

Extraction d'entités nommées en Unitex-CasEN, cascade de transducteurs

Alexane Jouglar

Le 08/10/2021

Plan

La tâche d'extraction d'entités nommées

Le logiciel CasEN

Installation et utilisation

Lien vers le GitHub : https://github.com/AlexaneJ/CasEN_priseEnMain.

La tâche d'extraction d'entités nommées

Entité nommée : Définition

« Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. » - Ehrmann 2008

Types d'entités nommées

Trois catégories d'EN¹ :

- **ENAMEX** (personnes, lieux, organisations, fonctions, etc.)
- **TIMEX** (dates, heures, et indications temporelles)
- **NUMEX** (quantités et informations chiffrées)

¹Entités nommées

Le logiciel CasEN

- outil de REN²
- sous licence LGPL-LR
- mis à disposition par l'Université de Tours
- créé dans le cadre des projets Ortolang et Istex (mise à disposition de grandes quantités de ressources langagières).

²Reconnaissance d'Entités Nommées

- Outil fonctionnant à base de règles (cascade de transducteurs)
- Utilise CaSys

Installation et utilisation

1. Télécharger le programme sur le site de l'Université de Tours : <https://tln.lifat.univ-tours.fr/version-francaise/ressources/casen>.
2. Décompresser le fichier téléchargé dans le répertoire personnel Unitex.

Utilisation avec l'interface Unitex

1. Ouvrir le texte 80jours et refuser le preprocessing
2. Dans "Text", cliquer sur "Apply lexical Resources" puis sélectionner les trois dictionnaires de la cascade (CasEN_Dico, CasEN_Ambiguites et ProlexUnitexBestOf_2_2_fra).
3. Passer la cascade d'analyse et garder le résultat.
4. Ouvrir le fichier csc.txt qui vient d'être créé en refusant à nouveau le preprocessing.
5. Passer la cascade de synthèse.

Utilisation avec les scripts

1. Télécharger les scripts sur CasEN, ou si vous avez Linux, télécharger le fichier zip que j'ai mis sur GitHub.
2. Décompresser. Le texte que l'on souhaite annoter se trouve dans exemple.
3. Dans LancementDeCasEN, modifier la version de CasEN par celle dont vous disposez.
4. Ouvrir un terminal depuis le fichier et marquer la commande suivante :
`"/lancementDeCasEN.sh"`
5. le résultat se trouve sous exemple_resultat

Merci pour votre attention.

References



Maud Ehrmann. “Les Entités Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation”. PhD thesis. Paris Diderot University, 2008.