RER – NLP Named Entity Recognition

Contexte :

Se familiariser avec les techniques de Deep Learning pour reconnaitre les entités nommées (NER).

Problématique(s) :

Quelles sont les techniques de deep learning pour reconnaitre les entités nommées (NER)?

Comment le NER permet-il d’extraire de l’information des données textuelles?

Mots clés :

* NER (Named Entity Recognition) : La reconnaissance d'entités nommées (NER ou Named Entity Recognition en anglais) est une technique de traitement automatique du langage naturel qui permet d'identifier et de classer des entités nommées telles que les noms de personnes, les noms de lieux, les noms d'organisations, etc., dans un texte donné. Cela peut aider à extraire des informations importantes à partir de grandes quantités de texte et à automatiser certaines tâches, telles que l'extraction de données ou la traduction automatique.
* Corpus : Un corpus est une collection de textes ou de données linguistiques qui sont sélectionnés et organisés pour être utilisés dans des études linguistiques ou pour entraîner des modèles d'apprentissage automatique. Les corpus peuvent être constitués de différents types de textes, tels que des journaux, des livres, des articles scientifiques, des conversations enregistrées, etc. Ils sont souvent utilisés en linguistique pour l'analyse de la langue naturelle et en traitement automatique du langage naturel pour entraîner des modèles d'apprentissage automatique.
* Annotation sémantique: L'annotation sémantique est le processus consistant à ajouter des informations sémantiques à un texte, telles que la signification des mots ou des phrases, les relations entre les entités nommées, etc. Ces informations peuvent être utilisées pour faciliter la recherche d'informations pertinentes, améliorer la compréhension des textes et aider à l'automatisation de certaines tâches, telles que la traduction automatique ou la réponse automatique aux questions. L'annotation sémantique est souvent réalisée manuellement par des annotateurs humains ou à l'aide de modèles d'apprentissage automatique.
* Ontologie: Une ontologie est une représentation formelle d'un ensemble de concepts et de relations dans un domaine particulier. Elle est utilisée pour organiser et structurer les connaissances dans un domaine donné afin de faciliter leur compréhension et leur utilisation. Les ontologies sont souvent utilisées en intelligence artificielle et en informatique pour faciliter la recherche d'informations et la résolution de problèmes complexes en représentant les connaissances de manière formelle et structurée.
* Web sémantique : Le web sémantique est une extension du World Wide Web (WWW) qui permet de structurer et d'organiser les informations de manière plus intelligente et plus significative pour les ordinateurs et les utilisateurs humains. Il repose sur des standards et des technologies permettant de décrire de manière formelle les données et les connaissances, telles que les ontologies, les langages de description de données, les protocoles d'échange de données, etc. L'objectif du web sémantique est de rendre les informations sur le web plus accessibles et plus utiles pour les utilisateurs en facilitant leur compréhension et leur interprétation par les machines, et en permettant aux applications informatiques de traiter les données de manière plus efficace et plus automatique.
* Catégorie d’une entité :
* Morphosyntaxique : La morphosyntaxe est la branche de la linguistique qui étudie la structure des mots et leur combinaison en phrases dans une langue donnée. Elle s'intéresse notamment aux règles qui gouvernent la formation des mots (morphologie) et à leur ordre et leur fonction dans une phrase (syntaxe). La morphosyntaxe permet de comprendre comment les mots se transforment et interagissent pour produire un sens dans une langue donnée. Par exemple, elle peut expliquer comment les verbes s'accordent avec les sujets en genre et en nombre, ou comment les adjectifs et les adverbes modifient le sens des mots auxquels ils se rattachent.
* Word Embedding : Le Word Embedding est une technique de traitement automatique du langage naturel qui permet de représenter les mots sous forme de vecteurs de nombres réels dans un espace vectoriel de grande dimension. Cette représentation permet de capturer les relations sémantiques et syntaxiques entre les mots, en codant des informations telles que la similarité entre les mots, leur contexte d'utilisation, leurs associations avec d'autres mots, etc. Les vecteurs obtenus à partir du Word Embedding peuvent être utilisés pour entraîner des modèles d'apprentissage automatique pour différentes tâches, telles que la classification de textes, la traduction automatique, la recommandation de contenu, etc.
* Word2vec : Word2vec est un algorithme de Word Embedding largement utilisé dans le domaine du traitement automatique du langage naturel. Il permet de représenter les mots sous forme de vecteurs de nombres réels en apprenant des relations sémantiques et syntaxiques à partir de grandes quantités de données textuelles. Word2vec utilise un modèle de réseau de neurones artificiels pour prédire le contexte d'un mot donné, ou pour prédire un mot à partir de son contexte. Les vecteurs obtenus grâce à Word2vec peuvent être utilisés pour mesurer la similarité sémantique entre les mots, pour résoudre des tâches de classification de textes, de traduction automatique, de recommandation de contenu, etc. Word2vec a contribué à de nombreuses avancées dans le traitement automatique du langage naturel depuis sa création en 2013.
* Skip-gram : Skip-gram est un modèle d'algorithme de Word Embedding qui utilise un réseau de neurones artificiels pour prédire le contexte d'un mot donné. Contrairement à d'autres modèles de Word Embedding, qui utilisent une fenêtre fixe de mots pour déterminer le contexte, Skip-gram permet de prendre en compte un contexte plus large et plus flexible en considérant les mots voisins à différents niveaux de distance. Les vecteurs obtenus grâce à Skip-gram peuvent être utilisés pour mesurer la similarité sémantique entre les mots, pour résoudre des tâches de classification de textes, de traduction automatique, de recommandation de contenu, etc.
* CBOW (Continuous Bag of Words) : CBOW (Continuous Bag of Words) est un modèle d'algorithme de Word Embedding qui utilise un réseau de neurones artificiels pour prédire un mot donné à partir de son contexte. Contrairement au modèle Skip-gram qui utilise le mot en entrée pour prédire le contexte, CBOW utilise le contexte pour prédire le mot. Le modèle prend en entrée un ensemble de mots (le "bag of words") et tente de prédire le mot central. Les vecteurs obtenus grâce à CBOW peuvent être utilisés pour mesurer la similarité sémantique entre les mots, pour résoudre des tâches de classification de textes, de traduction automatique, de recommandation de contenu, etc.
* LSA : LSA (Latent Semantic Analysis) est une méthode d'analyse sémantique de textes qui permet de découvrir des relations sémantiques entre les mots. LSA repose sur une technique d'analyse de la matrice de co-occurrence des mots dans un corpus de texte, qui permet de déterminer les dimensions latentes (ou concepts) qui sous-tendent les relations entre les mots. Les vecteurs obtenus grâce à LSA peuvent être utilisés pour mesurer la similarité sémantique entre les mots, pour résoudre des tâches de classification de textes, de recommandation de contenu, etc.
* LDA : LDA (Latent Dirichlet Allocation) est une méthode d'analyse de texte utilisée pour découvrir les sujets abordés dans un corpus de texte. LDA repose sur une modélisation probabiliste des thèmes à partir des mots présents dans les documents. La méthode consiste à considérer chaque document comme une combinaison de différents sujets, et chaque sujet comme une distribution de probabilité sur les différents mots du corpus. L'objectif est de découvrir les distributions de probabilité des sujets et des mots, en utilisant des techniques d'inférence bayésienne. Les résultats obtenus grâce à LDA permettent d'identifier les thèmes récurrents dans un corpus de texte, et peuvent être utilisés pour résoudre des tâches de classification de textes, de recommandation de contenu, etc.
* LSTM : LSTM (Long Short-Term Memory) est un type de réseau de neurones récurrents (RNN) utilisé pour le traitement de séquences de données. Contrairement aux RNN classiques, les LSTM ont une capacité de stockage à long terme et peuvent conserver des informations importantes sur une séquence sur une longue période de temps. Les LSTM sont composés de plusieurs "portes" qui contrôlent le flux d'informations à l'intérieur du réseau, permettant ainsi de mémoriser des informations pertinentes et d'oublier des informations inutiles. Les LSTM sont couramment utilisés pour des tâches de traitement de langage naturel telles que la traduction automatique, la génération de texte, la reconnaissance de la parole, etc.
* FastText: FastText est un modèle de Word Embedding développé par Facebook qui permet de représenter chaque mot d'un vocabulaire sous forme de vecteurs. FastText utilise une approche basée sur des n-grammes de caractères pour générer des représentations vectorielles de mots. Cette approche permet de prendre en compte des préfixes et suffixes de mots, ce qui permet de mieux gérer les mots inconnus et de généraliser les représentations de mots à des formes inédites. Les vecteurs obtenus grâce à FastText peuvent être utilisés pour mesurer la similarité sémantique entre les mots, pour résoudre des tâches de classification de textes, de recommandation de contenu, etc. FastText est considéré comme l'un des modèles de Word Embedding les plus performants et est largement utilisé dans l'industrie et la recherche.
* Text Mining : Text Mining est une méthode d'analyse de texte utilisée pour extraire des informations utiles et significatives à partir de grands volumes de données textuelles. Text Mining utilise des techniques d'analyse de données pour traiter automatiquement des textes en langage naturel et identifier des motifs, des tendances ou des relations dans les données textuelles. Les tâches couramment réalisées avec le Text Mining comprennent l'analyse de sentiments, la catégorisation de textes, l'extraction d'informations, la recommandation de contenu, etc. Les applications du Text Mining sont nombreuses, allant de l'analyse des médias sociaux à la surveillance des opinions des consommateurs, en passant par la recherche scientifique et la médecine.
* Gensim: Gensim est une bibliothèque open-source en Python pour la modélisation de textes et la transformation de données de texte. Elle offre des outils pour la création de modèles de Word Embedding, tels que Word2Vec, FastText, ainsi que des algorithmes pour la modélisation de thèmes tels que LDA. Gensim permet également de réaliser des tâches telles que la similarité de documents, la classification de documents, l'indexation de documents, la recherche de similarités, etc. Gensim est une bibliothèque largement utilisée dans l'industrie et la recherche pour l'analyse de texte.
* Pos tag : Pos tag (Part of Speech tagging) est une technique de traitement automatique de langage naturel qui consiste à étiqueter chaque mot d'un texte avec sa catégorie grammaticale correspondante, telle que verbe, nom, adjectif, adverbe, etc. Le Pos tagging est utilisé pour aider à l'analyse sémantique, la reconnaissance de texte et la traduction automatique, entre autres tâches. En assignant des étiquettes grammaticales à chaque mot, le Pos tagging permet de mieux comprendre la signification des mots dans un texte, et donc d'effectuer des tâches de traitement de texte plus précises.
* Transformer : Transformer est une architecture de réseau de neurones pour le traitement automatique de langage naturel (NLP). Elle est conçue pour transformer les entrées textuelles en représentations vectorielles utiles et de haute qualité pour les tâches de NLP. Le Transformer a été introduit par Vaswani et al. en 2017 et est devenu l'une des architectures les plus couramment utilisées dans le domaine de la NLP. Il utilise une attention multi-têtes pour capturer les relations entre les différents mots dans un texte, ce qui lui permet d'obtenir des représentations de texte plus précises et plus expressives. Les modèles Transformer sont utilisés dans de nombreuses applications de NLP telles que la traduction automatique, la génération de texte, la classification de textes, etc.

Hypothèses :

1. LSTM a été remplacé par Transformer. Adrien
2. Le NER est la seule technique en NLP pour reconnaitre des entités nommées. Osman
3. Le NER au même titre que la normalisation est une étape de preprocessing pour le NLP. Solenn
4. Le NER peut servir à la détection de fraudes. Jean Paul 1
5. Les catégories se rapportent à la nature des mots ou à la fonction grammaticale. Aude
6. Le NER utilise la réduction de dimensions afin de faire converger la matrice de vecteur représentant les mots. Etienne
7. LDA est le modele de word embedding le plus performant. Nicolas
8. NER est un modèle de Deep Learning attaché au traitement NLP. Adeline 1ere
9. La plupart des modèles statistiques NER ne sont applicables que pour la langue anglaise. Adeline 2eme
10. Le NER permet par entrainement d’un modèle de pouvoir classer des mots de manière sémantique. Briand
11. LSTM rend NER plus performant. Tetyana
12. On peut se servir du NER pour classer les documents. Axel

VRAI

1. L'utilisation de techniques de prétraitement améliore la qualité de la reconnaissance d'entités nommées. Seydou
2. Il existe une convention de reconnaissance (nommage) des entités. Jean Paul 2
3. CBOW est appelé CBOW parce qu’il ressemble à un arc. Etienne 2

Plan d'action :

Explorer les ressources

Définir les mots clés

Répondre aux problématiques

Faire les workshops Scktit-learn, SpaCy, Gensim, LDA, Text Mining

Rendre les livrables : RER et au moins 2 workshops (Gensim est obligatoire, Word2vec)

QCM Mardi 18/04