Rapport PSTALN

Cléa Han, Yanis Labeyrie et Adrien Zabban

janvier 2024

1 Introduction

Le but de ce projet consistait à entraîner un modèle de langage à prédire certains attributs des mots d'une phrase. Une des tâches sur lesquelles nous nous sommes concentrées est appelée le balisage de séquence prototypique. Cette tâche consiste à attribuer des étiquettes à chaque élément d'une séquence en fonction de son prototype ou modèle. Par exemple, dans le domaine de la reconnaissance d'entités nommées, on pourrait avoir une phrase en anglais comme :

"John works at Microsoft in Paris."

Dans cette phrase, la tâche de balisage de séquence consisterait à attribuer des étiquettes à chaque mot, indiquant s'il s'agit d'un nom propre, d'une entreprise, d'une localité, etc. Un exemple de balisage de séquence prototypique (POS-tagging) pour cette phrase pourrait être :

"John [name] works [verb] at [pronoun] Microsoft [noun] in [pronoun] Paris[noun]."

Ici, chaque mot est étiqueté en fonction de son prototype ou de sa catégorie, ce qui permet de comprendre le rôle de chaque élément dans la séquence.

Nous nous sommes également concentrés sur une autre tâche : la prédiction des traits morpho syntaxiques. Par exemple prédire à quel personne est un verbe, si un nom est au pluriel, ect..

2 Dataset d'entraînement et préparation des données

Parler ici du dataset, comment il est traité, le padding, les mots inconnus ....

3 Méthodologie et Modèle

Pour réaliser cette tâche nous avons opté pour l'utilisation d'un réseau de neurone récurent de type LSTM (Long-Short Term Memory) bidirectionnel. En effet, cette approche classique est intéressante car le résau LSTM bidirectionnel permet d'analyser les dépendances entre les mots potentiellement éloignés d'une phrase, or ceci est nécessaire pour démêler les ambiguïtés lors de l'étiquetage morpho-syntaxie. Considérons l'exemple suivant :

Phrase : "Je ne peux pas ouvrir le fichier que tu m'as envoyé."

Sans la capacité de capturer les dépendances à long terme, une approche classique pourrait avoir du mal à démêler l'ambiguïté dans la phrase, en particulier en ce qui concerne le mot "que". Est-ce une conjonction de subordination introduisant une proposition subordonnée relative, ou est-ce une conjonction de coordination ?

Avec un réseau LSTM bidirectionnel, on peut mieux gérer ces dépendances à long terme. Le réseau peut prendre en compte le contexte des mots précédents et suivants pour déterminer que "que" est utilisé comme une conjonction de subordination dans ce cas précis. L'utilisation de la bidirectionnalité permet au modèle d'avoir une compréhension contextuelle plus riche, en analysant à la fois les mots antérieurs et postérieurs pour chaque position dans la séquence.

Le réseau LSTM que nous avons employé est constitué d'un module LSTM bidirectionnel (comprenant 2 couches avec 2 directions opposées) ainsi que d'une couche dense de $[n_neurone]$ qui récupère l'état caché de la dernière cellule LSTM est produit le vecteur des prédictions.

Nous utilisons la CrossentropyLoss pour comparer les vecteurs de densité de probabilité de classe prédit par le modèle (un pour chaque mot de la phrase avec le vecteur one-hot correspondant à la vrai classe du mot). Nous prédisons un nombre $[n_classes]$ d'attributs de POS-tagging possibles.

Pour la tâche de prédiction des traits morphologiques, la difficulté est supérieure. En effet pour cette tâche il faut pour chaque mot de la phrase prédire à la fois son rôle dans la phrase (Verbe, Nom, ect..) mais aussi pour chacune de ses classes prédire des attributs (singulier, pluriel, ...).

Pour cela une première méthode serait d'entraîner un modèle distinct par attributs, mais cela demanderai un assez long temps d'entraînement. Nous avons donc opté pour une première méthode différente : prédire non pas un vecteur comme pour la première tâche mais une matrice [classe,attribut] sans changer l'architecture du réseau, simplement en changeant le nombre de neurones de la couche dense. Nous avons par la suite tenté une deuxième approche : diviser la sortie de la couche dense en la faisant passer dans un nombre n de couche dense égal au nombre de classes, chacune chargée de prédire un attribut et reformer la matrice après passage dans ces couches distinctes. L'intéret de cette approche est que l'on utilise les features extraites du bloc LSTM de manière commune pour les sous-tâches mais chaque sous-tâche de classification est réalisé par une couche dense distincte.

4 Métriques et Résultats

Nous avons donc lancé un entraînement de notre réseau LSTM-Bidirectionel sur 30 epochs pour le POS-tagging et nous avons obtenu une accuracy de validation d'environ 9O% ce qui dénote que le réseau à réussi à apprendre correctement cette tâche d'étiquetage.

En revanche pour la tâche de prédiction des traits morphologiques les résultats se sont avérés bien inférieurs avec une accuracy de validation de seulement 20

Parler ici des différentes métriques macro/micro/accuracy

Mettre ici le graphe d'entrainement

Mettre ici des exemples d'inférences (un cas simple, un cas difficile ambigu)