Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique

Filière Génie Informatique et Digitalisation

Text Analytics

Dr. H. SEBBAQ *h.sebbaq@gmail.com*

Pr. N. EL FADDOULI *nfaddouli@gmail.com*

2023-2024

NLP : Annotation Linguistique vs NLP vs Text Analytics

L'annotation linguistique est un processus de marquage de texte avec des informations linguistiques, utilisé comme composant dans le NLP pour la compréhension et l'analyse du langage. Le Text Analytics est un domaine plus vaste qui se concentre sur l'extraction de connaissances à partir de données textuelles, souvent en utilisant des méthodes d'analyse avancées et exploitant les résultats du NLP pour des insights plus approfondis

- **Annotation Linguistique :** Ajoute des informations linguistiques (syntaxe, entités, etc.) à un texte pour le rendre compréhensible pour les machines.
- **Traitement Automatique du Langage Naturel (NLP) :** Utilise des techniques informatiques pour comprendre et traiter le langage naturel, incluant souvent l'annotation comme étape initiale.
- **Text Analytics :** Explore, extrait et analyse des informations à partir de grandes quantités de texte, en utilisant souvent les résultats du NLP pour des insights avancés.

NLP : Objectifs de l'annotation linguistique

L'annotation linguistique désigne **l'enrichissement des données textuelles non structurées en y intégrant des informations organisées et spécifiques**, telles que la syntaxe, la sémantique, les parties du discours, les entités nommées ou d'autres marqueurs linguistiques → Rendre le langage naturel compréhensible pour les systèmes informatiques en les dotant de repères et de métadonnées, facilitant ainsi la compréhension et l'analyse automatisée du texte

- ❑ Faciliter la compréhension du langage par les machines en ajoutant des informations grammaticales et sémantiques.
- ❑ Analyser les structures linguistiques pour identifier des modèles et des tendances linguistiques.
- ❑ Développer des outils de traitement du langage pour une interprétation plus précise et fonctionnelle du langage naturel.
- ❑ Établir des normes et des références pour la représentation formelle du langage, favorisant la cohérence des données linguistiques.

NLP : Rôle de l’annotation linguistique dans le Text Analytics

- **Structuration des grandes quantités de données textuelles non structurées**: Ajout d'informations linguistiques (syntaxe, sémantique, entités nommées) pour organiser et enrichir le contenu.
- **Facilitation de l'extraction d'informations** : Permet l'identification et l'extraction d'éléments pertinents (entités, relations, concepts).
- **Classification des textes** : Aide à catégoriser les documents en fonction de thèmes ou de sujets spécifiques.
- **Entraînement des modèles d'apprentissage automatique** : Utilisation des données annotées pour entraîner des modèles pour diverses tâches d'analyse textuelle.
- **Amélioration de la précision des analyses** : Fournit des repères linguistiques pour une interprétation plus précise et des résultats plus fiables lors des analyses de texte.

NLP : Pourquoi des plateformes d'annotations linguistiques ?

- Standardisation et cohérence → Une structuration uniforme et des directives claires pour des annotations de qualité, comparativement à l'approche manuelle qui peut être sujette à des interprétations variables et à des incohérences.

- Chaîne de traitement des données textuelles → préparent, structurent et fournissent des données annotées pour les étapes successives, assurant ainsi la cohérence et la qualité tout au long du processus.

- Gestion efficace des données→ offrent des fonctionnalités pour stocker, organiser et gérer de grandes quantités de données annotées, ce qui est difficile à réaliser en utilisant uniquement des scripts de programmation.

- Facilitation du travail collaboratif sur un même corpus, facilitant ainsi la collaboration, la révision et le consensus sur les annotations.

- Interface conviviale pour les non-programmeurs

NLP : Les plateformes d'annotations linguistiques existantes

UIMA (Unstructured Information Management Architecture) : https://uima.apache.org/

GATE (General Architecture for Text Engineering) : https://gate.ac.uk/

Unitex : https://unitexgramlab.org/fr

NOOJ : https://nooj.univ-fcomte.fr/index.html

Ogmios (Alvis-NLPPlatform) : https://metacpan.org/release/THHAMON/Alvis-NLPPlatform-0.6

GATE (General Architecture for Text Engineering)

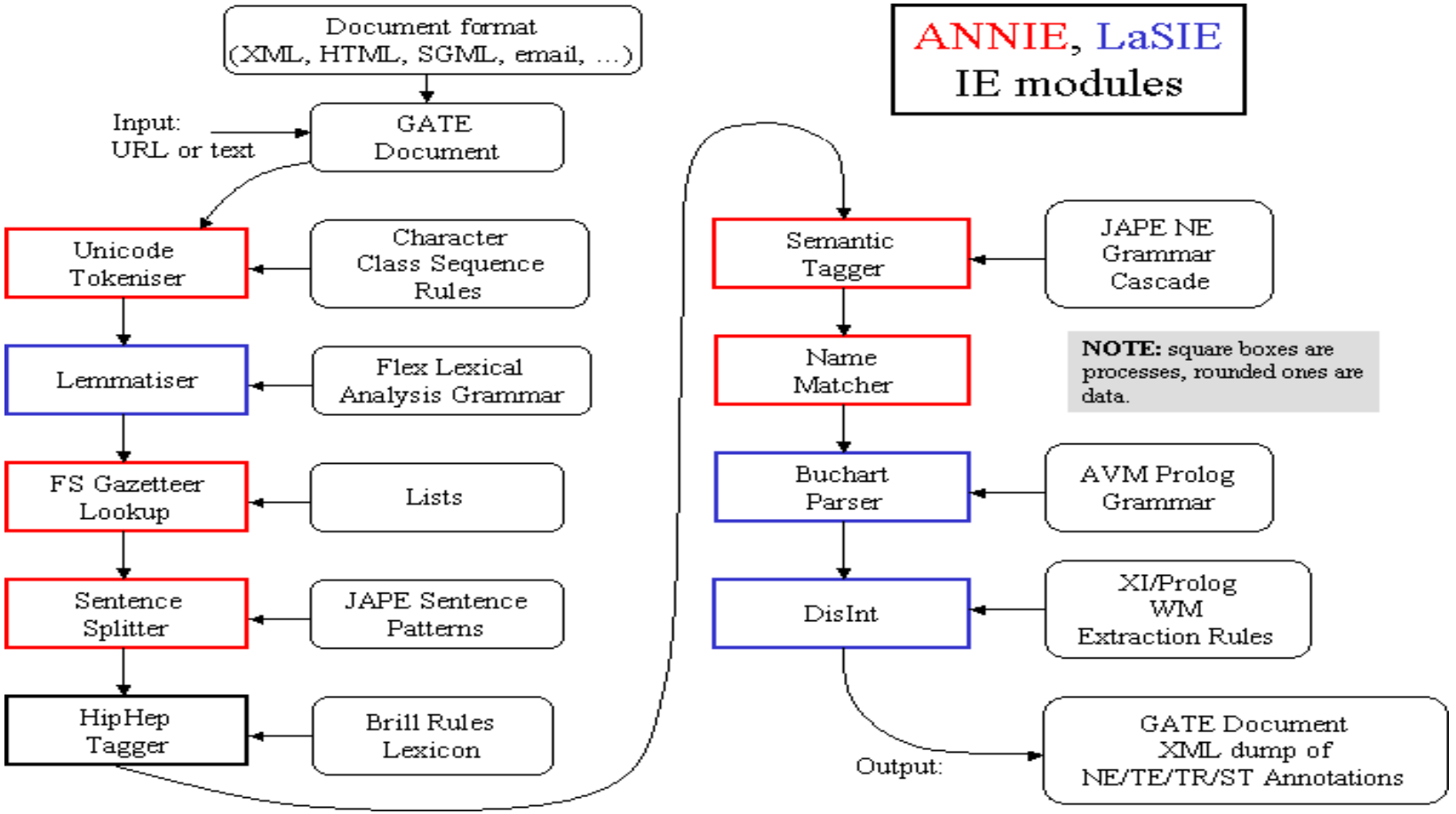
- Plateforme développée depuis 1995 à l'Université de Sheffield

- Infrastructure permettant le développement et le déploiement de composants pour le TALN

- GATE propose
  - Une architecture
  - Un framework en Java (incluant de nombreux modules) un environnement de développement autonome

Institution : Université de Sheffield
Site web : http://gate.ac.uk
Licence : open source
Plateforme : Multiplateforme
Langage : Java
Référence(s) : Cunningham, Maynard, Bontcheva, et Tablan (2002)

GATE : pipeline de modules autonomes



Fonctionne comme pipeline purement linéaire de modules autonomes

GATE : éléments de base

- **Language Ressource (LR) :** Lexiques, taxonomies, ontologies, corpus et autres ressources (support de plusieurs formats XML, RTF, HTML, SGML…)
- **Processing Ressources (PR) :** algorithmes et composants effectuant un traitement ayant pour but d'ajouter ou de transformer des annotations (sous forme d'attributs/valeurs)
- **Application ou contrôleur :** agrégation de plusieurs PR sous forme de pipeline
- **Visual Ressource** : permet la présentation des résultats à l'intérieur de l'environnement de développement GATE
- **Plugins :** composition de plusieurs LR et PR spécialisés dans l'exécution d'une tâche précise
- **CREOLE (Collection of REusable Objects for Language Engineering) :** large inventaire de ressources (LR, PR…) qui fonde GATE

GATE : Charger un document

- Conversion vers un format spécial de tout document après son chargement
- Support de presque tous les formats
- Traitement par défaut des annotations déjà existantes dans les documents (HTML tags …) grâce au paramètre **markupAware**
- **Export des documents en différents formats ou stocker dans le DataStore**
- **Paramètres d'initialisation vs paramètres d**'exécution

GATE : Annotations

- Les annotations sont des métadonnées associées au document
- « L'objectif de GATE est d'annoter les documents. Alors que les applications peuvent être utilisées pour annoter les documents de manière entièrement automatique, l'annotation peut également être effectuée manuellement, ou de manière semi-automatique, en exécutant une application sur le corpus, puis en corrigeant/ajoutant de nouvelles annotations manuellement. »
- Les annotations sont créées, supprimées et gérées par des ensembles d'annotations (Annotation set).

Lab1 : Créer un corpus

Définition du Corpus : Un corpus est une importante collection de matériel linguistique naturel écrit ou parlé, stocké sur un ordinateur. Il est compilé de manière systématique et utilisé pour l'analyse linguistique, fournissant des données linguistiques authentiques pour étudier comment la langue est utilisée.

- Créez un nouveau corpus vide, sans y ajouter de documents pour l'instant. Language ressources → New → Gate Corpus
- Cliquez avec le bouton droit de la souris sur le nom du corpus dans le panneau des ressources et sélectionnez "Populate".
- Utilisez l'icône du navigateur de fichiers pour sélectionner le nom du répertoire contenant vos documents (…/news-texts)
- Tous les documents seront chargés en une seule fois
- Afficher le contenu du corpus

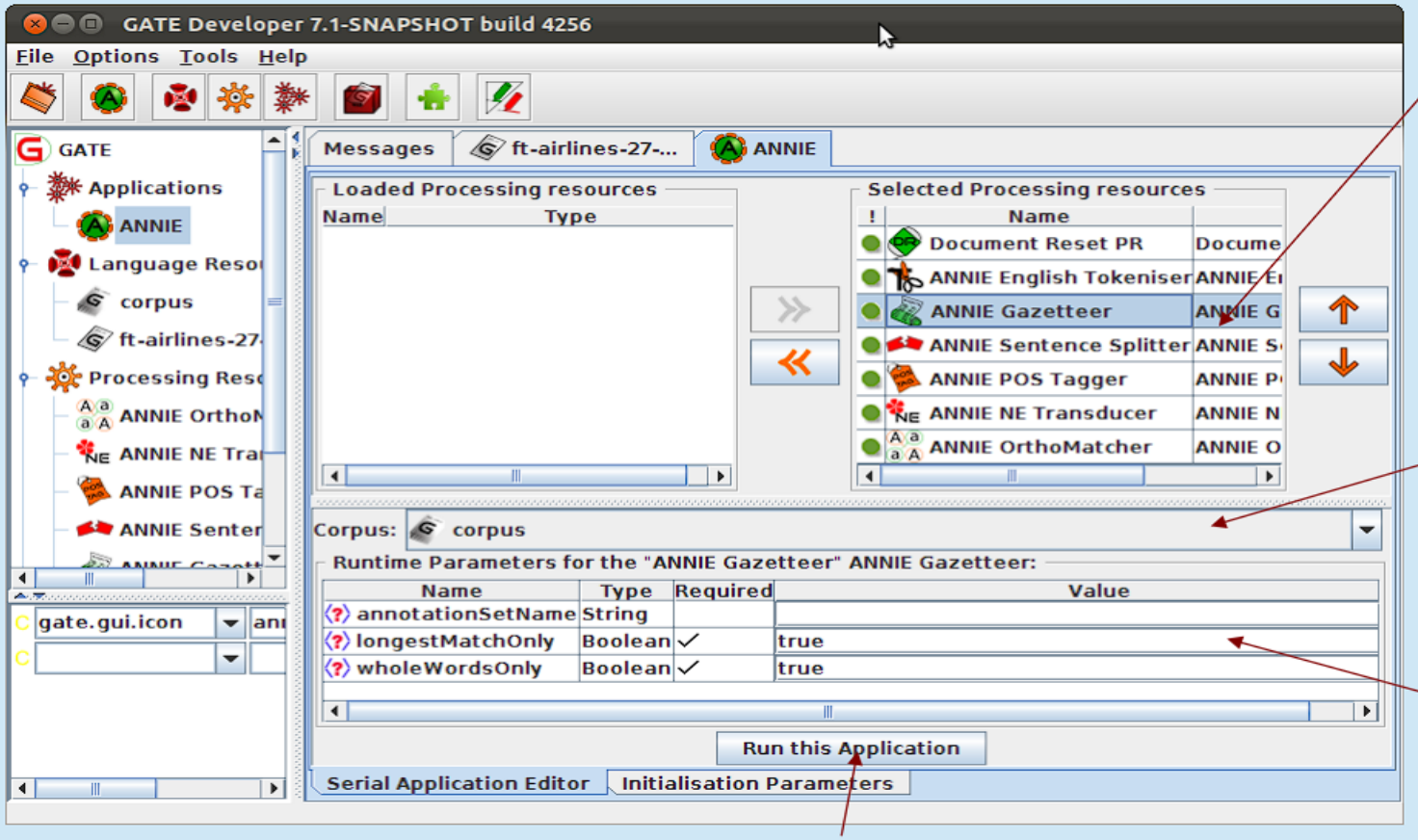
Lab2 : Charger une application existante (ANNIE)

ANNIE est une collection prête à l'emploi de PRs qui effectue l'extraction d'informations sur des textes non structurés.

- Une explication détaillée de ANNIE sera donnée dans la prochaine séance. Pour l'instant, nous allons simplement l'utiliser comme exemple d'application.
- Plus tard, nous vous montrerons comment créer votre propre application à partir de zéro.
- Cliquez sur l'icône dans le menu supérieur de GATE OU sélectionnez File → Load ANNIE system.
- Sélectionnez “with defaults”
- Chargez n'importe quel document de la documentation pratique et ajoutez-le à un corpus.

Lab2 : Charger une application existante (ANNIE)

View the ANNIE application by double clicking on it



PRs selected in application (in order of their execution)

Corpus on which the application is executed

Runtime parameters of the selected PR

Execute the application

Lab3 : Afficher les annotations

Visualisation des résultats

- Lorsqu'un message apparaît dans le coin inférieur gauche de votre fenêtre GATE, du type “ANNIE run in 1.3 seconds”, l'application est terminée.
- Double-cliquez sur le document pour le visualiser
- Affichez les annotations en sélectionnant les ensembles d'annotations et en cliquant sur n'importe quel type d'annotation dans l'ensemble par défaut (sans nom) et aussi dans l’ensemble d’annotations KEY
- L'annotation set KEY spécifique à ANNIE fait référence à l'ensemble d'annotations généré par les modules ANNIE lors du traitement textuel.
- Elle permet de référencer et de manipuler les annotations extraites par les composants ANNIE, facilitant ainsi la gestion et l'exploitation des informations linguistiques extraites à partir du texte.
- Vous pouvez également afficher le tableau des annotations.

Lab3 : Editer les annotations

- Sélectionnez un type d'annotation dans la vue set d'annotations et survolez une annotation en surbrillance dans le texte
- Une fenêtre contextuelle affiche plus d'informations à ce sujet : il s'agit de l'éditeur d'annotations.
- Cliquez sur le symbole de l'épingle à dessin en haut de l'éditeur. La fenêtre est ainsi "épinglée" (vous pouvez toujours déplacer la fenêtre sur votre écran si vous le souhaitez).
- Essayez de modifier l'annotation : vous pouvez changer le type d'annotation, les noms et les valeurs des caractéristiques, la portée de l'annotation (en cliquant sur les flèches gauche et droite en haut de la boîte) ou supprimer l'annotation ou ses caractéristiques
- Fermez l'éditeur d'annotations, puis visualisez l'annotation modifiée dans la liste des annotations.

Lab4 : Ajouter un « annotation set » / Changer les « Runtime parameters »

❑ **Input / Output annotation sets :**

- Quelques PRs utilisent les résultats des PRs précédents dans la chaine de traitement. Par exemple, le PR ‘sentence splitter’ se base sur les ‘Token annotations’ produites par le tokenizer
- Le ‘inputAS’ (annotation set) du PR ‘sentence splitter’ est le nom du set d’annotation ou il va trouver les Token annotations.
- Le ‘OutputAS’ est le nom du set ou il va produire ses annotations
- Dans ANNIE, le ‘inputAS’ est toujours égal au ‘OutputAs’, ca peut être différents dans d’’autres cas
- Certains PRs qui ajoutent plutôt de l’information a une annotation existante plutôt que de créer une nouvelle : les paramètres ‘inputAS’ et ‘outputAS’ sont fusionnés dans un seul paramètre ‘annotationSetName’

Lab4 : Ajouter un « annotation set » / Changer les « Runtime parameters »

Changer les « Runtime parameters » :

❑ **Changer le nom du set d'annotation vers 'ANNIEresult'**

- Double-cliquer sur ANNIE pour visualiser l'application et ses PRs
- Pour chaque PR, cliquer dessus et vérifier s'il a des paramètres : 'inputAS' , 'OutputAS' et 'annotationSetName'
- Editer toutes ces paramètres par la valeur 'ANNIEresult'
- Revérifier que vous n'avez rien oublie, car c'est important pour la réussite de l' exécution de l'application
- Ré-exécuter l'application ANNIE

Lab6 : Ajouter/supprimer une ressource de traitement (PR)

❑ Ajout de nouveaux PR (1)

- Ajoutons un PR '**Verb Phrase Chunker**' à ANNIE.
- Tout d'abord, nous devons charger le plugin qui le contient, puis charger le PR dans GATE, avant de pouvoir l'ajouter à l'application.
- Utilisez le gestionnaire de plugins pour charger le plugin Tools.
- Cliquez avec le bouton droit sur 'Processing Resources' et sélectionnez "New" → "ANNIE VP Chunker" - Laissez tous les paramètres par défaut et cliquez sur "OK".

Lab6 : Ajouter/supprimer une ressource de traitement (PR)

❑ Ajout de nouveaux PR (2)

Nous devons maintenant ajouter le nouveau PR à l'application.

- Double-cliquez sur ANNIE.
- Vous verrez que le **VP chunker** figure dans la liste des PR chargés. Cela signifie qu'il est disponible dans GATE, mais qu'il n'est pas encore contenu dans l'application.
- Ajoutez-le à l'application en le sélectionnant et en utilisant la flèche droite pour le transférer.
- Utilisez maintenant la flèche vers le haut pour le déplacer au bon endroit dans l'application. Il doit être placé après (en dessous) le 'POS tagger' mais avant (au-dessus) le 'NE transducer'.
- Exécutez l'application et visualisez les résultats sur le document.
- Vous devriez voir un nouveau type d'annotation **VG**

Lab6 : Ajouter/supprimer une ressource de traitement (PR)

❑ Supprimer un PR existant

- Double-cliquez sur ANNIE.
- Supprimer les PRs un par un en partant de 'ANNIE OrthoMatcher' → vérifier les changements dans les set d'annotations après chaque suppression
- Recharger tous les PRs de l'application ANNIE puis supprimer le PR 'ANNIE English Tokeniser' → re-exécuter ANNIE. Que se passe t-il? Expliquer?
- Changer l'ordre des PRs en mettant le PS tagger avant le sentence splitter. Que se passe ti-il? Expliquer?

Lab7 : Créer un Datastore

- Sauvegarder les documents
  - En utilisant les datastores
  - Sauvegarder les documents pour utilistaion en dehors de GATE

- 2 types de Datastore:
  - ‘Serial Datastores’ sauvegarde les donnes directement dans un répertoire
  - ‘Lucene Datastores’ fournit un référentiel consultable avec ‘Lucene-based indexing’ (full-featured search engine library written entirely in Java : https://lucene.apache.org/core/)

Lab7 : Créer un Datastore

- Pour l'instant, nous allons nous intéresser aux Datastores en série Créer un nouveau Datastore en série

  - Cliquez avec le bouton droit de la souris sur "Datastores" dans le panneau Ressources et sélectionnez "Create Datastore«
  - Sélectionnez "Serial Datastore".
  - Créez un nouveau répertoire vide en cliquant sur l'icône "Create New Folder" et donnez un nom à votre nouveau répertoire.
  - Sélectionnez ce répertoire et cliquez sur "Ouvrir".
  - Votre Datastore est maintenant prêt à stocker vos documents.