M2 IASD - HAI922I - Langage naturel 2 - 2023-24
Projet d'analyseur sémantique du français / extraction de relations

mathieu.lafourcade@lirmm.fr

Le projet vise à programmer un analyseur sémantique (simple) de textes français. L'analyse sera une structure de graphe calculée à partir du texte et sur laquelle on pourra extraire des relations sémantiques.

Les relations sémantiques typiques sont celles que l'on peut trouver dans le projet JeuxDeMots (JDM) : http://www.jeuxdemots.org/jdm-about.php

Implémentation dans le langage de votre choix. Donnez un nom à votre système. Groupe de 2-3 personnes max. Vous n'êtes pas en concurrence entre les groupes, vous pouvez coopérer. Vous êtes invités à la fin du document à partager vos idées, algo et code.

Lien vers le discord de l'UE : https://discord.gg/8UkepQQktr

Le graphe du travail

Le texte est d'abord transformé en une chaîne linéaire de mots relié par une relation (r_succ)

"le petit chat boit du lait"

```
[_START] →  [le] → [petit] → [chat] → [boit] → [du] → [lait] → [_END]
```

Des traitements (que vous devez préciser) vont rajouter des nœuds et des arc à ce graphe. Les arcs sont des relations typées, orientées et pondérées. Un poids peut être négatif. Les nœuds portent une chaîne de caractère et un poids (éventuellement un type si nécessaire).

Toutes les infos nécessaires doivent être représentées dans le graphe. Par exemple la ou les natures morphosyntaxiques des termes :

```
→ [petit] → [chat] →
      |         |
      |         |
   [Adj:]    [

]
```

La relation entre *petit* et *Adj:* est de type r_pos, orientée de 'petit' vers 'Adj:', avec un poids positif.

Autre exemple 'le' peut être un Det; (déterminant) ou un pronom (Pro:). Au début, vous rajoutez dans le graphe les deux possibilités. Ensuite avec vos règles vous déterminez laquelle est la bonne (pour la phrase ci-dessus) et vous négativez le nœud et les relations associées pour la mauvaise.

```
→  [le] → [petit] → [chat] →

 [le] - r_pos   ->  [Det:]
 [le] - r_pos<0 ->  [Pro:]
```

Quand vous parcourez votre graphe, le comportement par défaut est de ne pas considérer les nœuds et les relations à poids négatifs. Ainsi, vous gardez dans la structure des solutions envisagées mais finalement rejetées. Il ne faut pas supprimer ces nœuds, mais les négativer.

tâches

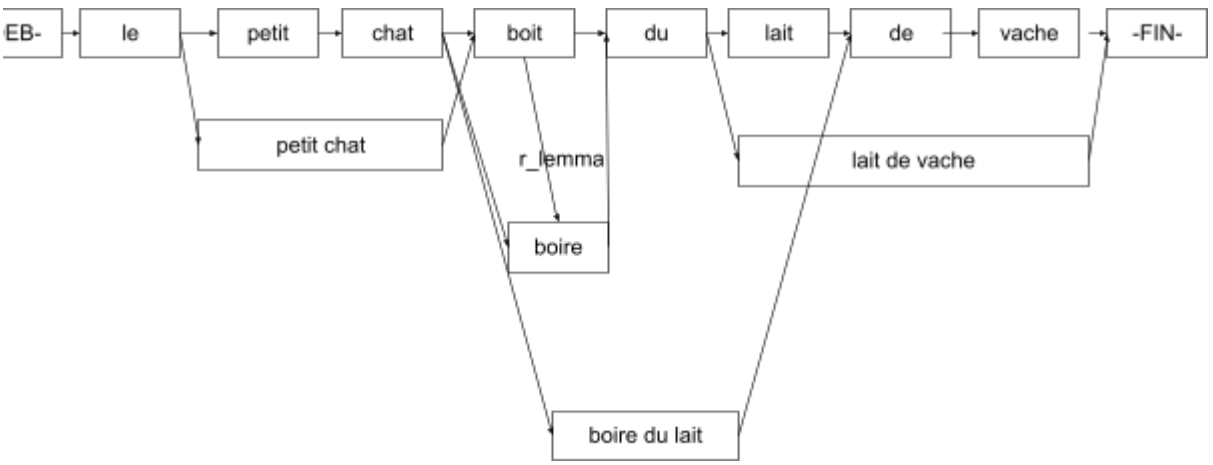
 * trouver les **termes composés** - liste dispo sur le site jdm

"le petit chat boit du lait de chèvre"

```
[_START] →  [le] → [petit] → [chat] → [boit] → [du] → [lait] → [de] →
[chèvre] → [_END]

[du] → [lait de chèvre] → [_END]
```

le chemin avec 'lait de chèvre' est parallèle au chemin d'origine, il ne le remplace pas.



* écrire un **ensemble de règles** permettant de compléter le graphe de travail

exemple de règles :

$x == GN & $y == GV & $x r_succ $y ⇒ $x r_agent-1 $y & $y r_agent $x

une règle a une partie gauche qui est une conjonction de conditions à vérifier dans le graphe, et une partie droite qui sont des actions à réaliser (des relations ou des nœuds à ajouter).

Partagez vos règles, évaluez leur pertinence à participer à l'analyse du texte.

* écrire le **moteur de règles**, qui prend votre ensemble de règles pour les exécuter

L'algo le plus basique est d'exécuter toutes les règles dans l'ordre, et de recommencer tant que le graphe est modifié (dans l'esprit d'un algorithme de algorithme de markov nlp Markov).

Assurez-vous donc que vos règles ne font pas partir en boucle votre système.

* il faut si possible **désambiguïser les termes** (désambiguïsation lexicale) qui ont plusieurs sens

```
[_START] →   → [frégate] → [coule] → [_END]

         →  [la] → [frégate>navire] →
```

Les sens disponibles sont ceux de JDM.
Comment vous y prendre ? Si vous avez par exemple A R B, et que A est un terme ayant plusieurs sens, cherchez dans JDM le A>x tel que A>x R B existe (avec un poids positif bien sûr).

Désambiguïser un terme c'est choisir le sens le plus probable dans l'inventaire de sens (toujours JDM, relation de type 1 - r_raff_sem)

* il faut si possible **résoudre les anaphores simples**. Par exemple avec les pronoms :

" le chien est tombé dans le puits. Il a aboyé toute la nuit"

on doit obtenir :

```
[Il] -r_reference     -> chien
```

```
[Il] -r_reference<0   -> puits
```

Ainsi vous pouvez extraire que le chien est bien l'agent de 'aboyer'

L'approche est sensiblement la même que pour la désambiguïsation lexicale, à savoir chercher dans JDM entre le chien et le puits, lequel est le plus susceptible d'aboyer.

Il des anaphores avec les pronom (il, elle, etc), les adjectifs possessifs (son, sa, etc.) et dans de nombreux autres cas (à vous de les déterminer)

* **résultat final** : des relations sémantiques

A fin d'analyse, on voudrait extraire des choses comme :

chat r_agent-1 boire
petit chat r_agent-1 boire

boire r_patient lait
boire r_patient lait de chèvre

petit chat r_agent-1 (boire r_patient lait)
petit chat r_agent-1 (boire r_patient lait de chèvre)

* quand la recherche d'une relation A R B ne donne rien dans JDM (la relation n'existe pas) alors vous pouvez tenter de faire une **inférence** (de type déductive).

Si on a A r_isa C et C R B alors on peut inférer A R B

d'autres schéma d'inférence sont possibles (par exemple avec les synonymes), mais attention à la polysémie.

* utilisation d'un **cache**

Si vous extrayez des information de JDM via les services proposés (voir - **Accès interactifs au réseau lexical de JeuxDeMots** dans A propos de JeuxDeMots) - pensez à mettre en place un cache de façon à ne pas demander plusieurs fois la même chose.

* **visualiser le graphe** de travail et outillage de votre système

utiliser BRAT : http://brat.nlplab.org/
NE CODEZ PAS quelque chose pour dessiner le graphe

Pensez à outiller votre système, combien de temps prennent les règles pour être évaluées ? Qu'est-ce qui prend du temps ? Dites-vous que le système doit être le plus rapide possible, donc essayez d'optimiser ce qui peut l'être. N'oubliez pas que des caches peuvent aider.

Tester pour de vrai

Essayez votre systeme sur des phrases et paragraphe issus de :
https://fr.wikisource.org/wiki/Vingt_mille_lieues_sous_les_mers/Texte_entier
(JULES VERNE VINGT MILLE LIEUES SOUS LES MERS en txt)

Idées, algo et Implémentation

(ajouter ici vos codes, remarques idées - indiquez qui vous êtes qd vous proposez quelque chose)

* graphe de travail (structure et fonction de manipulation)

* interrogation de JDM

* arbre préfixe de mots

* moteur de règles

* règles

* inférences

on se demande si x R Y ?
si on trouve x r_isa z et z R y alors on peut répondre positivement

* phrases de tests (avant d'appliquer sur le texte proposé)

le chat boit du lait

le chat boit du lait de chèvre
le petit chat n'aime pas le lait de chèvre
la souris est mangée par le chat
le chat lèche sa queue
la chatte allaite ses petits
la frégate sombre rapidement
la frégate a attrapé un poisson
la frégate a volé un poisson au pêcheur
le chat de la voisine a uriné sur le paillasson
la fusée a explosé au décollage
le garçon regarde sa voisine avec un télescope
le garçon regarde la fille avec le chien
l'enfant lèche la glace avec délice

le chien est tombé dans le puits. Il a pleuré toute la nuit.
le chien est tombé dans le puits. Il est profond et le récupérer sera compliqué.

la religieuse entre dans la pâtisserie pour acheter une religieuse

le missile fonce sur le croiseur et le coule.

les vers mangent les chairs des cadavres
les vers embellissent les cadavres exquis

la voisine vient s'excuser

Question & Réponses

Dans les mots composés, on peut supprimer les mots contenant des | et des > comme ici :

```
12209617;"Catégorie mère - Portail - Café>29592";Catégorie mère - Portail - Café>film;
12209618;"Catégorie mère - Portail - Café>100640";Catégorie mère - Portail - Café>roman;
12209619;"What's the New Mary Jane";
12209620;"Saint Louis Blues";
12209621;"Step Inside Love";
12209634;"DVD de l'édition Event V";
12209635;"mesure d'(alternaria alternata+aspergillus fumigatus+cladosporium herbarum+penicillium notatum) ac ige | multidisque | sérum |
12209636;"mesure d'(alternaria alternata+aspergillus fumigatus+cladosporium herbarum+penicillium notatum) ac ige | multidisque | sérum |
```

Version modifiée sans les pipes et chevrons :
https://torisu.fr/MODIFIED_20220927-LEXICALNET-JEUXDEMOTS-ENTRIES-MWE.txt
Écrire une cinquantaine de règles