LETTRES SORBONNE UNIVERSITÉ

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 donnés 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'EN[1] :

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

[1] Entités nommées

Le logiciel CasEN

CasEN : informations

- outil de REN[2]
- 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).

[2] Reconnaissance d'Entités Nommées

CasEN : fonctionnement

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

Principe

« La reconnaissance des entités nommées par la cascade CasEN utilise des ressources lexicales et des descriptions locales de motifs, des transducteurs qui agissent sur le texte par des insertions, remplacement ou suppressions. [...] Le principe d'une cascade est de pouvoir utiliser dans les descriptions suivantes les motifs déjà détectés ou, au contraire, d'éviter un étiquetage non souhaité pour un motif déjà reconnu. L'ordre de passage de ces transducteurs est donc un paramètre important. » Maurel et al. 2011

Exemple de sortie

DANS LEQUEL <persName ><forename>PHILEAS </forename><surname >FOGG ET PASSEPARTOUT </surname></persName>S'ACCEPTENT RÉCIPROQUEMENT L'UN COMME <roleName type="office">MAÎTRE </roleName>, L'AUTRE COMME <roleName type="office">DOMESTIQUE </roleName>

Installation et utilisation

Installation

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 Entitées Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguıisation”. PhD thesis. Paris Diderot University, 2008.

Denis Maurel et al. “Cascades de transducteurs autour de la reconnaissance des entités nommées”. In: *Traitement automatique des langues* 52.1 (2011), pp. 69–96.