation des mots par leur texte environnant (Akbik *et al.*, 2018). Autrement dit, un mot peut avoir plusieurs plongements selon son utilisation contextuelle. Ces propriétés sont intégrées dans une architecture d'étiquetage de séquences. Le concept global consiste à faire passer une séquence de texte au modèle de langage de caractères et récupérer les plongements contextuels en sortie de façon à ce que le modèle d'étiquetage de séquence puisse classer les entités comme le montre la figure 4.3. De plus, le modèle Flair Embeddings utilise des empilements de plongements des mots, et ce, en combinant plusieurs plongements (Akbik *et al.*, 2018). Aussi, il combine les plongements des mots contextuels avec les plongements Glove pour représenter un mot pour l'étiquetage de séquence (Pennington *et al.*, 2014). Enfin, l'empilement des plongements des mots représentant les phrases est transmis à une architecture LSTM-CRF pour classer les entités nommées.

Le modèle Flair Embeddings atteint des performances de pointe et surpasse les travaux antérieurs sur la reconnaissance d'entités nommées (93 de F_1_score sur les données CoNLL 2003 (Tjong Kim Sang et De Meulder, 2003) en anglais et en allemand).

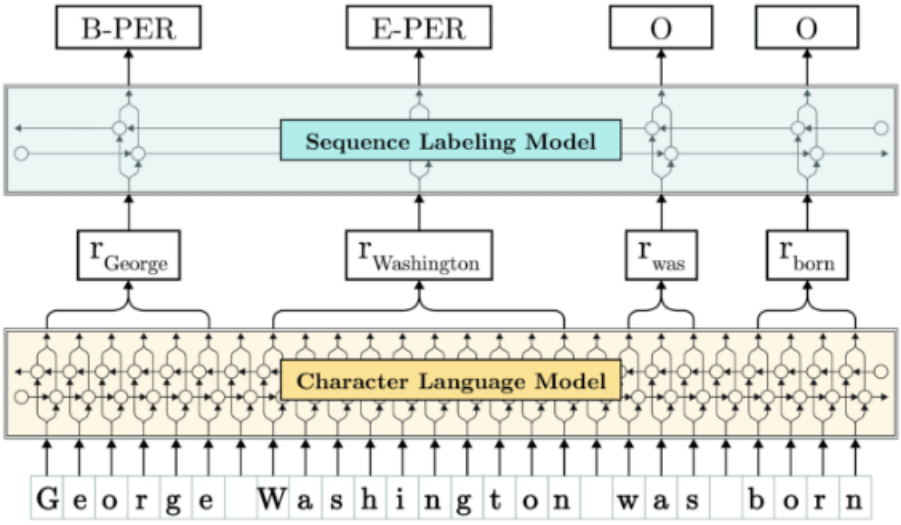

FIGURE 4.3 – Flair Embeddings. (Source : (Akbik *et al.*, 2018))

Une fois que les entités sont identifiées dans la langue source (riche en ressources), elles doivent être projetées vers la langue cible (peu dotée). en appliquant des méthodes et en utilisant des outils. Dans notre cas, la méthode basée sur les règles et la méthode basée sur les plongements des mots.

4.1.2 Alignement des mots

Dans le traitement du langage naturel, il existe de nombreuses applications pour l'alignement des mots. Les performances de la majorité de ces applications dépendent fortement de la qualité de l'alignement des mots (Och et Ney, 2000a; Yarowsky et Wicentowski, 2000). L'extraction automatique des dictionnaires bilingues à partir des corpus parallèles est un système d'application fréquemment utilisé pour l'alignement des mots (Tian *et al.*, 2011). Dans le but de garantir une bonne qualité de projection des entités nommées de l'anglais vers l'inuktitut, nous utilisons et comparons, dans cette approche, différents outils d'alignement

des mots.

4.1.2.1 Modèles d'alignement des mots utilisés

Dans cette section, nous faisons un survol sur les outils d'alignement des mots utilisés dans notre projet.

GIZA++

GIZA++ est un outil d'alignement qui implémente les modèles les plus connus accompagné par l'amélioration. GIZA++ (Och et Ney, 2000b) est un outil basé sur l'entraînement des modèles génératifs tels que IBM 1 à 5 (Brown *et al.*, 1993) et les modèles basés sur le modèle de Markov caché (Vogel *et al.*, 1996a). De plus, les paramètres dans ce modèle sont estimés en utilisant l'algorithme d'espérance-maximisation (EM). Ainsi, les résultats d'alignement des mots sont obtenus après l'entraînement des corpus parallèles sur plusieurs itérations en deux directions (langue source vers langue cible et langue cible vers langue source).

Fast-Align

Le modèle Fast-Align est une variante de la traduction lexicale du modèle proposé par (Brown *et al.*, 1993). Étant donné deux phrases en langue source et en langue cible, le modèle génère un alignement qui indique la traduction de chaque mot dans la phrase source qui se trouve dans la phrase cible. La configuration du modèle d'alignement consiste en la paramétrisation log-linéaire du modèle IBM 2. La formulation est écrite sous la forme $\delta(a_i = j, i, m, n)$ comme illustré dans

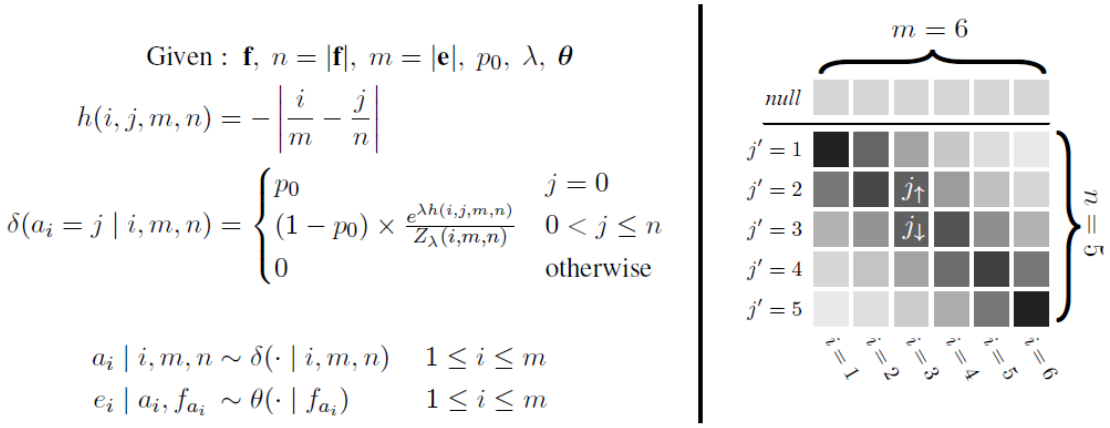

FIGURE 4.4 – Le modèle Fast-align. (Source : (Dyer *et al.*, 2013))

la figure 4.4. Cette dernière montre une traduction e et son alignement a sur une phrase source f, le paramètre de d'alignement p_0 et la précision λ qui contrôle le degré de la mesure des points d'alignement près de la diagonale. Le côté droit de la figure 4.4 montre une illustration graphique de la distribution d'alignement dans laquelle les carrés plus sombres indiquent une probabilité plus élevée.

Eflomal

Eflomal (en anglais *Efficient Low-Memory Aligner*) est un outil basé sur Efmaral qui est basé sur l'échantiollonage de Gibbs, une autre méthode de Monte-Carlo par chaînes de Markov, ou MCMC (en anglais *Markov chain Monte Carlo*) (Östling et Tiedemann, 2016). Cette classe d'algorithmes fait des compromis entre la précision et l'efficacité de calcul. De plus, il a été montré que les antérieurs hiérarchiques améliorent la précision mais restent coûteux (Östling, 2015). Pour cette raison, Eflomal utilise un modèle HMM pour l'ordre des mots ainsi qu'un modèle de fertilité.

4.1.3 Détection des groupes nominaux.

Dans notre projet, nous avons utilisé l'outil d'analyse morphologique UQAILAUT pour identifier les groupes nominaux dans la langue inuktitut (The UQAILAUT Project, 2012).

4.1.3.1 UQAILAUT

UQAILAUT est un analyseur morphologique inuktitut réalisé dans le cadre d'un projet de recherche par l'institut de technologie de l'information du CNRC (Conseil national de recherches du Canada) (The UQAILAUT Project, 2012). Dans notre approche, nous avons utilisé l'analyseur morphologique inuktitut proposé dans UQAILAUT. Cet analyseur consiste à décomposer grammaticalement un mot inuktitut en racine, suffixes et terminaison. L'information lexicale contenue dans la base de données de cet analyseur morphologique comporte :

- 2000 racines ;
- plusieurs centaines de mots lexicalisés ;
- plus de 330 suffixes ;
- 300 terminaisons nominales ;
- et 1200 terminaisons verbales.

La figure 4.5 représente un exemple d'analyse morphologique du mot "*ilinniaqtulirinirmut*". L'outil décompose le mot en ses éléments constitutifs (morphèmes) et retourne plusieurs décompositions possibles. L'analyse correcte se trouve généralement parmi les trois premières lignes. Dans notre projet, nous considérons la première ligne de décomposition.

Format de décomposition : La décomposition se fait sous forme d'une séquence

FIGURE 4.5 – Exemple d'analyse morphologique avec l'outil UQAILAUT. (The UQAILAUT Project, 2012)

d'une ou de plusieurs paires $x : y$, où x représente la forme de surface du morphème dans le mot et y est l'identifiant unique du morphème dans la base de données linguistique. La figure 4.7 montre un exemple de décomposition à l'aide de l'analyseur morphologique inuktitut UQAILAUT.

Format de l'identifiant unique :

L'identifiant unique est formulé sous la forme *<unique id> : <name>/<nb><type>*, où *<nb>* est un entier et les types sont classifiés dans le tableau 4.1.

4.1.4 Projection

La projection des entités nommées de l'anglais vers l'inuktitut à l'aide des corpus parallèles s'effectue en plusieurs étapes.

1. Tout d'abord, nous effectuons la reconnaissance d'EN dans le corpus anglais en utilisant le modèle Flair Embeddings.
2. Ensuite, nous identifions les noms dans la langue inuktitut dépendamment

TABLEAU 4.1 – Les types de morphèmes (The UQAILAUT Project, 2012).

v	racine verbale
n	racine du nom
a	adverbe
c	conjonction
q	suffixe de terminaison
nn	suffixe nom à nom
nv	suffixe nom-verbe
vn	suffixe verbe-nom
vv	suffixe verbe à verbe
tv-<mode>-<subject's person & number>[-<object's person & number>-[prespas|fut]]	terminaison verbale
tn-<case>-<number>[-<possessor's person & number>]	terminaison nominale
tad-<case>	terminaison d'adverbe
tpd-<case>-<number>	terminaison de pronom démonstratif
pd-<location>-<number>	pronom démonstratif
ad-<location>	adverbe démonstratif
rad	racine de l'adverbe démonstratif
rpd	racine du pronom démonstratif

TABLEAU 4.2 – Description des composantes des morphèmes (The UQAILAUT Project, 2012).

<mode>	dec	déclaratif
	ger	gerundive
	int	interrogatif
	imp	impératif
	caus	causatif
	cond	conditionnel
	freq	fréquentatif
	dub	dubitatif
	part	participatif
<case>	nom	nominatif
	acc	accusatif
	gen	génitif
	dat	datif
	abl	ablatif
	loc	locatif
	sim	similaire
<number>	s	singulier
	d	double
	p	pluriel
<location>	sc	statique ou court
	ml	dynamique or long

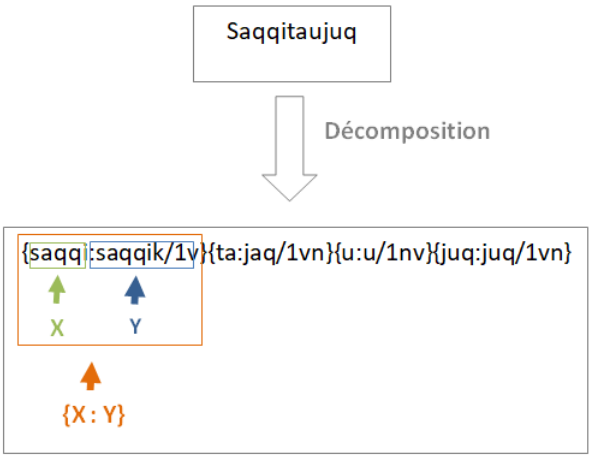

FIGURE 4.6 – Exemple de format de décomposition avec l'analyseur morphologique.

de l'analyse morphologique. Exemple : l'analyse morphologique du mot *Titiraqsimaningit* qui signifie en anglais *Premier* est :

{titiraq :titiraq/1v}{sima :sima/1vv}{ni :niq/2vn}{ngit :ngit/**tn**-nom-p-4s}. La terminaison du mot est *tn*, ce qui signifie que c'est une terminaison d'un nom.

3. Après, nous identifions les groupes nominaux en sélectionnant dans la phrase les segments qui contiennent des successions de noms.
4. Enfin, nous projetons l'entité de l'anglais vers l'inuktitut en utilisant les appariements d'indices obtenus par l'alignement des mots. Les indices des EN et les indices des groupes nominaux sont regroupés pour obtenir des appariements d'EN en sortie. Exemple :

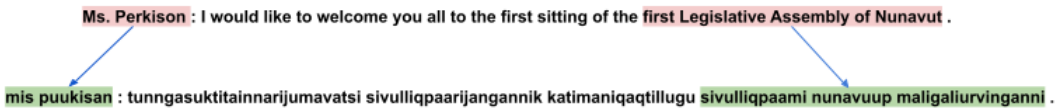

FIGURE 4.7 – Exemple de projection d'entité.

4.2 Approche basée sur les plongements des mots bilingues

Les entités nommées translingues sont obtenues par le transfert de connaissances à partir d'une langue riche en ressources admettant de nombreuses étiquettes d'entités nommées vers une langue peu dotée (Ehrmann *et al.*, 2011). Dans cette approche, nous adoptons la méthode de transfert nonsupervisé basée sur l'espace de plongements des mots bilingues. Cette approche s'opère en deux étapes : résoudre le problème d'ordre des mots entre les langues et effectuer efficacement le mappage lexical entre les deux langues.

Cette approche comporte :

- La construction d'un dictionnaire bilingue.
- La détection des entités nommées en utilisant le modèle Flair embeddings.
- L'entraînement des plongements des mots Monolingues sur chaque corpus (anglais, inuktitut) sur des données inuktitut segmentées.
- La projection translingue en effectuant un mappage linéaire entre les deux espaces monolingues dans le même espace et en utilisant un dictionnaire bilingue.
- Pour chaque entité dans la langue source à partir des représentations apprises :
  - Calculer la distance entre les vecteurs des EN bilingues en utilisant la métrique CSLS.

— Sélectionner le voisin le plus proche comme l'entité traduite.

La figure 4.8 montre le pipeline de notre méthode basée sur les plongements des mots bilingues.

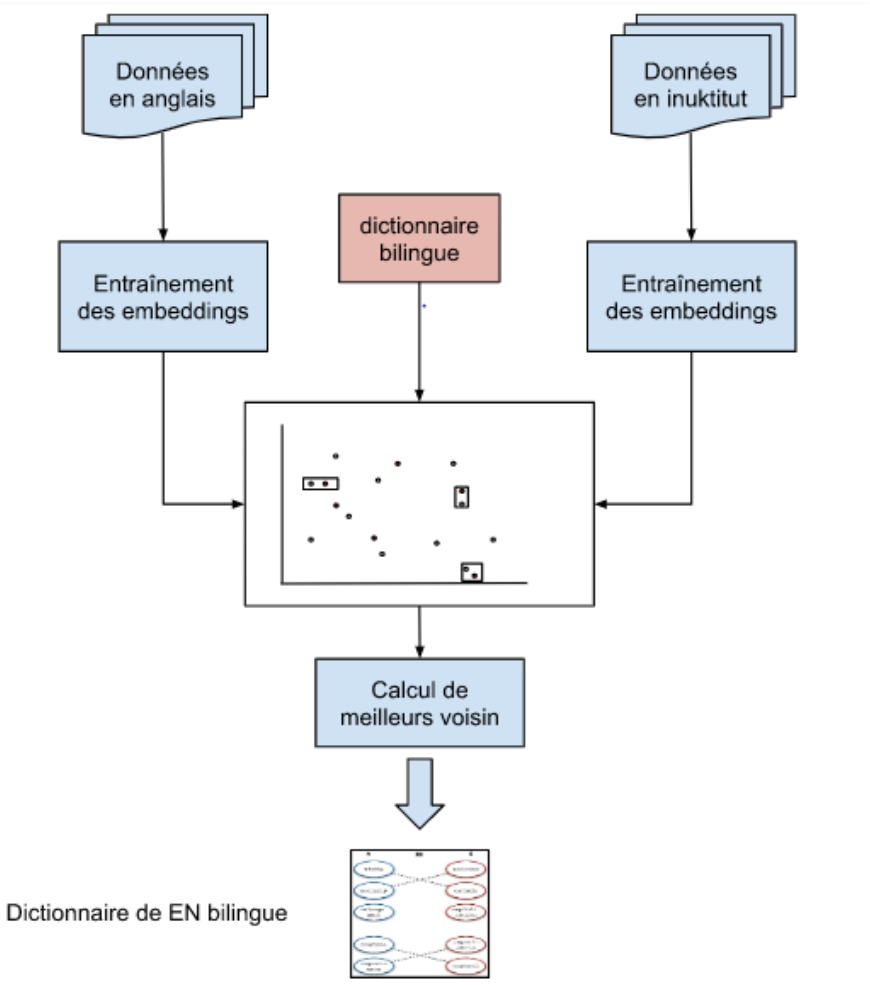

FIGURE 4.8 – Méthode basée sur les plongements des mots bilingues.

4.2.1 Entraînement des plongements des mots monolingues

Étant donné que dans notre travail nous utilisons des corpus parallèles, nous entraînons en premier lieu les plongements des mots sur la langue source (anglais)

ainsi que sur la langue cible (inuktitut) indépendamment.

4.2.1.1 Prétraitement

En général, les données textuelles ne sont pas totalement nettoyées, car ses données proviennent des sources qui ont des caractéristiques différentes. En ce qui concerne le prétraitement du texte, il constitue l'une des étapes les plus importantes du pipeline d'entraînement des plongements des mots. De plus, nos résultats de traduction dépendent fortement de la qualité des plongements des mots que nous entraînons séparément. Afin d'assurer une bonne précision quant à la traduction, nous effectuons les étapes prétraitement suivantes.

Suppression des contractions : Les contractions sont des mots qui s'écrivent avec des apostrophes. Exemple : le mot *can't* devient *cannot*. L'extension de ces mots est nécessaire vu que nous voulons normaliser le texte.

Suppression de ponctuation et nettoyage des caractères spéciaux : Puisque les caractères spéciaux et la ponctuation n'admettent pas de représentations vectorielles, il est nécessaire de les éliminer pour les deux langues.

Segmentation des mots : En raison de la polysémie de la langue inuktitut, l'alignement des mots sources avec les mots cibles dans l'espace commun serait erroné, ce qui affectera les performances de la traduction automatique. Pour remédier à ce problème, nous utilisons l'analyseur morphologique (UQAILAUT) pour séparer les mots en plusieurs morphèmes, ce qui facilite l'appariement des mots dans l'espace des plongements des mots bilingues.

4.2.1.2 FastText

L'algorithme FastText est un modèle supervisé, similaire à l'architecture CBOW de word2vec. FastText prédit les balises à travers le contexte qui consiste au type de texte déterminé par une annotation manuelle (Ma *et al.*, 2020). L'architecture de ce modèle comporte trois couches :

— couche d'entrée ;
— couche cachée ;
— et couche de sortie.

Le modèle FastText utilise *Hierarchical Softmax* (Goodman, 2001) basé sur l'arbre de codage de Huffman (Mikolov *et al.*, 2013a). Les données d'entrée sont un nombre de mots et leurs caractéristiques n-grammes utilisées pour représenter un seul document. La couche cachée représente la moyenne de plusieurs vecteurs de caractéristiques qui construit un arbre de Huffman selon les poids et les paramètres de chaque catégorie et utilise l'arbre de Huffman comme sortie (Ma *et al.*, 2020).

4.2.2 Plongements des mots bilingues (MUSE)

L'intégration de mots multilingues non supervisée ou supervisée (en anglais *Multilingual Unsupervised or Supervised Word Embedding*, MUSE) de Facebook est une bibliothèque Python libre de mots multilingues supervisés et non supervisés. Cette bibliothèque aligne les espaces des plongements des mots bilingues. L'alignement se fait d'une manière supervisée en utilisant un dictionnaire, ou d'une manière non supervisée qui établit un dictionnaire entre les deux langues en alignant les espaces des plongements des mots bilingues et n'utilise aucune donnée parallèle. Étant basé sur FastText, le modèle MUSE possède la fonction d'intégration de mots multilingues dans plus de 30 langues et dispose de 110 dictionnaires

bilingues (Ma *et al.*, 2020).

MUSE, pour sa part, utilise des vecteurs de mots non supervisés formés à l'aide de FastText (plongements monolingues de dimension 300 entraînés sur des corpus Wikipédia). De ce fait, l'association W a la taille 300×300. Le modèle d'architecture de MUSE est basé sur un perceptron multicouche avec deux couches cachées de taille 2048. et fonctions d'activation Leaky-ReLU. MUSE est motivé par la production des appariements fiables entre deux langues, et ce, en améliorant la métrique de comparaison de sorte que le voisin le plus proche d'un mot source, dans la langue cible, soit plus susceptible d'avoir ce même mot source comme le voisin le plus proche. Cette métrique de distance est appelée CSLS (*Cross-domain Cimilarity Local Ccaling*) (Conneau *et al.*, 2017).

La figure 4.9 illustre le processus de cette méthode.

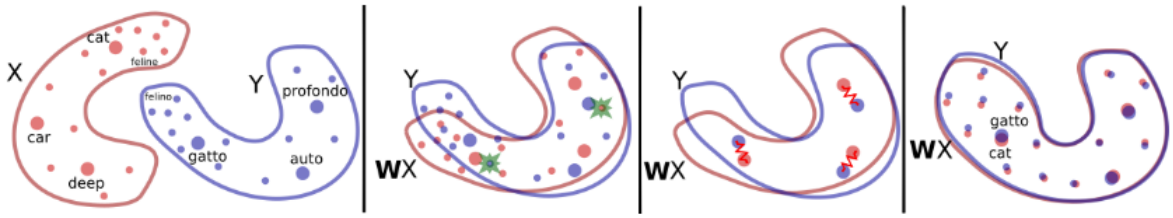

FIGURE 4.9 – Illustration de MUSE. (Source : (Conneau *et al.*, 2017))

Dans la figure 4.9, chaque point désigne un mot avec une taille proportionnelle à la fréquence du mot dans le corpus. Une matrice de rotation W qui aligne deux distributions est formée en utilisant l'apprentissage contradictoire. L'association W est affiné en utilisant des mots fréquents alignés lors de l'étape précédente comme points d'ancrage. Cette association est ensuite utilisé pour mapper tous les mots du dictionnaire. Enfin, la traduction est effectuée en utilisant l'association W et la métrique CSLS (en anglais *Cross-Domain Similarity Local Scaling*) qui

est définie par.

$$CSLS(x_i, y_j) = 2\cos(x_i, y_j) - r_T(x_i - r_S(y_j).$$

Où $r_T(x_i)$ désigne la similarité cosinus moyenne entre x_i et ses K voisins y_t. La traduction pour chaque mot source s en sélectionnant le mot cible $\hat{t}_s$, ou $\hat{t}_s = argmaxCSLS(x_i, y_j)$.

4.2.3 Transfert d'annotation

Afin de transférer l'annotation (entités nommées) de l'anglais vers l'inuktitut, nous effectuons la traduction mot-à-mot en recherchant le voisin le plus proche dans l'espace de plongements des mots communs en se basant sur la méthode proposée dans (Xie *et al.*, 2018). Tel que mentionné ci-haut, la métrique CSLS est utilisée. L'entrée est représentée par des phrases annotées en anglais. La sortie est la traduction des phrases mot-à-mot en inuktitut annoté.

4.3 Reconnaissance des entités nommées pour la traduction automatique

L'identification correcte des entités nommées (EN) représente un défi pour la tâche de traduction automatique neuronale, à savoir dans la recherche et le développement de traducteurs commerciaux. Souvent, la traduction des noms propres nécessite des approches et des méthodes différentes de celles des autres mots (Newmark, 1981). Ainsi, les systèmes de TAN montrent une faiblesse significative dans l'apprentissage de la traduction des EN et les noms communs. Par conséquent, le contexte, la structure syntaxique et lexicale de la traduction seront affectés.

Dans notre projet, nous décrivons l'effet de l'utilisation des entités nommées dans la traduction automatique neuronale. Nous évaluons trois systèmes de traduction automatique. Le premier représente le système de base qui consiste à utiliser uni-

quement les corpus parallèles. Le deuxième utilise le dictionnaire des entités nommées extraites à l'aide de notre première approche, qui est basée sur les règles en plus des corpus parallèles. Le troisième utilise le dictionnaire des entités nommées extraites à travers notre deuxième approche, qui est basée sur les plongements des mots bilingues ajoutés au corpus parallèles. Cette approche permet d'améliorer les performances de la TAN pour la langue inuktitut.

4.4 Conclusion

Dans ce chapitre, nous avons présenté nos modèles d'extraction des entités nommées dans la langue inuktitut. En prenant en entrée les données labellisées dans l'anglais et des corpus parallèles (anglais, inuktitut), nous avons proposé une première approche d'extraction des entités nommées basée sur les règles. Par la suite, nous avons présenté une deuxième méthode basée sur les plongements des mots bilingues. Dans le chapitre suivant, nous présentons les résultats d'évaluations et nous établissons une étude comparative entre les deux méthodes proposées. À notre connaissance, il n'existe pas de travaux qui explorent cette voie qui concerne les EN pour la langue inuktitut.

CHAPITRE V

ÉVALUATION ET RÉSULTATS

Dans ce chapitre, nous effectuons une étude comparative des deux approches proposées au chapitre IV : une approche basée sur les règles et une approche basée sur les plongements des mots bilingues. Pour ce faire, nous commençons par présenter les données utilisées lors de l'évaluation de nos modèles. Par la suite, nous définissons les métriques utilisées pour l'étude des performances.

L'évaluation des systèmes de TALN peut être classée en méthodes intrinsèques et extrinsèques (Clark et Lappin, 2010). Par contre, dans une évaluation intrinsèque, les qualités des sorties des systèmes sont évaluées par rapport à une référence prédéterminé (ensemble de données annotées). Dans une évaluation extrinsèque, la qualité des sorties est mesurée en fonction de leur impact sur les performances des autres systèmes. Vers la fin du chapitre, nous discutons les résultats de l'évaluation intrinsèque, c'est-à-dire pour la tâche de REN, ainsi que l'évaluation extrinsèque, c'est-à-dire l'impact des EN sur la traduction automatique.

5.1 Données d'évaluation

Pour notre projet, nous avons utilisé le corpus parallèle de *Nunavut Hansard Inuktitut–English* version 3.0, un écrit rassemblé et aligné sur les peines des déli-

Tableau 5.1 – Description des données.

	Anglais	Inuktitut
Entraînement	1 293 346 phrases	1 293 346 phrases
Validation	5 433 phrases	5 433 phrases
Test	6 139 phrases	6 139 phrases
Total	1 304 918 phrases	1 304 918 phrases

bérations de l'assemblée législative du Nunavut (Nunavut Maligaliurvia) (Joanis *et al.*, 2020). Ce corpus comprend les actes des 687 journées de débats avec 8−068 977 mots en inuktitut et 17 330 271 mots et couvertures en anglais, ce qui donne environ $1,3$ million paires de phrases alignées. Ce corpus a été utilisé dans plusieurs travaux de recherche notamment dans la tâche partagée (Knowles *et al.*, 2020).

5.1.1 Dictionnaires

À l'aide de la base de données du projet UQAILAUT, nous avons pu construire un dictionnaire bilingue de 1 560 mots. Ces derniers constituent des significations des mots racines ainsi que les significations des suffixes. Aussi, nous avons utilisé le traducteur Microsoft Bing [1] pour traduire les mots anglais les plus fréquents dans les corpus parallèles vers l'inuktitut.

1. https ://www.bing.com/translator (consulté en septembre 2022)

5.2 Métriques d'évaluation

Afin d'analyser les performances de nos systèmes et étudier leur capacité à identifier et extraire correctement les entités nommées, les prévisions de nos modèles sont comparées avec les résultats des prévision faites par nos soins. Dans la littérature, il existe plusieurs métriques d'évaluation des performances d'un système de reconnaissance d'entités nommées. Dans le cadre de notre travail, nous avons utilisé les mesures de rappel, précision, F-score et SacreBLEU (Post, 2018). Ces métriques sont calculées en se basant sur :

— **Faux positif (FP) :** représente les entités prédites par le système de reconnaissance d'entités nommées, mais qui n'apparaissent pas dans la vérité de référence (*ground truth*).
— **Faux négatif (FN) :** représente les entités non prédites par le système REN, mais qui apparaissent dans la vérité de terrain.
— **Vrai positif (VP) :** représente les entités prédites par le système REN, et qui apparaissent dans la vérité de terrain (Li *et al.*, 2022).

5.2.1 Rappel

Le rappel est le pourcentage de toutes les entités nommées correctement reconnues par notre système REN. Cette masure est obtenue par le rapport des entités correctement reconnues et la somme des entités étiquetées reconnues et non reconnues. Cette métrique nous permet de mesurer la capacité de notre modèle à reconnaître toutes les entités dans le corpus (Dekhili, 2020). Le rappel est calculé comme suit (Li *et al.*, 2022) :

$$Rappel = \frac{VP}{VP + FN}.$$

5.2.2 Précision

La précision est le rapport entre les entités correctement reconnues par le système REN et toutes les entités nommées. Ce rapport donne le pourcentage des résultats correctement reconnus, ce qui nous permet d'analyser la capacité de notre modèle à prédire que les entités correctes (Dekhili, 2020). Le rapport de précision est formulé comme suit (Li *et al.*, 2022) :

$$Prcision = \frac{VP}{VP + FP}.$$

5.2.3 F-mesure

Cette mesure est la moyenne de la précision et du rappel, elle est calculée en combinant les deux mesures et est formulée comme suit :

$$F1 - mesure = 2 \times \frac{Prcision \times Rappel}{Prcision + Rappel}.$$

Les métriques **F-mesure macro-moyenne** et **F-mesure micro-moyenne** sont utilisées pour évaluer les systèmes REN car elles tiennent compte des performances de plusieurs types d'entités.

- F-mesure macro-moyenne prend la moyenne des F-mesure calculée indépendamment sur différents types d'entités.
- F-mesure micro-moyenne donne les statistiques calculées à partir des sommes des faux négatifs, faux positifs et vrais positifs sur tous les types d'entités.

5.3 Évaluation de la méthode basée sur les règles

Dans cette section, nous définissions la méthode ainsi que les métriques d'évaluation qui concernent chaque approche proposée.

5.3.1 Évaluation d'alignement des mots

La table 5.2 représente les résultats d'alignement des mots testés à l'aide de plusieurs outils d'alignement. Nous comparons aussi nos résultats à ceux de la Shared Task (Koehn *et al.*, 2005) obtenus par (Langlais *et al.*, 2005) à savoir NUKTI et JAPA dans la table 5.3. Les outils d'alignement ont été entraînés sur les corpus parallèles de Nunavut Hansard Inuktitut–English et ont été évalués sur un ensemble de données alignées de référence utilisé dans la tâche partagée.

Les performances obtenues avec l'outil Eflomal montrent une amélioration considérable quant au taux d'erreur d'alignement par rapport aux autres outils. Ceci est expliqué par la méthode d'échantillonnage par itération qu'utilise ce modèle.

TABLEAU 5.2 – Les performances des outils d'alignement utilisés.

	AER	**Précision**	**Rappel**	**F-mesure**
Fast Align	0,643	0,25	0,623	0,25
GIZA	0,669	0,32	0,33	0,33
Eflomal	**0,474**	**0,367**	**0.930**	**0,367**
Eflomal (IBM+HMM)	0,499	0,351	0,874	0,351
Eflomal (IBM1)	0,596	0,281	0,721	0,281

Les résultats d'alignement enregistrent un taux d'erreur d'alignement plus élevé comparé aux outils d'alignement de la shared Task. Nos résultats (TEA) se rapprochent des résultats obtenus par l'outil NUKTI combiné au modèle JAPA mais restent tout de même moins performants que l'outil NUKTI admettant un TEA égal à 30.6.

TABLEAU 5.3 – Comparaison des performances d'alignement.

	AER	**Précision**	**Rappel**	**F1_mesure**
Eflomal	**0,474**	0,367	0.930	0,367
NUKTI	**30,6**	0,6309	0,6587	0,6445
NUKTI + JAPA	0,4646	0,5134	0,5360	0,5244
JAPA	0,7127	0,2617	0,7449	0,3873

5.3.2 Évaluation de la projection des EN (méthode intrinsèque)

Afin d'évaluer les performances de projection des EN, nous avons construit un corpus d'EN en inuktitut. L'annotation de ce corpus a été effectuée manuellement sur un ensemble de phrases sur lequel on a effectué la reconnaissance des entités nommées on prenant comme référence les résultats du traducteur Bing. La table 5.4 illustre les statistiques des entités nommées de cet ensemble de données d'évaluation ainsi que les 4 types d'entités nommées : LOC (*location*), ORG (*organization*), PER (*person*) et MISC (autre) qui n'appartiennent à aucun type d'EN.

Le tableau 5.5 illustre les résultats d'évaluation de notre approche basée sur les règles. Le tableau 5.5 résume les performances en termes de rappel, précision et F-mesure du modèle basé sur les plongements des mots. Les résultats de cette approche s'avèrent intéressants, en particulier pour l'entité "PER" proportion-

TABLEAU 5.4 – Statistiques d'EN.

Entité	**Nombre**
LOC	45
PER	111
ORG	38
MISC	11
Total	105

nellement au nombre total d'entités. en raison de la morphologie de la langue inuktitut qui est très différente de celle de l'anglais, l'outil d'alignement des mots pourrait être induit en erreur. Contrairement aux langues admettant la même typologie morphologique, le taux d'erreur d'alignement est beaucoup moins élevé. De plus, les parties du texte qui représentent une entité PER constituées de n mots admettent généralement une traduction de n mots (traduction mot-à-mot). Exemple : la traduction de l'entité *PER* "Glenn McLean" est "gilin maklain". Par contre, la traduction de l'entité *LOC* "Whale Cove" est "tikirarjuaq".

TABLEAU 5.5 – Résultats d'évaluation d'EN (approche basée sur les règles).

	Précision	**Rappel**	**F1_mesure**
PER	0.84	0.73	0.78
MISC	1	0.20	0.33
LOC	0.95	0.59	0.73
ORG	0.81	0.54	0.65

5.4 Évaluation de la méthode basée sur les plongements des mots bilingues

Afin d'évaluer les performances de traduction dans l'espace de plongements des mots commun, nous avons construit un dictionnaire d'évaluation bilingue constitué de 30 paires de mots. Comme montré dans le tableau 5.6. L'évaluation s'est faite en calculant la précision de la traduction des mots dans le voisinage de $k = 1, 5, 10$. La traduction se fait en calculant la similarité entre le mot à traduire et les mots-voisins.

TABLEAU 5.6 – Résultats de la traduction des mots.

K	**Précision**
1	36,67
5	40,00
10	43,33

Le tableau 5.6 illustre les résultats de la traduction mot-à-mot par la méthode des plongements des mots bilingues. Nous remarquons que les performances en termes de précision quant au voisinage de $k = 10$ sont plus élevées. Ceci est expliqué par le fait que la probabilité de tomber sur la bonne traduction du mot est élevée lorsque le nombre de voisins est grand.

5.5 Évaluation par la traduction automatique

5.5.1 Architecture et configuration

Concernant la tâche de traduction automatique neuronale et afin d'évaluer nos trois modèles, nous avons utilisé l'outil Fairseq (Ott *et al.*, 2019) pour entraîner le

modèle basé sur les transformeurs avec les paramètres mentionnés dans la table 5.7. Pour le prétraitement, nous avons utilisé l'outil de *tokenisation* de Moses (Koehn *et al.*, 2007). En outre, dans la segmentation en sous-mots BPE (en anglais *Byte-Pair Encoding*), nous avons utilisé l'outil Subword-NMT (Sennrich *et al.*, 2015) pour créer un vocabulaire de 20k. Enfin, Nous évaluons nos trois modèles avec la métrique SacreBLEU définie comme suit.

5.5.2 SacreBLEU

BLEU (*bilingual evaluation understudy*) est une métrique d'évaluation de traduction automatique. Elle mesure la qualité d'un texte qui a été traduit automatiquement d'une langue source à une autre langue. De son côté, SacreBLEU (Post, 2018) fournit un calcul facile de la mesure BLEU partageable, comparable et reproductible.

Tableau 5.7 – Configuration du système de TAN.

Nom du paramètre	**Choix**
Architecture	transformeur
Optimiseur	Adam
Taux d'apprentissage	0.00005
Dropout	0.3
Batch-size	32
Epoch	40

Le tableau 5.8 illustre l'évaluation de nos deux approches et leur comparaison à notre baseline.

Le tableau 5.8 montre les performances des trois modèles de TAN, à savoir le modèle de base, l'approche basée sur l'alignement des mots (Modèle 1) et l'approche

Modèle	SacreBLEU	Temps d'exécution (m)
Référence	31.31	116
Modèle 1	**32.84**	132
Modèle 2	31.70	144

TABLEAU 5.8 – Résultats obtenus avec la TAN.

basée sur les plongements des mots bilingues (Modèle 2). Le modèle 1 obtient des performances plus élevées que celles du modèle de base et le modèle 2 en termes de SacreBLEU. La raison est que le modèle 1 réussit à aligner les entités dans le corpus parallèle malgré le taux d'erreur d'alignement. Contrairement au modèle 2 qui effectue la traduction des EN mot par mot dans l'espace de plongements des mots bilingues en sélectionnant le voisin le plus proche. Ce qui fausse parfois la traduction des EN, particulièrement les mots inuktituts représentant des phrases.

5.6 Analyse des erreurs

Concernant la méthode basée sur l'alignement des mots, Les performances sont plus élevées pour les entités "PER" et "LOC". Cela est expliqué par la morphologie de la langue. Or, les noms propres sont généralement traduits mot-à-mot, tandis que les autres entités telles que "ORG" et "MISC" représentent des phrases dont la traduction en inuktitut est juste un seul mot. Exemple : la traduction de l'entité *PER* "Hunter Tootoo" est "Hanta tutu", la traduction de l'entité *ORG* "Legislative Assembly" est "Maligaliurvik". La différence morphologique entre les deux langues a causé des erreurs d'alignement des mots, ce qui a entraîné la projection erronée des entités nommées.

Les résultats d'évaluation des trois modèles énumérés dans le tableau 5.8 montrent

que le modèle 1 qui est basé sur l'alignement des mots est le plus performant, vient en deuxième position le modèle 2 qui est basé sur les plongements des mots bilingues. La raison est que le modèle 1, hormis les erreurs d'alignements, il est tout de même capable d'aligner les entités nommées dans les deux langues. Par contre, le modèle 2 effectue les traductions des entités mot-à-mot. Or, comme expliqué précédemment, la langue inuktitut, étant une langue polysynthétique, une phrase peut être représentée par un seul mot. Les erreurs de projections sont illustrées avec les exemples suivants :

— Les erreurs de projection dues aux erreurs d'alignement

  Exemple :

  (iu) Uqausiksait **jain sutuuatmut**, maligaliuqti , inulirijituqakkunnut ministarijaujuq.

  (en) Presentation by the Hon. **Jane Stewart**, MP, Minister of Indian Affairs and Northern Development.

  L'entité *PER* "**Jane Stewart**" est alignée avec "**sutuuatmut ,**" au lieu de "**jain sutuuatmut**"

— Erreurs d'identification des groupes nominaux.

  Parfois un nom qui suit ou précède une entité nommée en inuktitut est considéré comme une partie de l'entité, vu qu'on a considéré les suites de noms comme étant des EN.

  Exemple :

  (iu) Nuqqausirutigilugu , uqausirikkannirumavakka katimajiuqatima ukausirisimajangit **taivit alakannuap**, sinnattuumajunnaiqpugut.

  (en) In closing, I would like to echo comments by my colleague **Ovide Alakannuark**, we are no longer dreaming .

  L'entité *PER* **Ovide Alakannuark**, a été alignée avec tout le groupe no-

minal **ukausirisimajangit taivit alakannuap** au lieu de **taivit alakannuap**.

— Les erreurs de traduction dues aux mots hors vocabulaire et le domaine restreint des données.

Ceci est dû au domaine des données utilisées qui concerne l'assemblée législative. Contrairement au dictionnaire construit à partir de la base de données du projet UQAILAUT, les paires de mots proviennent du domaine général, ainsi que les mots hors vocabulaire qui n'admettent pas de traduction.

Exemples :

(en) "Legislative Assembly Of Nunavut". $\Longrightarrow$ (iu) "maligaliurvia Ralaa Jumaar Nunavut", au lieu de **nunavut maligaliurvia**.

(en) "South Baffin" $\Longrightarrow$ (iu) "Nginni baffin", au lieu de "qikiqtaaluup niggiani".

À travers l'analyse des erreurs, nous avons constaté des lacunes dans nos deux modèles. Nous avons tout de même constaté que la méthode basée sur les plongements des mots est moins performante que la méthode basée sur les règles à cause du changement qu'elle apporte à la traduction des EN. Une liste d'entités nommées (anglais-inuktitut) se trouve dans l'annexe de notre mémoire.

5.7 Conclusion

Dans ce chapitre, nous avons introduit les données utilisées pour l'évaluation de nos deux approches proposées et nous avons présenté les différentes mesures d'évaluations que nous utilisons pour analyser les performances de nos modèles. Par la suite, nous avons effectué une analyse comparative de deux méthodes en évaluant

la qualité de l'alignement, la projection des entités nommées, la traduction des mots et la traduction automatique. Enfin, nous avons présenté une analyse des erreurs.

CONCLUSION

Dans ce mémoire, nous avons construit un système de reconnaissance des entités nommées en inuktitut, une langue inuite du Canada. Comptée parmi les quatre grands ensembles dialectaux des langues inuites, l'inuktitut est écrit à l'aide du syllabaire autochtone canadien. En effet, c'est une langue peu dotée qui ne dispose pas de données labellisées. Cela présente un grand défi quant à la construction de notre système REN. Aussi, la langue inuktitut a une grammaire particulière et des compositions de mots assez complexes qui la différencient des autres langues. Pour pallier à ces problèmes, l'idée principale de notre approche est d'utiliser l'anglais, étant donné que c'est une langue riche en ressource et que les données labellisées sont disponibles. Nous avons construit un modèle capable de détecter les entités nommées en inuktitut, et ce, en transférant les caractéristiques linguistiques de l'anglais vers l'inuktitut.

La reconnaissance des entités nommées est une tâche cruciale dans le domaine du traitement automatique du langage. En effet, elle permet d'améliorer la qualité des systèmes comme pour la traduction automatique. Notre contribution est définie par la proposition de deux méthodes de détection d'entités nommées dans la langue peu dotée (inuktitut) et la construction d'un corpus inuktitut annoté. Dans la première approche basée sur les règles, nous avons extrait des entités nommées du corpus anglais et nous avons effectué une analyse morphologique des phrases inuktitut. De plus, nous avons identifié les groupes nominaux du texte inuktitut et nous avons utilisé l'alignement des mots pour filtrer les groupes nominaux. La deuxième méthode est basée sur les plongements des mots bilingues. Dans cette approche, nous avons entraîné les plongements des mots monolingues sur chaque

corpus (anglais, inuktitut) sur des données inuktitut segmentées et nous avons effectué une projection translingue en réalisant une association linéaire entre les deux espaces monolingues dans le même espace et en utilisant un dictionnaire bilingue.

Par ailleurs, nous avons évalué les performances des deux approches proposées et nous avons effectué une étude comparative entre les deux méthodes proposées pour l'extraction des entités nommées. Les résultats d'évaluation ont montré que les performances de la traduction automatique neuronale s'améliorent.

Ce projet, contribue à la préservation d'une langue et culture autochtone. Et ce, en construisant une base de connaissance en langue inuktitut qui contribuera à la réalisation des travaux futurs qui touchent différents sous-domaines du TALN. L'enjeu dans notre cas était de relever les défis de la langue et de se familiariser avec sa morphologie puisque c'est une langue polysynthétique. De plus, nous avons été confrontés au manque de ressources telles que les données d'évaluation et les dictionnaires de projection et d'évaluation des plongements des mots, que nous nous avons dû construire.

Dans nos travaux futurs, nous proposons d'intégrer des bases de connaissances telles que celles liées à la toponymie et des données provenant du savoir et des connaissances autochtones dans l'entraînement des plongements des mots et l'amélioration des performances de notre système. Pour cela, une collaboration avec une communauté autochtone du Nunavut dont la langue maternelle est l'inuktitut est primordiale. Aussi, nous proposons d'utiliser, en plus des résultats obtenus, des méthodes d'apprentissage profond. Plus précisément, entraîner un modèle basé sur les transformeurs, afin de détecter les entités nommées dans les corpus de la langue inuktitut autres que *Nunavut Hansard Inuktitut–English*, et ce, en utilisant les corpus annotés que nous avons construits à l'aide des modèles présentés dans

ce mémoire. Cela contribuera à l'enrichissement des données étiquetées dans la langue inuktitut, ce qui permet de faciliter l'exploitation du domaine de recherche lié au TALN et ses tâches en aval.

ANNEXE A

TABLEAU 5.9 – Listes d'entités nommées (anglais-inuktitut).

Anglais	**Inuktitut**	**Tag**
1st Session	sivuliqpaat katimaniq	MISC
1st Assembly	sivuliqpaat maligaliurvik	ORG
HANSARD	katimautigisimajangitta	ORG
Legislative Assembly of Nunavut	nunavut maligaliurvia	ORG
Ovide Alakannuark	uuviq alakkannuaq	PER
Akulliq	akulliq	PER
Enoki Irqittuq	inuuki iqqittuq	PER
Arviat	arviat	LOC
Glenn McLean	gilin maklain	PER
Baker Lake	qamaniqtuaq	LOC
Kelvin Ng	kiulvin ing	PER
Cambridge Bay	iqaluttuutsiaq	LOC
Peter Kattuk	pita kattuq	PER
Hudson Bay	sanikiluaq	LOC
Hunter Tootoo	hanta tutu	PER
Iqaluit Centre	iqaluit qitingani	LOC
Ed Picco	ituaq piiku	PER
Iqaluit East	iqaluit kanannangani	LOC
Paul Okalik	paal ukaliq	PER

Iqaluit West	iqaluit uannangani	LOC
Donald Havioyak	taanut haviujaq	PER
James Arvaluk	jaimisi arvaallu	PER
Uriash Puqiqnak	juuraijas pukiqnaq	PER
Nattilik	natsilik	LOC
Peter Kilabuk	piita qilavva	PER
Pangnirtung	panniqtuuq	LOC
Levi Barnabas	liivai paanapaasi	PER
Jack Anawak	jak anaruaq	PER
Rankin Inlet North	kangiqlhiniup tuniviani	LOC
Thompson	taamsan	PER
Rankin Inlet South	kangiqlhiniup niggiani	LOC
Olayuk Akesuk	ulaajuk akisuk	PER
South Baffin	qikiqtaaluup niggiani	LOC
Jobie Nutarak	juupi nutaraq	PER
David Iqaqrialu	taiviti ikkarrialuk	PER
John Quirke	jaan kuark	PER
Levi Barnabas	livai paanapas	PER
Kelvin Ng	kiavin ning	PER
Olayuk Akesuk	ulajuk akisuk	PER
Ovide Alakannuark	uviti alakkannuaq	PER
Government of Canada	kanataup	ORG
Rankin Inlet South	kangiqiniup niggiani	LOC
Whale Cove	tikirarjuaq	LOC
Olayuk Akesuk	ulaajuk akisuk	PER
South Baffin	qikiqtaaluup niggiani	LOC

Jobie Nutarak	juupi nutaraq	PER
David Iqaqrialu	taiviti ikkarrialuk	PER
John Quirke	jaan kuaq	PER
Rhoda Perkison	ruta pukisan	PER
Clerk of Committees	katimajiralaanut alaati	ORG
Nancy Tupik	naansi tupiq	PER
Susan Cooper	suusan kuupa	PER
Arms	paliisinga	ORG
Simon Nattaq	saiman nataa	PER
Innirvik Support Services	iniqvi ikajuqtiit	ORG
Iqaluit	iqaluit	LOC
Nunavut	nunavut	LOC
IQALUIT	iqaluit	LOC
NUNAVUT	nunavut	LOC
Legislative Assembly	maligaliurvik	ORG
Akesuk	akisu	PER
Alakannuark	ulakannuaq	PER
Anawak	anugaaq	PER
Mr. Arvaluk	jaim arvaalluk	PER
Mr. Barnabas	vaanavaas	PER
Mr. Havioyak	Haviujaq	PER
Mr. Iqaqrialu	ikkarrialuk	PER
Mr. Irqittuq	inuuki	PER
Mr. Kattuk	katuk	PER
Mr. Kilabuk	qilavvaq	PER
Mr. McLean	maklain	PER

Mr. Ng	kauvin ing	PER
Mr. Nutarak	nutaraq	PER
O'Brien	uuvuraian	PER
Mr. Okalik	ukaliq	PER
Mr. Picco	piiku	PER
Mr. Puqiqnak	uraias pukirnaq	PER
Mrs . Thompson	taamsin	PER
Mr. Tootoo	tuutuu	PER
Quirke	kuak	PER
Prayer	tutsiarniq	PER
Tatiigiit Group	tatigiikkut	ORG
Iqaluit Drummers	iqaluit qilaujjaqtingit	ORG
Drum Dancing	qilaujjaqtullu	MISC
Legislative Assembly of Nunavut	nunavuup maligaliurvinganni	ORG
Alakannuark	alakannuaq	PER
Rules of the Legislative Assembly	maligaliurviup	ORG
House	katimavvingmi	ORG
Canada	kanatami	LOC
Inuk Charlie	inuk saali	PER
Joseph Suqslaq	juusipi	PER
Paul Malliki	paa maliki	PER
Mariano Aupilardjuk	miarianu aupilaarjuk	PER
Mathew Nuqingaq	maatiu ningiungat	PER
Sam Pitsiulak	saam pitsiulaaq	PER
Assembly	maligaliurviup	ORG
Territory	avittuqsimajumik	LOC

Inuit	inuttigut	MISC
Inuit	inuulluta	MISC
English	qallunaatitut	MISC
French	uiviititullu	MISC
Land Claim	nunataarutimik	MISC
Manitoba	maanituuvvamik	LOC
Yukon Territory	juukaanmik	LOC
Saskatchewan	saskaatsuan	LOC
Alberta	auvvuuta	LOC
Manitoba	maanituuvva	LOC
Northwest Territories	nunattiarmi	LOC
Canadians	kanatamiut	MISC

RÉFÉRENCES

ACL (2022). Acl 2022. https ://www.2022.aclweb.org/papers. En ligne. Consulté en août 2022.

Adelani, D. I., Hedderich, M. A., Zhu, D., Berg, E. v. d. et Klakow, D. (2020). Distant supervision and noisy label learning for low resource named entity recognition : A study on hausa and yorùbá. http://dx.doi.org/10.48550/ARXIV.2003.08370. Récupéré de https://arxiv.org/abs/2003.08370

Akbik, A., Blythe, D. et Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. Dans *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649., Santa Fe, New Mexico, USA. Association for Computational Linguistics. Récupéré de https://aclanthology.org/C18-1139

Aone, C., Halverson, L., Hampton, T. et Ramos-Santacruz, M. (1998). SRA : Description of the IE2 system used for MUC-7. Dans *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. Récupéré de https://aclanthology.org/M98-1012

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., Kehler, A., Martin, D., Myers, K. et Tyson, M. (1995). SRI International FASTUS SystemMUC-6 test results and analysis. Dans *Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. Récupéré de https://aclanthology.org/M95-1019

Argetsinger, T. (2017). National inuit position paper on federal legislation in relation to inuktut.

Artetxe, M., Labaka, G. et Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. Dans *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2289–2294., Austin, Texas. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D16-1250. Récupéré de https://aclanthology.org/D16-1250

Artetxe, M., Labaka, G. et Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. 451–462. http://dx.doi.org/10.18653/v1/P17-1042

Asgari, E., Sabet, M. J., Dufter, P., Ringlstetter, C. et Schütze, H. (2020). Subword sampling for low resource word alignment. http://dx.doi.org/10.48550/ARXIV.2012.11657. Récupéré de https://arxiv.org/abs/2012.11657

Azmat, A., Xiao, L., Yating, Y., Rui, D. et Turghun, O. (2020). Constructing Uyghur name entity recognition system using neural machine translation tag projection. Dans *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, 1006–1016., Haikou, China. Chinese Information Processing Society of China. Récupéré de https://aclanthology.org/2020.ccl-1.93

Babych, B. et Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. Dans *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Bari, M. S., Joty, S. et Jwalapuram, P. (2020). Zero-resource cross-lingual na-

med entity recognition. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7415–7423.

Battiste, M. (2013). Promising practices in indigenous languages research and reconciliation in canadian education. Dans *D. Newhouse, J. Orr, & T. A. Program (Eds.), Aboriginal knowledge for economic development (pp. xii-xvii).*, Black Point, Nova Scotia and Winnipeg, Manitoba : Fernwood.

Bengio, Y., Simard, P. et Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Bharadwaj, A., Mortensen, D., Dyer, C. et Carbonell, J. (2016). Phonologically aware neural model for named entity recognition in low resource transfer settings. 1462–1472. http://dx.doi.org/10.18653/v1/D16-1153

Bird, S. (2009). Natural language processing and linguistic fieldwork. *Computational Linguistics*, *35*, 469–474. http://dx.doi.org/10.1162/coli.35.3.469

Bird, S. (2020). Decolonising speech and language technology. Dans *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519., Barcelona, Spain (Online). International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.313. Récupéré de https://aclanthology.org/2020.coling-main.313

Black, W. J., Rinaldi, F. et Mowatt, D. (1998). FACILE : Description of the NE system used for MUC-7. Dans *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. Récupéré de https://aclanthology.org/M98-1014

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. et Mercer, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Com-*

putational Linguistics, *19*(2), 263–311. Récupéré de https://aclanthology.org/J93-2003

Bustamante, G., Oncevay, A. et Zariquiey, R. (2020). No data to crawl ? monolingual corpus creation from PDF files of truly low-resource languages in Peru. Dans *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2914–2923., Marseille, France. European Language Resources Association. Récupéré de https://aclanthology.org/2020.lrec-1.356

Chandar, Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. et Saha, A. (2014). An autoencoder approach to learning bilingual word representations. http://dx.doi.org/10.48550/ARXIV.1402.1454. Récupéré de https://arxiv.org/abs/1402.1454

Chen, C., Sun, M. et Liu, Y. (2020a). Mask-align : Self-supervised neural word alignment. http://dx.doi.org/10.48550/ARXIV.2012.07162. Récupéré de https://arxiv.org/abs/2012.07162

Chen, E. et Schwartz, L. (2018). A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. Dans *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). Récupéré de https://aclanthology.org/L18-1416

Chen, Y., Liu, Y., Chen, G., Jiang, X. et Liu, Q. (2020b). Accurate word alignment induction from neural machine translation. Dans *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 566–576., Online. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.emnlp-main.42. Récupéré de https://aclanthology.org/2020.emnlp-main.42

Cheng, P. et Erk, K. (2019). Attending to entities for better text understanding. http://dx.doi.org/10.48550/ARXIV.1911.04361. Récupéré de https://arxiv.org/abs/1911.04361

Chiruzzo, L., Amarilla, P., Ríos, A. et Giménez Lugo, G. (2020). Development of a Guarani - Spanish parallel corpus. Dans *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2629–2633., Marseille, France. European Language Resources Association. Récupéré de https://aclanthology.org/2020.lrec-1.320

Chiu, J. P. C. et Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. *CoRR*, *abs/1511.08308*. Récupéré de http://arxiv.org/abs/1511.08308

Clark, A. et Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*, (p. 197 – 220).

Collignon, B. (2002). Les toponymes inuit, mémoire du territoire : étude de l'histoire des inuinnait. *Anthropologie et Sociétés*, *26*(2-3), 45–69. http://dx.doi.org/https://doi.org/10.7202/007048ar

Collignon, B. (2004). Recueillir les toponymes inuit. pour quoi faire ? *Études/Inuit/Studies*, *28*(2), 89–106. Récupéré le 2022-12-16 de http://www.jstor.org/stable/42870185

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. et Kuksa, P. (2011). Natural language processing (almost) from scratch. http://dx.doi.org/10.48550/ARXIV.1103.0398. Récupéré de https://arxiv.org/abs/1103.0398

comité de coordination de la recherche au Canada (2020). Établir de nouvelles orientations à l'appui de la recherche et de la

formation en recherche autochtone au canada 2019 - 2022. Récupéré de https://www.canada.ca/fr/comite-coordination-recherche/priorites/recherche-autochtone/plan-strategique-2019-2022.html

Compton, R. (2016, dernière modification 2021). Inuktitut. Dans *The canadian encyclopedia*. Récupéré de https://www.thecanadianencyclopedia.ca/en/article/inuktitut

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. et Jégou, H. (2017). Word translation without parallel data. http://dx.doi.org/10.48550/ARXIV.1710.04087. Récupéré de https://arxiv.org/abs/1710.04087

Dekhili, G. (2020). *Apport d'une approche hybride à base de réseaux de neurones et de statistiques pour la reconnaissance des entités nommées en domaines général et restreint*. (Thèse de doctorat).

Deng, Y. et Byrne, W. (2008). Hmm word and phrase alignment for statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(3), 494–507. http://dx.doi.org/10.1109/TASL.2008.916056

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. Dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186., Minneapolis, Minnesota. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1423. Récupéré de https://aclanthology.org/N19-1423

Drapeau, L. (2011). *Les langues autochtones du Québec*. Dossiers du Conseil de la langue française. Presses de l'Université du Québec. Récupéré de https://books.google.ca/books?id=pYlQnyYxF0UC

Du, Lim et Tan (2019). A novel human activity recognition and prediction in smart home based on interaction. *Sensors*, *19*, 4474. http://dx.doi.org/10.3390/s19204474

Duan, M., Fasola, C., Rallabandi, S. K., Vega, R., Anastasopoulos, A., Levin, L. et Black, A. W. (2020). A resource for computational experiments on Mapudungun. Dans *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2872–2877., Marseille, France. European Language Resources Association. Récupéré de https://aclanthology.org/2020.lrec-1.350

Dufter, P., Zhao, M., Schmitt, M., Fraser, A. et Schütze, H. (2018). Embedding learning through multilingual concept induction. Dans *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 1520–1530., Melbourne, Australia. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P18-1141. Récupéré de https://aclanthology.org/P18-1141

Dyer, C., Chahuneau, V. et Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. Dans *NAACL*.

Ehrmann, M., Turchi, M. et Steinberger, R. (2011). Building a multilingual named entity-annotated corpus using annotation projection. 118–124.

Fang, M. et Cohn, T. (2017). Model transfer for tagging low-resource languages using a bilingual dictionary. Dans *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, 587–593., Vancouver, Canada. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P17-2093. Récupéré de https://aclanthology.org/P17-2093

Feldman, I. et Coto-Solano, R. (2020). Neural machine translation models

with back-translation for the extremely low-resource indigenous language Bribri. Dans *Proceedings of the 28th International Conference on Computational Linguistics*, 3965–3976., Barcelona, Spain (Online). International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.351. Récupéré de https://aclanthology.org/2020.coling-main.351

Feng, X., Feng, X., Qin, B., Feng, Z. et Liu, T. (2018). Improving low resource named entity recognition using cross-lingual knowledge transfer. Dans *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4071–4077. International Joint Conferences on Artificial Intelligence Organization. http://dx.doi.org/10.24963/ijcai.2018/566. Récupéré de https://doi.org/10.24963/ijcai.2018/566

Fong, Y. S., Ranaivo-Malançon, B. et Yeo, A. W. (2011). Nersil - the named-entity recognition system for iban language. Dans *PACLIC*.

Gers, F. A. et Schmidhuber, J. (2001). Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, *12 6*, 1333–40.

Goodman, J. (2001). Classes for fast maximum entropy training. http://dx.doi.org/10.48550/ARXIV.CS/0108006. Récupéré de https://arxiv.org/abs/cs/0108006

Gouvernement du Canada (2019). Government of canada introduces historic legislation on indigenous languages. https ://www.canada.ca/en/canadian-heritage/news/2019/02/government-of-canadaintroduces-historiclegislationonindigenous-languages.html. En ligne. Consulté en août 2022.

gouvernement du Canada (2021a). Inuit. Récupéré de https://www.rcaanc-cirnac.gc.ca/fra/1100100014187/1534785248701

gouvernement du Canada (2021b). Métis. Récupéré de https://www.rcaanc-cirnac.gc.ca/fra/1100100014427/1535467913043

gouvernement du Canada (2021c). Premières nations. Récupéré de https://www.rcaanc-cirnac.gc.ca/fra/1100100013791/1535470872302

gouvernement du Canada (2022a). Récupéré de https://www.sshrc-crsh.gc.ca/home-accueil-fra.aspx

gouvernement du Canada (2022b). Peuples et communautés autochtones. Récupéré de https://www.rcaanc-cirnac.gc.ca/fra/1100100013785/1529102490303

gouvernement du Canada Conseil de recherches en sciences humaines (2022). Recherche autochtone. Récupéré de https://www.sshrc-crsh.gc.ca/society-societe/community-communite/indigenous_research-recherche_autochtone/index-fra.aspx

Gouvernement du Québec (2012). Banque de noms de lieux du québec. https ://toponymie.gouv.qc.ca/ct/ToposWeb/recherche.aspx ?avancer=oui. En ligne. Consulté en décembre 2022.

gouvernement du Québec (2022). Promotion des noms géographiques utilisés par les autochtones du québec. Récupéré de https://toponymie.gouv.qc.ca/ct/toponymie-autochtone/promotion-noms-geographiques-utilises-autochtones-quebec/

Graves, A. (2013). Generating sequences with recurrent neural networks. Récupéré de https://arxiv.org/abs/1308.0850

Guo, J., Xu, G., Cheng, X. et Li, H. (2009). Named entity recognition in query. Dans *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 267–274.

Gutierrez-Vasques, X. et Mijangos, V. (2017). Low-resource bilingual lexicon extraction using graph based word embeddings. *ArXiv*, *abs/1710.02569*.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R. et Fluck, J. (2005). Prominer : rule-based protein and gene entity recognition. *BMC bioinformatics*, *6*(1), 1–9.

Hatami, A., Mitkov, R. et Corpas Pastor, G. (2021). Cross-lingual named entity recognition via FastAlign : a case study. Dans *Proceedings of the Translation and Interpreting Technology Online Conference*, 85–92., Held Online. INCOMA Ltd. Récupéré de `https://aclanthology.org/2021.triton-1.10`

Hedderich, M. A. et Klakow, D. (2018). Training a neural network in a low-resource setting on automatically annotated noisy data. `http://dx.doi.org/10.48550/ARXIV.1807.00745`. Récupéré de `https://arxiv.org/abs/1807.00745`

Ho, A. K. N. et Yvon, F. (2020). Neural baselines for word alignment. `http://dx.doi.org/10.48550/ARXIV.2009.13116`. Récupéré de `https://arxiv.org/abs/2009.13116`

Hochreiter, S. et Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. `http://dx.doi.org/10.1162/neco.1997.9.8.1735`

Huang, Z., Xu, W. et Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. `http://dx.doi.org/10.48550/ARXIV.1508.01991`. Récupéré de `https://arxiv.org/abs/1508.01991`

Huck, M., Dutka, D. et Fraser, A. (2019). Cross-lingual annotation projection is effective for neural part-of-speech tagging. Dans *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, 223–233., Ann Arbor, Michigan. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W19-1425. Récupéré de https://aclanthology.org/W19-1425

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. et Wilks, Y. (1998). University of Sheffield : Description of the LaSIE-II system as used for MUC-7. Dans *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. Récupéré de https://aclanthology.org/M98-1007

Hunt, B., Chen, E., Schreiner, S. L. et Schwartz, L. (2019). Community lexical access for an endangered polysynthetic language : An electronic dictionary for St. Lawrence Island Yupik. Dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 122–126., Minneapolis, Minnesota. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-4021. Récupéré de https://aclanthology.org/N19-4021

IDIL2022-2032 (2022). 2022- 2032 décennie internationale des langues autochtones : à propos de l'idil2022-2032. https ://fr.idil2022-2032.org/about-2022-2032/. En ligne. Consulté en août 2022.

Jain, A., Paranjape, B. et Lipton, Z. C. (2019). Entity projection via machine translation for cross-lingual ner. http://dx.doi.org/10.48550/ARXIV.1909.05356. Récupéré de https://arxiv.org/abs/1909.05356

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556.

Jiao, Q. et Zhang, S. (2021). A brief survey of word embedding and its recent development. Dans *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, 1697–1701. http://dx.doi.org/10.1109/IAEAC50856.2021.9390956

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D. et Micher, J. (2020). The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. Dans *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2562–2572., Marseille, France. European Language Resources Association. Récupéré de https://aclanthology.org/2020.lrec-1.312

Joseph, S., Sedimo, K., Kaniwa, F., Hlomani, H. et Letsholo, K. (2016). Natural language processing : A review. *Natural Language Processing : A Review*, *6*, 207–210.

Kim, J.-H. et Woodland, P. C. (2000). A rule-based named entity recognition system for speech input. Dans *Sixth International Conference on Spoken Language Processing*.

Knowles, R., Stewart, D., Larkin, S. et Littell, P. (2020). NRC systems for the 2020 Inuktitut-English news translation task. Dans *Proceedings of the Fifth Conference on Machine Translation*, Online.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007). Moses : Open source toolkit for statistical machine translation. Dans *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180., Prague, Czech Republic. Association for Computational Linguistics. Récupéré de https://aclanthology.org/P07-2045

Koehn, P., Martin, J., Mihalcea, R., Monz, C. et Pedersen, T. (dir.) (2005). *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan. Association for Computational Linguistics. Récupéré de https://aclanthology.org/W05-0800

Krupka, G. R. et Hausman, K. (1998). IsoQuest inc. : Description of the NetOwl$^{\text{tm}}$ extractor system as used for MUC-7. Dans *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. Récupéré de https://aclanthology.org/M98-1015

Kuhn, R., Davis, F., Désilets, A., Joanis, E., Kazantseva, A., Knowles, R., Littell, P., Lothian, D., Pine, A., Running Wolf, C., Santos, E., Stewart, D., Boulianne, G., Gupta, V., Maracle Owennatékha, B., Martin, A., Cox, C., Junker, M.-O., Sammons, O., Torkornoo, D., Thanyehténhas Brinklow, N., Child, S., Farley, B., Huggins-Daines, D., Rosenblum, D. et Souter, H. (2020). The indigenous languages technology project at NRC Canada : An empowerment-oriented approach to developing language software. Dans *Proceedings of the 28th International Conference on Computational Linguistics*, 5866–5878., Barcelona, Spain (Online). International Committee on Computational Linguistics. Récupéré de https://aclanthology.org/2020.coling-main.516

Kuru, O., Can, O. A. et Yuret, D. (2016). CharNER : Character-level named entity recognition. Dans *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, 911–921., Osaka, Japan. The COLING 2016 Organizing Committee. Récupéré de https://aclanthology.org/C16-1087

Langlais, P., Gotti, F. et Cao, G. (2005). NUKTI : English-Inuktitut word alignment system description. Dans *Proceedings of the ACL Workshop on Building*

and Using Parallel Texts, 75–78., Ann Arbor, Michigan. Association for Computational Linguistics. Récupéré de https://aclanthology.org/W05-0810

Le, N. T. et Sadat, F. (2021). Towards a first automatic unsupervised morphological segmentation for Inuinnaqtun. Dans *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 159–162., Online. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2021.americasnlp-1.17. Récupéré de https://aclanthology.org/2021.americasnlp-1.17

Le, T. N. et Sadat, F. (2020). Low-resource NMT : an empirical study on the effect of rich morphological word segmentation on Inuktitut. Dans *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1 : Research Track)*, 165–172., Virtual. Association for Machine Translation in the Americas. Récupéré de https://aclanthology.org/2020.amta-research.15

Li, J., Sun, A., Han, J. et Li, C. (2018). A survey on deep learning for named entity recognition. *CoRR*, *abs/1812.09449*. Récupéré de http://arxiv.org/abs/1812.09449

Li, J., Sun, A., Han, J. et Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. http://dx.doi.org/10.1109/TKDE.2020.2981314

Lin, D. et Cherry, C. (2003). Proalign : Shared task description. Dans *Proceedings of the HLT/NAACL 2003 Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics. Récupéré de https://www.microsoft.com/en-us/research/publication/proalign-shared-task-description/

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C. et Junker,

M.-O. (2018). Indigenous language technologies in Canada : Assessment, challenges, and successes. Dans *Proceedings of the 27th International Conference on Computational Linguistics*, 2620–2632., Santa Fe, New Mexico, USA. Association for Computational Linguistics. Récupéré de https://aclanthology.org/C18-1222

Luoma, J. et Pyysalo, S. (2020). Exploring cross-sentence contexts for named entity recognition with BERT. Dans *Proceedings of the 28th International Conference on Computational Linguistics*, 904–914., Barcelona, Spain (Online). International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.78. Récupéré de https://aclanthology.org/2020.coling-main.78

Luong, M.-T., Pham, H. et Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. http://dx.doi.org/10.48550/ARXIV.1508.04025. Récupéré de https://arxiv.org/abs/1508.04025

Ma, W., Yu, H., Zhao, K., Zhao, D. et Yang, J. (2020). Tibetan-chinese cross-lingual word embeddings based on muse. Dans *Journal of Physics : Conference Series*, volume 1453, p. 012043. IOP Publishing.

Mager, M., Çetinoğlu, Ö. et Kann, K. (2020). Tackling the low-resource challenge for canonical segmentation. Dans *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5237–5250., Online. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.emnlp-main.423. Récupéré de https://aclanthology.org/2020.emnlp-main.423

Mager, M., Mager, E., Medina-Urrea, A., Meza Ruiz, I. V. et Kann, K. (2018). Lost in translation : Analysis of information loss during machine translation between polysynthetic and fusional languages. Dans *Proceedings of the Workshop on*

Computational Modeling of Polysynthetic Languages, 73–83., Santa Fe, New Mexico, USA. Association for Computational Linguistics. Récupéré de https://aclanthology.org/W18-4808

Makarov, P. et Clematide, S. (2018). Imitation learning for neural morphological string transduction. Dans *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2877–2882., Brussels, Belgium. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D18-1314. Récupéré de https://aclanthology.org/D18-1314

Marchisio, K., Xiong, C. et Koehn, P. (2021). Embedding-enhanced giza++ : Improving alignment in low- and high- resource scenarios using embedding space geometry. http://dx.doi.org/10.48550/ARXIV.2104.08721. Récupéré de https://arxiv.org/abs/2104.08721

Mayhew, S., Tsai, C.-T. et Roth, D. (2017). Cheap translation for cross-lingual named entity recognition. Dans *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2536–2545., Copenhagen, Denmark. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D17-1269. Récupéré de https://aclanthology.org/D17-1269

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013a). Efficient estimation of word representations in vector space. http://dx.doi.org/10.48550/ARXIV.1301.3781. Récupéré de https://arxiv.org/abs/1301.3781

Mikolov, T., Le, Q. V. et Sutskever, I. (2013b). Exploiting similarities among languages for machine translation.

Mollá, D., Van Zaanen, M. et Smith, D. (2006). Named entity recognition for question answering. Dans *Proceedings of the Australasian language technology workshop 2006*, 51–58.

Moore, R. C. (2004). Improving ibm word alignment model 1. Dans *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 518–525.

Nadeau, D., Turney, P. D. et Matwin, S. (2006). Unsupervised named-entity recognition : Generating gazetteers and resolving ambiguity. Dans L. Lamontagne et M. Marchand (dir.). *Advances in Artificial Intelligence*, 266–277., Berlin, Heidelberg. Springer Berlin Heidelberg.

Newmark, P. (1981). *Approaches to translation / Peter Newmark.* Language teaching methodology series. Oxford ; New York : Pergamon Press.

Ni, J., Dinu, G. et Florian, R. (2017). Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. http://dx.doi.org/10.18653/v1/p17-1135. Récupéré de http://dx.doi.org/10.18653/v1/P17-1135

Och, F. J. (2003). Minimum error rate training in statistical machine translation. Dans *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, p. 160–167., USA. Association for Computational Linguistics. http://dx.doi.org/10.3115/1075096.1075117. Récupéré de https://doi.org/10.3115/1075096.1075117

Och, F. J. et Ney, H. (2000a). A comparison of alignment models for statistical machine translation. Dans *COLING 2000 Volume 2 : The 18th International Conference on Computational Linguistics*. Récupéré de https://aclanthology.org/C00-2163

Och, F. J. et Ney, H. (2000b). Improved statistical alignment models. Dans *Proceedings of the 38th Annual Meeting of the Association for Computatio-*

nal Linguistics, 440–447., Hong Kong. Association for Computational Linguistics. http://dx.doi.org/10.3115/1075218.1075274. Récupéré de https://aclanthology.org/P00-1056

Olah, C. (2015). Understanding lstm networks. [En ligne; publié le 27 Août 2015].

Östling, R. (2015). Bayesian models for multilingual word alignment.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D. et Auli, M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. http://dx.doi.org/10.48550/ARXIV.1904.01038. Récupéré de https://arxiv.org/abs/1904.01038

Pennington, J., Socher, R. et Manning, C. (2014). GloVe : Global vectors for word representation. Dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543., Doha, Qatar. Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/D14-1162. Récupéré de https://aclanthology.org/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. et Zettlemoyer, L. (2018). Deep contextualized word representations. http://dx.doi.org/10.48550/ARXIV.1802.05365. Récupéré de https://arxiv.org/abs/1802.05365

Petkova, D. et Croft, W. B. (2007). Proximity-based document representation for named entity retrieval. Dans *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 731–740.

Post, M. (2018). A call for clarity in reporting BLEU scores. Dans *Proceedings of the Third Conference on Machine Translation : Research Papers*, 186–191., Brussels, Belgium. Association for Computational Linguistics. http://

dx.doi.org/10.18653/v1/W18-6319. Récupéré de https://aclanthology.org/W18-6319

Pourdamghani, N., Ghazvininejad, M. et Knight, K. (2018). Using word vectors to improve word alignments for low resource machine translation. Dans *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, 524–528., New Orleans, Louisiana. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N18-2083. Récupéré de https://aclanthology.org/N18-2083

Ruzsics, T. et Samardžić, T. (2017). Neural sequence-to-sequence learning of internal word structure. Dans *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 184–194., Vancouver, Canada. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/K17-1020. Récupéré de https://aclanthology.org/K17-1020

Sabet, M. J., Dufter, P. et Schütze, H. (2020). Simalign : High quality word alignments without parallel training data using static and contextualized embeddings. Dans *FINDINGS*.

Sachan, D. S., Xie, P., Sachan, M. et Xing, E. P. (2018). Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. Dans *Machine learning for healthcare conference*, 383–402. PMLR.

Sak, H., Senior, A. et Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. http://dx.doi.org/10.48550/ARXIV.1402.1128. Récupéré de https://arxiv.org/abs/1402.1128

Sapci, A. O. B., Tastan, O. et Yeniterzi, R. (2021). Focusing on possible named en-

tities in active named entity label acquisition. http://dx.doi.org/10.48550/ARXIV.2111.03837. Récupéré de https://arxiv.org/abs/2111.03837

Schweter, S. et Baiter, J. (2019). Towards robust named entity recognition for historic german. http://dx.doi.org/10.48550/ARXIV.1906.07592. Récupéré de https://arxiv.org/abs/1906.07592

Sennrich, R., Haddow, B. et Birch, A. (2015). Neural machine translation of rare words with subword units. http://dx.doi.org/10.48550/ARXIV.1508.07909. Récupéré de https://arxiv.org/abs/1508.07909

Sharma, A., Katrapati, G. et Sharma, D. M. (2018). IIT(BHU)–IIITH at CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. Dans *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task : Universal Morphological Reinflection*, 105–111., Brussels. Association for Computational Linguistics. Récupéré de https://aclanthology.org/K18-3013

Sierra Martínez, G., Montaño, C., Bel-Enguix, G., Córdova, D. et Mota Montoya, M. (2020). CPLM, a parallel corpus for Mexican languages : Development and interface. Dans *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2947–2952., Marseille, France. European Language Resources Association. Récupéré de https://aclanthology.org/2020.lrec-1.360

Sigl, S. (2020). Text classification demystified : An introduction to word embeddings. [En ligne; publié le 20-Janvier-2020].

Statistique Canada (2017). Les langues autochtones des premières nations, des métis et des inuits. Récupéré de https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-fra.cfm

Stengel-Eskin, E., Su, T.-R., Post, M. et Van Durme, B. (2019). A discriminative neural model for cross-lingual word alignment. http://dx.doi.org/10.48550/ARXIV.1909.00444. Récupéré de https://arxiv.org/abs/1909.00444

Tang, X., Cheng, S., Do, L., Min, Z., Ji, F., Yu, H., Zhang, J. et Chen, H. (2018). Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages. Récupéré de https://arxiv.org/abs/1810.08740

The UQAILAUT Project. (2012). *Inuktitut Computing : The UQAILAUT Project*. Récupéré de https://www.inuktitutcomputing.ca/Uqailaut/info.php

Tian, L., Wong, F. et Chao, S. (2011). Word alignment using giza++ on windows. Dans *MTSUMMIT*.

Tiedemann, J. (2003). Combining clues for word alignment. Dans *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Tiedemann, J., Agić, Ž. et Nivre, J. (2014). Treebank translation for cross-lingual parser induction. Dans *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 130–140., Ann Arbor, Michigan. Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/W14-1614. Récupéré de https://aclanthology.org/W14-1614

Tjong Kim Sang, E. F. et De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. Dans *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147. Récupéré de https://aclanthology.org/W03-0419

Tran, Q., MacKinlay, A. et Jimeno Yepes, A. (2017). Named entity recognition with stack residual LSTM and trainable bias decoding. Dans *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Vo-*

lume 1 : Long Papers), 566–575., Taipei, Taiwan. Asian Federation of Natural Language Processing. Récupéré de https://aclanthology.org/I17-1057

Unesco (2019). International conference language technologies for all (lt4all) : Enabling linguistic diversity and multilingualism worldwide. https ://en.unesco.org/events/international-conference-language-technologies-all-lt4all-enabling-linguistic-diversity-and. En ligne. Consulté en décembre 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. et Polosukhin, I. (2017). Attention is all you need. Récupéré de https://arxiv.org/abs/1706.03762

Vogel, S., Ney, H. et Tillmann, C. (1996a). HMM-based word alignment in statistical translation. Dans *COLING 1996 Volume 2 : The 16th International Conference on Computational Linguistics*. Récupéré de https://aclanthology.org/C96-2141

Vogel, S., Ney, H. et Tillmann, C. (1996b). *HMM-based Word Alignment in Statistical Translation Archived*. The 16th International Conference on Computational Linguistics.

Wang, H. et Lepage, Y. (2016). Combining fast_align with hierarchical subsentential alignment for better word alignments. Dans *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, 1–7., Osaka, Japan. The COLING 2016 Organizing Committee. Récupéré de https://aclanthology.org/W16-4501

Wang, M. et Manning, C. D. (2013). Cross-lingual pseudo-projected expectation regularization for weakly supervised learning. http://dx.doi.org/10.48550/ARXIV.1310.1597. Récupéré de https://arxiv.org/abs/1310.1597

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F. et Tu, K. (2021). Automated concatenation of embeddings for structured prediction. Dans *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, 2643–2660., Online. Association for Computational Linguistics. Récupéré de https://aclanthology.org/2021.acl-long.206

Xie, J., Yang, Z., Neubig, G., Smith, N. A. et Carbonell, J. (2018). Neural cross-lingual named entity recognition with minimal resources. Récupéré de https://arxiv.org/abs/1808.09861

Yadav, V. et Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *CoRR*, *abs/1910.11470*. Récupéré de http://arxiv.org/abs/1910.11470

Yamada, I., Asai, A., Shindo, H., Takeda, H. et Matsumoto, Y. (2020). LUKE : Deep contextualized entity representations with entity-aware self-attention. Dans *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6442–6454., Online. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.emnlp-main.523. Récupéré de https://aclanthology.org/2020.emnlp-main.523

Yarowsky, D. et Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. Dans *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 207–216., Hong Kong. Association for Computational Linguistics. http://dx.doi.org/10.3115/1075218.1075245. Récupéré de https://aclanthology.org/P00-1027

Yohannes, H. M. et Amagasa, T. (2022). Named-entity recognition for a low-resource language using pre-trained language model. Dans *Proceedings of the*

37th ACM/SIGAPP Symposium on Applied Computing, SAC '22, p. 837–844., New York, NY, USA. Association for Computing Machinery. http://dx.doi.org/10.1145/3477314.3507066. Récupéré de https://doi.org/10.1145/3477314.3507066

Young, T., Hazarika, D., Poria, S. et Cambria, E. (2017). Recent trends in deep learning based natural language processing. http://dx.doi.org/10.48550/ARXIV.1708.02709. Récupéré de https://arxiv.org/abs/1708.02709

Zenkel, T., Wuebker, J. et DeNero, J. (2020). End-to-end neural word alignment outperforms GIZA++. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1605–1617., Online. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.146. Récupéré de https://aclanthology.org/2020.acl-main.146

Zhang, S. et Elhadad, N. (2013). Unsupervised biomedical named entity recognition : Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, *46*(6), 1088–1098. Special Section : Social Media Environments, http://dx.doi.org/https://doi.org/10.1016/j.jbi.2013.08.004. Récupéré de https://www.sciencedirect.com/science/article/pii/S1532046413001196

Zhang, S., Frey, B. et Bansal, M. (2020). Chren : Cherokee-english machine translation for endangered language revitalization. http://dx.doi.org/10.48550/ARXIV.2010.04791. Récupéré de https://arxiv.org/abs/2010.04791

Zhang, Y., Gaddy, D., Barzilay, R. et Jaakkola, T. (2016). Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. Dans *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*,

1307–1317., San Diego, California. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N16-1156. Récupéré de https://aclanthology.org/N16-1156

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. et Liu, Q. (2019). Ernie : Enhanced language representation with informative entities. *arXiv preprint arXiv :1905.07129*.

Östling, R. et Tiedemann, J. (2016). Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, *106*. http://dx.doi.org/10.1515/pralin-2016-0013