**Résumé de l’article : Remplacement de mentions pour l’adaptation d’un corpus de reconnaissance d’entités nommées à un domaine cible**

**Contexte du travail**

La reconnaissance d’entités nommées est une tâche de traitement automatique du langage naturel bien étudiée et utile dans de nombreuses applications, comme l’extraction d’informations. Grâce aux avancées en apprentissage profond, ce domaine a connu des progrès significatifs, notamment avec des modèles de langue neuronaux contextuels pré-entraînés, tels que BERT, qui ont montré des performances remarquables sur des corpus bien annotés et standardisés. Cependant, ces corpus se concentrent sur des domaines spécifiques, ce qui réduit la performance d’un modèle entraîné sur un corpus lié à un domaine (domaine source) lorsqu'il est appliqué à un autre domaine (domaine cible).

**Problématique**

Les modèles de reconnaissance d’entités nommées entraînés ou réentraînés sur un corpus d’un domaine spécifique (ex. journalistique) sont moins performants lorsqu’ils sont appliqués à un autre domaine (ex. fantasy) en raison du manque d’entités nommées propres au domaine cible. L’article cherche à répondre à la question suivante : comment adapter un corpus annoté d'un domaine source vers un domaine cible moins bien couvert ?

**Détail des solutions**

L’article propose une méthode d’augmentation des données, le remplacement de mentions. Cette méthode consiste à générer de nouveaux exemples en remplaçant une entité nommée par une autre du même type (ex. remplacer « Emmanuel Macron » par « Gandalf »).

L’article applique cette technique pour adapter le jeu de données CoNLL-2003, qui représente généralement le domaine journalistique, au domaine de la fantasy représenté par le jeu de données Dekker, selon plusieurs configurations :

• Sans augmentation.

• Avec augmentation interne (version du CoNLL-2003 modifiée).

• Avec augmentation externe hors-domaine (WikiGold).

• Avec augmentation externe intra-domaine (listes d’entités de fantasy).

• Avec augmentation parfaite (listes d’entités extraite du domaine cible directement).

**les observations/apports**

* L’augmentation de F-mesure est notable avec des données augmentées par une augmentation externe intra-domaine (fantasy) avec un taux d’augmentation de 5 %, ainsi que pour le rappel dans les mêmes configurations, mais avec un taux d’augmentation de 50 %.
* Les données pertinentes liées au domaine cible jouent un rôle crucial dans l’amélioration des performances, comme l’indiquent les configurations idéales et ‘externes intra-domaines’.
* L’ajout de noms non pertinents pour le domaine cible n’apporte pas d’amélioration significative des performances.

**Limites du travail**

L’étude effectuée dans cet article se limite uniquement aux noms de personnes et à un seul domaine spécifique. De plus, la technique proposée pourrait être enrichie par d'autres méthodes d'augmentation de données afin d’apporter plus de variété dans les exemples générés.