Annotation manuelle et automatique de corpus dans GATE

Cours : Sémantique computationnelle

Didier Kouamé

Date de fin du projet : 12/12/2022

Professeure : Victoria Eyharabide

1. Présentation du projet

L'objectif de ce projet est d'annoter manuellement et automatiquement les unités linguistiques d'un corpus composé d'un livre en langue française et en langue anglaise avec le logiciel GATE. C'est une plateforme d'ingénierie linguistique utilisée pour l'extraction d'information dans différentes langues. Il s'agira donc d'utiliser certaines fonctionnalités intéressantes de cet outil comme :

- La création d'annotation manuelles ou avec des expressions régulières permettant de matcher des motifs de caractères dans tout le corpus
- La détection des langues des textes du corpus
- L'utilisation de règles à partir de transducteurs JAPE

2. Présentation du corpus

L'ensemble des chaines de traitement a été fait sur deux corpus, l'un en anglais et l'autre en français. Il y a au total 5 textes pour chaque corpus. Ce qui représente au minimum 15 000 mots pour l'ensemble.

3. Annotations manuelles

Les annotations manuelles ont été faites sur les deux corpus selon plusieurs schémas d'annotations avec parfois des expressions régulières créées pour les unités qui se répètent. Les deux premières annotations qui figurent dans la consigne sont : *les titres et les dialogues.* Mais j'ai ajouté d'autres ensembles annotations. Pour le corpus en français :

- Les titres
- Les noms propres
- Les dates ou les périodes les dates de 3 chiffres et celles de 4 chiffres
- Auteur
- Les Pays
- Les lieux (villes, pays…)
- Religion
- Roi
- Royauté
- Date ou période
- Nom propre
- Peuple
- Dynastie
- Empereur
- Loi
- Code
- Législation
- Nombre

4. Annotations avec expressions régulières

Des expressions régulières ont été utilisées pour annoter :

- **Les phrases avec les ponctuations . !?**
- **Certaines subordonnées**
- **Les bigrammes et trigrammes**

Certains mots qui se répètent plusieurs fois dans le corpus ou qui s'écrivent de la manière mais qu'il s'agit du même mot :

- **Nom de pays :** regex : Allemagnes* = chap1
- **Nom propre**s (on a désactivé l'option case sensitive)
  a. regex = Jean\sSchilter = chapitre 1 FR[1]
  b. regex = \bjustinien\b = chapitre 2 FR
  c. regex = frédéric\s\le\sgrand = chapitre 4 FR
  d. regex = frédéric\sguillaume\s1er | guillaume\s1er = chapitre 4 FR
  e. regex = karl\sgotlieb\svariez = chapitre 4 FR
  f. regex = \bsavigny[^(]\b = chapitre 5 FR
- **Années**
  Expressions pas très précises mais qui marchent quand même parce qu'ils matchent d'autres motifs qui ne sont pas des années :
  g. \b\d{4}\b
  h. \b\d{3}\b
- **Dates du type** *jour mois année*
  i. \d{2}\.\d{2}\.\d{4}
  j. **Période du type (**1265-1321**),** regex **= \d{4}+-\d{4}** les deux corpus
  k. **Jour ou période en anglais :** regex = \d{2}+th
  l. **Nombres ou chiffres:** regex = \b\d\d\b|\b\d\b = chapitre 4 EN
- **Les phrases du corpus**
  L'expression n'est pas optimale mais fait correspondre des phrases.
  m. regex = [A-Z][^. !?]*[. !?]
- Les subordonnées avec marqueur ''que'' regex = \bque\b[^.!?]*[.!?]
- Les subordonnées avec marqueur ''dont'' regex = \bdont\b[^.!?]*[.!?]
- **Certains groupes nominaux de type déterminant + substantif :**
  J'ai essayé avec la fenêtre recherche de GATE elle fonctionne mais pas avec la fenêtre d'annotation.
  \b(^un|des|le|la|les|au|aux|mon|ma|mes|ton|tes|son|sa|ses|notre|votre|vos|leur|[d|D]u|leurs|ceci|cette|combien\sde|aucun|aucune|chaque|peu\sde|plus\sde|quelque|aucunes|certaines|différentes|nulles|bon\snombre\sde|la\splupart\sdes|peu\sde|plus\sde|tellement\sde|ce\s[^a-z]{4})\s[a-zA-ZÀ-ÿ][a-zA-ZÀ-ÿ-]*{1,26}\b

- **Les phrases interrogatives :** [A-Z][^.!?]*[?]
  Comme il s'agit de livre relatant essentiellement une histoire ou présentant des faits qui se sont déroulés dans le passé, il n'y aucune phrase de type exclamative.

[1] Corpus français

On pourrait reprendre l'expression précédente pour trouver les phrases simples. Mais comme elle n'est pas optimale, elle trouvera des éléments qui ne sont pas des phrases. Dès que l'expression trouve un point, elle l'identifiera comme une phrase déclarative alors que ce n'en est pas une.

5. Manipulation de la chaine de traitement ANNIE

Le fichier tokeniser.rules par défaut de ANNIE a été modifié pour trouver : les mots, les nombres, les symboles et les espaces. On a utilisé quand même le Annie English Tokeniser qui fonctionne sur le corpus en français.

6. Dictionnaires

Un dictionnaire de lexique du droit et de certains mots récurrents dans le corpus français a été créé. Il faut noter que les mots ont une certaine correspondance avec le titre de chaque chapitre. Chaque chapitre a donc été annoté selon ses mots les plus fréquents. Mais étant donné que c'est un livre qui a un certain sujet, certains mots peuvent apparaître dans plusieurs chapitres. Le nom de l'annotation est ''Lexique de mots importants''. Voici donc le lexique de chaque chapitre.

Chapitre 1 : droit ancien germanique

Droit, loi, famille, tribunal, mariage, père, siècle, coutumier, enfants, seigneur, autorité, roi, coutume, INDIUM, femme, mari, droits, romain, rédaction.

Chapitre 2 : renaissance des droits savants

Droit, romain, église, empereur, justice, textes, juristes, tribunal, canon, juridiction, coutume, évêque, procédure, autorité.

Chapitre 3 : mutations juridiques

Droit, romain, mariage, divorce, conseil, coutume, tribunal, empereur, ville, territorial, tutelle, procédure, adultère, coutume.

Chapitre 4 : exaltation du droit naturel

Droit, code, civil, naturel, codification, loi, civile, législation, romain.

Chapitre 5 : école historique de la marche vers l'unité

Droit, loi état, peuple, Allemagne, juridique, code, commerce.

Puis un dictionnaire contenant des lieux, des pays et de leurs capitales de sorte qu'on puisse annoter les deux types de lieux. La liste est longue. L'*annotation set name* est ''Pays et Capitales'' pour le corpus FR et ''Liste de lieux'' pour le corpus EN.

- Un dictionnaire comprenant les lieux (pays et leurs capitales) FR
- Un dictionnaire de verbes EN
- Un dictionnaire d'adverbes EN
- Un dictionnaire de lieux : pays et capitales FR et pays pour EN.

7. Dictionnaires flexibles

Il semble que Flexible Gazetteer ne fonctionne pas sur le corpus en français. Les dictionnaires flexibles ont été appliqués sur le corpus en anglais. Les dictionnaires flexibles vont permettre non seulement de trouver les verbes dans leurs formes de base mais aussi les formes lemmatisées ou conjuguées de ces verbes. L'ordre des traitements est le suivant :

- *ANNIE English Tokenizer :* on commence par découper tokens les éléments,
- *ANNIE Sentence Splitter :* puis on découpe en phrases,
- *ANNIE POS tagger :* récupérer les parties du discours,
- *Gate Morphological Analyser :* exécution d'une analyse morphologique des tokens,
- *Hash Gazetteer :* appliquer le dictionnaire contenant les verbes en anglais,
- *Flexible Gazetteer :* Gate détecte automatiquement les verbes et leurs différentes formes.

Ici, j'aurais dû mettre le data store des dictionnaires flexibles dans un dossier dédié, mais j'avais fait l'analyse morphologique alors le résultat se trouve dans le dossier ''Annotations manuelles, dictionnaires et expressions régulières''.

8. Grammaires JAPE

JAPE est un transducteur à états finis qui fonctionne sur des annotations créées à partir d'expressions régulières. Les grammaires qui sont créées ici vont permettre d'automatiser certaines annotations avec des règles. On commence d'abord par créer une règle simple JAPE pour voir si elle fonctionne sur le corpus français. Etant donné que les pays sont beaucoup évoqués dans les chapitres comme *Allemagne, Rome* et bien d'autres on commence par détecter les pays et on fait bien attention de ne pas supprimer les annotations existantes :

- Règle 1 : TrouverLesVillesEtPays : Détecte les pays et les villes de tous les chapitres. FR

- Règle 2 : Passages barbares : la grammaire récupère certains passages sur le peuple barbare, un peuple non pacifique dans le chapitre 1 : droits anciens germaniques. FR

- Règle 3 : Rôle de l'église dans le droit canon : règle qui récupère les passages en ce qui concerne le rôle de l'église dans le droit canon pour le chapitre 2 : le droit canon et les universités. FR
  D'autres Lookup qui ne sont pas dans le dictionnaire apparaissent avec les passages qui parlent du rôle de l'église. Il semble y avoir une erreur quelque part.