Projet final

Sémantique computationnelle

OBJECTIF

L'objectif de ce projet est d'annoter manuellement et automatiquement dans GATE deux livres, un en français et un en anglais.

DESCRIPTION

1. **Sélection des livres** : Commencez par choisir un livre en anglais et un livre en français. Enregistrez chaque chapitre des livres dans un fichier séparé. Les livres doivent avoir les conditions suivantes : comporter au minimum 15.000 mots et au minimum 5 chapitres. Vous pouvez choisir de prendre des extraits de livres (par exemple les 5 premiers chapitres) du moment qu'ils répondent aux critères précédents.
2. **Création des corpus** : Créez deux corpus dans GATE, un corpus pour livre en français et un autre pour le livre en anglais. Dans ces corpus, vous allez déposer les différents chapitres que vous avez créés à l'étape précédente.
3. **Création d'un data store** : Créez un data store afin de sauvegarder vos corpus avec les documents annotés dans ce data store.
4. **Création des schémas d'annotation** : Créez le schéma d'annotation suivant et annotez manuellement vos documents suivant ce schéma :
   - Les titres
   - Les dialogues

   Proposez au moins deux autres schémas d'annotation et annotez vos documents en conséquence. Toutes les annotations créées avec des schémas ou avec des expressions régulières (voir point suivant) doivent être regroupées dans l'ensemble « *Annotations manuelles* ».
5. **Création des annotations avec expressions régulières** : Créez des annotations en utilisant des expressions régulières depuis la fenêtre de recherche, par exemple en ajoutant des annotations à toutes les occurrences d'un certain mot.
6. **Création des dictionnaires** : Créez au moins quatre dictionnaires permettant d'annoter votre corpus. Vous êtes libres de choisir les dictionnaires que vous considérez pertinents, par exemple sur :
   - Les personnages
   - Les lieux
   - Les évènements
   - Les références temporelles (p. ex. « ce matin », « durant la guerre »…)

   Toutes les annotations créées avec des dictionnaires (ce point et le suivant) doivent être regroupées dans le groupe « Annotations avec dictionnaires ».
7. **Création des dictionnaires flexibles :** Exécutez une analyse morphologique sur votre corpus et annotez des verbes pertinents indépendamment du temps verbal utilisé. C'est à vous de déterminer quels sont les verbes pertinents du livre, par exemple :
   - *Verbes de mouvement* : marcher, courir, se déplacer, trotter, avancer, se promener, etc.
   - *Verbes du discours*: dire, raconter, affirmer, ajouter, expliquer, répondre, ordonner, etc.

8. **Création des grammaires JAPE** : Créez des grammaires pour annoter de façon automatique des passages importants dans l'histoire racontée. Par exemple : « Quand le feu a commencé, Marie a décidé de … », « Après avoir constaté qu'elle était morte, Jacques a pris l'arme et …. », « L'histoire commence au moment où Ulysse, après bien des péripéties, se trouve retenu… », etc.

   Une fois ces types de passages annotés, voici le type d'information que l'on souhaite pouvoir extraire : a) le ***personnage*** (Marie, Jacques, Ulysse), b) le ***verbe*** qui déclenche l'action (décider, prendre, se trouver retenu), c) le ***moment*** ou contexte de la situation, s'il est precisé (quand le feu a commencé, après avoir constaté qu'elle était morte, après bien des péripéties). Dans votre rapport, indiquez quelles sont les situations que vous avez décidé d'annoter en justifiant votre choix. Toutes ces annotations doivent être regroupées dans le groupe « Annotations automatiques ».

9. **Impression d'annotations** : Créez une page HTML avec une feuille CSS afin de mettre en valeur les annotations. Visualisez cette page HTML dans votre navigateur et imprimez-la au format PDF grâce à la fonction « imprimer » du navigateur.

10. **Détection des langues** : Détectez la langue du texte de votre corpus et ajoutez-la en tant que caractéristique du document.

11. **Création d'une application avec des traitements conditionnels** : Chargez la chaîne de traitements ANNIE en choisissant les options par défaut. Créez une application à partir d'un « Conditional corpus pipeline » qui s'exécute différemment en fonction de la langue des documents. Pour les textes en anglais, vous pouvez utiliser les traitements d'ANNIE. Pour les textes en français, il faut utiliser les traitements appropriés, comme par exemple, le POS tagger TreeTagger et les dictionnaires et grammaires créés précédemment. Exécutez-la sur votre corpus. **ATTENTION** ! Pensez à supprimer le traitement « **Document reset PR** » avant d'exécuter ANNIE afin de ne pas éliminer vos annotations manuelles.

12. **Rédaction du rapport** : Ecrivez un rapport détaillant le travail effectué et en discutant l'évaluation obtenue précédemment: quelles sont les forces et faiblesses des annotations manuelle et automatique ?

ÉVALUATION

L'ensemble des documents à fournir doit être contenu dans un même dossier portant votre nom, par exemple « Dupond ». Ce dossier principal doit comporter :

- Un dossier contenant les fichiers source de votre livre.
- Un dossier contenant le data store.
- Un dossier contenant les schémas.
- Un dossier contentant les dictionnaires.
- Un dossier contenant les grammaires JAPE.
- L'application créée avec les traitements conditionnels.
- Le rapport écrit, au format PDF.

Le dossier principal doit ensuite être compressé et envoyé via la plateforme pédagogique de notre faculté :

https://moodle.paris-sorbonne.fr/

- Groupes : **Travail individuel.**
- Date : **lundi 14 décembre 2020**