

HAI922 - TALN 2 mini PROJET

"A de B"

année 2025-2026

mathieu.lafourcade@lirmm.fr / gmail.com

Le projet vise à créer un système permettant de prédire la relation sémantique entre les termes A et B dans une forme : A de B.

Par exemple :

- saucisse de toulouse ⇒ r_lieu>origine
- désert d'Algérie ⇒ r_lieu OU r_holo
- tabouret de bois ⇒ r_object>mater
-

A lire :

<https://aclanthology.org/2024.cogalex-1.4/>

https://hal.science/hal-04638014v1/file/Actes_IC_PFLIA2024.pdf (page 25)

résumé : Cette étude concerne les relations sémantiques portées par les entités sous forme génitive « A de B ». Après identification des types sémantiques pertinents, nous construisons, à l'aide d'une IA générative, un corpus annoté. Nous proposons un algorithme de découverte des règles permettant de sélectionner la relation entre A et B. Ces règles correspondent à la sélection dans une base de connaissances du voisinage adéquat d'un terme donné. Soit « désert d'Algérie », portant la relation de lieu, le terme désert identifié comme lieu géographique et Algérie comme pays. Ces contraintes aboutissent par calcul à une règle permettant de sélectionner la relation de lieu.

A lire si pas déjà fait ou si pas compris :

<https://www.jeuxdemots.org/jdm-about.php>

Point clés

- Faire aussi bien ou mieux en termes de performances que ce qui est présenté dans les deux articles.
- L'approche doit être similaire, cad avec de l'apprentissage
- Il faut utiliser les 15 types de relations sémantiques définis dans les articles, mais vous pouvez en rajouter si besoin
- Constituez vos corpus d'exemples et de tests avec des IA génératives - vérifier/contrôler les résultats. Vous pouvez utiliser : chatGPT, gemini, claud, perplexity, DeepSeek, Qwen, etc

L'apprentissage doit être relativement rapide (quelques minutes), l'inférence (l'utilisation) doit être instantanée (moins de 1 seconde)

Vous pouvez coder dans le langage de programmation de votre choix

IL FAUT IMPÉRATIVEMENT UTILISER LES DONNÉES DE JEUXDEMOTS via son API

API jeuxDeMots ici : <https://jdm-api.demo.lirmm.fr/schema>

Le calcul des vecteurs (pseudo embeddings) doit se faire au moins avec JDM, et si vous le souhaitez combiner cela avec d'autres ressources (word2vec, etc).

Vous DEVEZ coopérer entre vous (groupes de 1 ou 2 personnes). Partagez vos ressources, en particulier le corpus d'apprentissage/test

Discutez de vos algo et performances

Lien **discord** <https://discord.com/channels/368735802806763520/694960575784026173>

Lien pour le **corpus commun** :

https://drive.google.com/drive/folders/1at2pk0933Qpc3T_AcEUpGvw4-bSkZaN0?usp=drivelink

Questions

voici un lien d'un dossier en commun, il serait utile de faire un fichier par relation, comme ça c'est plus simple à modifier et pour s'organiser.

Lien pour le **corpus commun** :

https://drive.google.com/drive/folders/1at2pko933Qpc3T_AcEUpGvw4-bSkZaN0?usp=drive_link

relation de la forme: A de B | r_type