lundi 17 mars 2025
Mélanie ZHUGE

Outils de traitement de corpus

Partie 1 - étude de cas CoNLL 2003:

1. QUELLE TYPE DE TÂCHE PROPOSE CONLL 2003 ?

CoNLL-2003 propose une tâche de Reconnaissance d'Entités Nommées (NER, Named Entity Recognition) . L'objectif est d'identifier et de classer des entités nommées dans un texte, telles que des noms de personnes, d'organisations, de lieux et d'autres expressions spécifiques.

2. QUEL TYPE DE DONNÉES Y A-T-IL DANS CONLL 2003 ?

Le corpus CoNLL-2003 contient des données textuelles annotées pour la reconnaissance d'entités nommées:

- PER (Personne)
- ORG (Organisation)
- LOC (Lieu)
- MISC (Divers)
- Articles
- Sentences
- Tokens

3. À QUEL BESOIN RÉPOND CONLL 2003 ?

CoNLL-2003 répond au besoin d'évaluer et de comparer les performances des modèles de reconnaissance d'entités nommées. Il fournit un benchmark standardisé pour la recherche en traitement automatique du langage naturel (TALN), permettant de mesurer l'efficacité des approches sur une tâche spécifique et bien définie.

4. QUELS TYPES DE MODÈLES ONT ÉTÉ ENTRAÎNÉS SUR CONLL 2003 ?

De nombreux modèles ont été entraînés et évalués sur CoNLL-2003, notamment :

- Modèles de Maximum Entropy (MaxEnt): la technique la plus utilisée avec cinq systèmes l’employant.
- Modèles de Markov cachés (HMM): 4 systèmes ont utilisé des HMM.
- Modèles des réseaux de neurones.
- Autres: Memory-based learning, Transformation-Based Learning,…etc.

5. EST UN CORPUS MONOLINGUE OU MULTILINGUE ?

C’est un corpus multilingue. Il est disponible en deux langues : anglais et allemand. Cependant, chaque version est traitée séparément, et il n'y a pas de mélange de langues dans le même ensemble de données. Le texte des données allemandes a été tiré du corpus de texte multilingue de l'ECI. Les données anglaises ont été tirées du Reuters Corpus.