Les utilisations possibles de l'Intelligence Artificielle dans la linguistique historique

3 étudiants de CPBx

0.1 Résumé

0.2 Abstract

0.3 Remerciements

Table des matières

0.4 Table des figures

0.5 Notations

Chapitre 1

Introduction

Mise en contexte pour arriver à la problématique, quel est le potentiel de l'intelligence artificielle dans la linguistique historique ? ... ?

Chapitre 2

La linguistique historique et l'Intelligence Artificielle

2.1 La linguistique historique

2.1.1 Introduction à la linguistique historique

Définir ce qu'est la linguistique historique, ce qu'elle étudie, et les mots de vocabulaires que nous allons rencontrer tout au long du mémoire.

2.1.2 Les différents principes

Évidemment cette science repose sur des concepts, allant des propriétés synchoniques des mots aux à leurs aspects diachroniques.

2.1.3 Les atouts de l'Intelligence Artificielle dans ce domaine

La linguistique historique fait face à de nombreux problèmes récurrents (traiter une grande quantité de textes pour l'homme, remarquer des motifs dans ces documents historiques). Alors que ce travail pourrait être effectué par une machine, grâce à sa capacité à traiter un grand nombres de données, et à chercher des similarités dans ces données. Avant, de voir les tâches où l'Intelligence Artificielle peut intervenir, il est d'abord nécessaire de voir en détail la conception des ces IA.

Résoudre des problèmes de Linguistique Historique avec un ordinateur nécessite de lui faire traiter du contenu textuel devant être abstrait sur des terrains parmi ceux de la **phonétique**, de la **sémantique**, de la **morphologie** ou encore de la **syntaxe**.

Développper un exemple pour illustrer ces 4 niveaux d'abstractions

La réalisation de ces abstractions s'inscrit dans le Traitement Automatisé du Langage Naturel (TAL), un domaine à cheval entre la Linguistique et l'Informatique. L'Intelligence Artificielle y occupe une place centrale pour sa capacité à effectuer des approximations améliorables avec de l'entraînement.

2.2 L'IA dans le Traitement Automatisé du Langage Naturel

2.2.1 Introduction à l'apprentissage automatique

Qu'est ce qu'une intelligence artificielle ?
Qu'est ce qu'un réseau de neurones ?
Quel est le principe derrière l'apprentissage automatique ?
Définition des apprentissages supervisés/non supervisés Définition de propagation avant. Définition rétro-propagation du gradient. Exemple de FFNN pour tâche de classification

De nombreux problèmes informatiques peuvent être résolus à travers la détermination d'une fonction mathématique f d'un $\mathbb{K}$-espace vectoriel E vers un $\mathbb{K}$-espace F (avec $\mathbb{K}$ correspondant à $\mathbb{R}$ ou $\mathbb{C}$).
Ainsi, lorsqu'une fonction informatique conventionnelle effectue un traitement sur une chaîne de caractères, une fonction f a déjà été implicitement déterminée pour réaliser la tâche. La séquence de n caractères encodés sous forme de bits forme un vecteur de l'espace $\mathbb{R}^n$ et la chaîne renvoyée par f est bien un élément d'un espace $\mathbb{R}^{n'}$.

Il est aussi très fréquemment difficile – voire impossible – de poser une expression mathématique ou un algorithme pour répondre à certains problèmes. Dans ce cas là, f est considérée comme hypothétique et on cherche à l'approcher à partir d'un **modèle**, qu'on construit à partir des informations qu'on dispose sur f, comme un ensemble de ses points $\{(x_k, y_k = f(x_k)), k \in S\}$.

Les **réseaux de neurones** sont d'excellents outils pour établir des modèles. Mathématiquement, ce sont des compositions d'applications non-linéaires et linéaires recevant un vecteur d'entrée représentant une donnée et sortant un vecteur de sortie représentant un résultat dans un format cohérent avec le problème.
Le neurone artificiel le plus élémentaire effectue la **somme pondérée** des coefficients du vecteur d'entrée, puis calcule l'image de cette somme à travers une **fonction d'activation**. La sortie du neurone est donc un réel ou un complexe. Si on la note y_j, qu'on note x le vecteur d'entrée dans $\mathbb{K}^n$, w_j le vecteur de **poids** associé au neurone et σ sa fonction d'activation, on a :

$$y_j = \sigma(\sum_{i=0}^{n} w_{ij} x_i) = \sigma(< w_j, x >) \tag{2.1}$$

introduire la matrice W pour le calcul d'une couche
dire que les informations dans chaque coeff du vecteur x sont souvent abstraites et n'ont de sens que pour la machine (teaser vers sous- section suivante) + leur importance est déterminée par les poids avec l'entraînement
décrire l'entraînement (régression logistique+fct perte)
Exemple de la tâche de classification (connotation textuelle)
dire que σ est un paramètre du réseau
on peut empiler plusieurs réseaux de neurones et plusieurs couches (paramètre architecturale)

2.2.2 Les pré-traitements nécessaires du texte.

tokenisation + normalisation des données
Vectorisation sémantique des mots (embeddings)
Encodage d'embeddings (statique ou contextuelle)

2.2.3 Architectures neuronales utiles au TAL

Quels sont les différents outils ? Suivant, comment les parties précédentes ont été traités, ou comment les parties futures seront discutées, cette partie pourrait ne pas être nécessaire. Sinon, elle regroupera l'idée de comment on passe de notre langue naturelle à celle de la machine, de passer aux mots à des vecteurs ? Quels traitements théoriques (théoriques pour ce distinguer de la pratique dans la partie future) doient être effectués sur les mots ? En fait cette partie fait référence aux chapitres 2 et 6 de Jurasky. Voir même le chapitre 9, en supprimant la sous partie précendente pour pouvoir parler directement des réseaux de neurones appliqués à la linguistique, en d'autres termes, des réseaux récurrents, des modèles séquentielles (encodeurs-décodeurs) avec l'attention, et des Transformers.

Réseaux de neurones récurrents

Transformeurs

Chapitre 3

Les contributions de l'IA dans la linguistique historique

C'est la partie 'Related Work', elle discute des différents aspects où la linguistique historique s'applique, à travers différents modèles.

3.1 Restoration de documents anciens

3.2 Déchiffrement de langues anciennes

Chapitre 4

Étude du cas de l'application de l'IA pour la reconstruction des proto-formes d'une langue

Ici, on se place dans un cas concret, pour montrer que ce n'est pas que de la théorie. En proposant une expérience.

4.1 État de l'art

4.1.1 Conceptualisation du problème

Définir clairement le problème du titre, énoncer et justifier le choix de notre modèle réseaux de neurones et des différents outils appliqués. Voir s'il est possible de faire apparaitre plusieurs démarches, c'est à dire, une approche statistique et une approche neuronale (toujours pour renforcer et montrer le potentiel de l'IA).

4.1.2 Dernières solutions neuronales

Solution supervisée + non supervisée

4.1.3 Limites d'applicabilité

Expliquer en quoi le non-supervisé donne plus d'espoir que le supervisé mais en quoi même cette approche présente des limites.
Transition avec la problématique de l'article scientifique

4.2 Observation expériementale d'une limite d'applicabilité d'une approche

4.2.1 Méthode

4.2.2 Récupération de la base de données

Il est fort possible que cette partie se regroupe avec la partie suivante, car il n'y aura pas grand chose à dire.

4.2.3 Potabiliser les données

Le choix de "potabiliser", et non pas normaliser, est volontaire. En effet, la normalisation de nos données s'effectura dans un second temps dans les différentes démarches. Il reste ici quelques sous parties à détailler.

4.2.4 Analyse

4.2.5 Critiques

Il reste ici quelques sous parties à détailler.

Chapitre 5

Conclusion

5.1 Synthèse

Résume tout ce qui a été dit.

5.2 Les différentes limites posées aujourd'hui

une partie des limites aura déjà été traitée dans le chapitre précédent. Cette sous partie se veut résumer ces limites, et aller dans les limites générales (voir acutelles) de l'IA dans la linguisitique historique.

5.3 Les perspectives de l'IA dans la linguistique historique

Ouverture, dépassement de certaines limites, évolution des modèles.

Chapitre 6

Références

6.1 Bibliographie