

Dafoe – Présentation de la plate-forme UIMA

Laurent Audibert

Laboratoire d'Informatique de l'université Paris-Nord (UMR CNRS 7030)
Institut Galilée - Université Paris-Nord
99, avenue Jean-Baptiste Clément
93430 Villetaneuse, France

11 juillet 2007

Information non structurée

Information non structurée

Source d'information (texte, audio, image ou vidéo) difficilement exploitable en raison de la nature non structurée du contenu.

- Objectifs d'une application de gestion de l'information non structurée (Unstructured Information Management, UIM) : Organiser un grand volume d'information non structurée pour extraire, structurer et diffuser l'information.
- Technologies mises en œuvre :
OCR, Reconnaissance de la parole, Outils TAL, Apprentissage, Ontologies, etc.

Problème : les technologies développées sont hétérogènes et non interopérables.

Information non structurée

Information non structurée

Source d'information (texte, audio, image ou vidéo) difficilement exploitable en raison de la nature non structurée du contenu.

- Objectifs d'une application de gestion de l'information non structurée (Unstructured Information Management, UIM) : Organiser un grand volume d'information non structurée pour extraire, structurer et diffuser l'information.
- Technologies mises en œuvre :
OCR, Reconnaissance de la parole, Outils TAL, Apprentissage, Ontologies, etc.

Problème : les technologies développées sont hétérogènes et non interopérables.

Plates-formes d'annotation

Objectif

Permettre d'intégrer des outils existants et faciliter leur interopérabilité par des mécanismes d'encapsulation.

- Unstructured Information Management Architecture (**UIMA**) d'IBM – *open source free SDK*
- General Architecture of Text Engineering (**GATE**) – *GNU License*
- Projet **SciBorg** de l'*University of Cambridge*.
- Architecture and Tools for Linguistic Analysis Systems (**ATLAS**) du *National Institute of Standards and Technology (NIST)*.
- Plate-forme d'annotation de documents ALVIS (**Ogmios**) du LIPN – *free software*
- **OpenNLP** – *Open Source projects*

UIMA

- Plate-forme d'intégration et de déploiement d'outils TAL (et autres).
- Kit de développement Java et C++ incluant un ensemble de plug-ins Eclipse.
- Annotations déportées.
- Possibilité d'intégrer/déploier des composants sous forme de services web.
- Pré et post traitements prévus par la plate-forme.
- Documentation importante et existence de tutoriaux.
- Plus puissant et général que des plates-formes comme GATE ou ATLAS.
- Passerelle possible avec GATE.
- Gratuit et open source.

UIMA vs GATE

	UIMA	GATE
Documentation	Bonne	Bonne
Langages de programmation	Java, C++	Java
Outils intégrés	Palette moyenne	Palette très riche
Type de documents	Tout	Texte
Déploiement	Prévu et flexible	Non prévu ?

La plateforme GATE fournit :

- Un adaptateur permettant d'utiliser sous forme de PR (*Processing Resource*) dans GATE un TAE (*Text Analysis Engine*) primitif ou composé d'UIMA ;
- Un adaptateur permettant d'utiliser sous forme de TAE dans UIMA un pipeline de traitement de GATE (*CorpusController*).

Analysis Engine (AE)

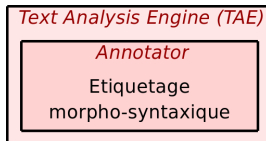
Analysis Engine (AE)

Composant fondamental de traitement.

AE primitif :

- Partie déclarative (spécifications) en XML ;
- *Annotator* : implémentation en Java, en C++ ou sous forme de service Internet.

Appelé TAE (*Text Analysis Engine*) quand il manipule des documents textuels.



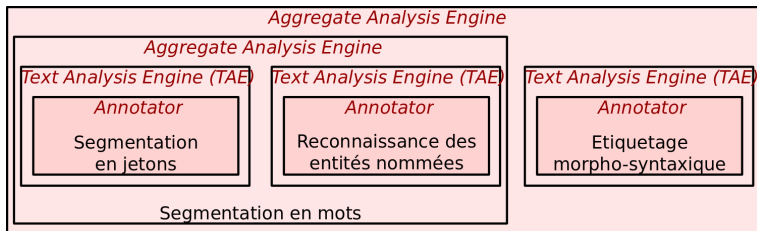
UIMA favorise la réutilisation et l'agrégation de composants en s'appuyant sur leur description XML.

Analysis Engine (AE)

Analysis Engine (AE)

Composant fondamental de traitement.

AE complexe : composition ordonnée d'un ensemble d'AE complexes ou primitifs (*Aggregate Analysis Engine*).



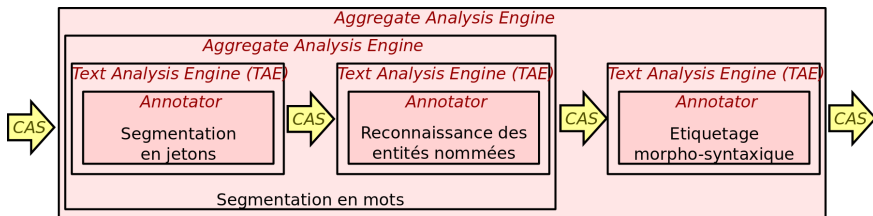
UIMA favorise la réutilisation et l'agrégation de composants en s'appuyant sur leur description XML.

Common Analysis System (CAS)

Common Analysis System (CAS)

Objet commun aux différents composants contenant le document original, ses méta-données (annotations) et une ou plusieurs interfaces pour accéder aux données.

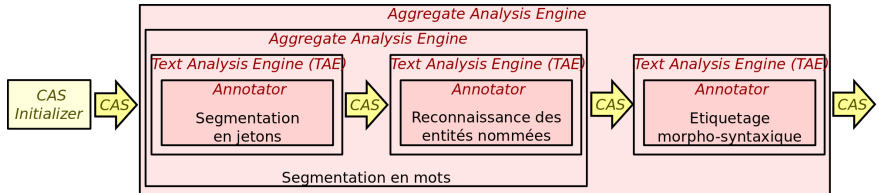
- Un TAE ne fait que compléter un CAS.
- Pour plus de flexibilité, les annotations sont déportées.



CAS Initialiser

CAS Initialiser

Un *CAS Initialiser* est propre à un format de document source et a pour tâche de produire un objet CAS.

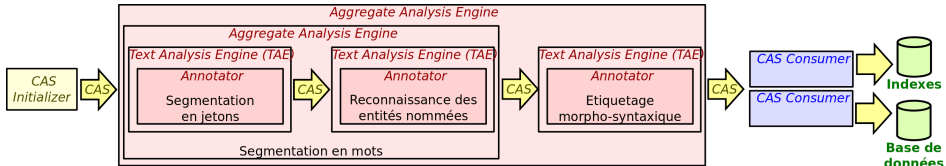


CAS Consumer

CAS Consumer

Intervient à la fin de la chaîne des différents AE pour produire, à partir des CAS, une ressource exploitable par une autre application (indexes, base de données, ...).

- Consomment des CAS mais n'en produisent pas.
- Le rôle peut aller de la simple mémorisation des CAS à des inférences portant sur la totalité des CAS consommés.

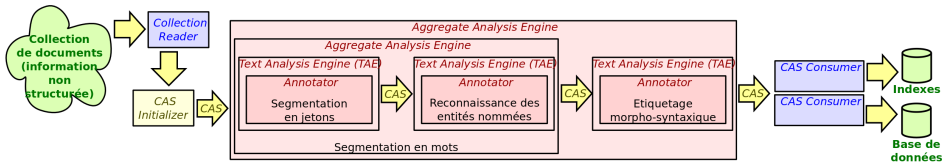


Collection Reader

Collection Reader

Itère sur la collection des documents pour alimenter les CAS *Initialiser*.

- La seule méthode d'un composant *Collection Reader* est « *passer au document suivant* ».

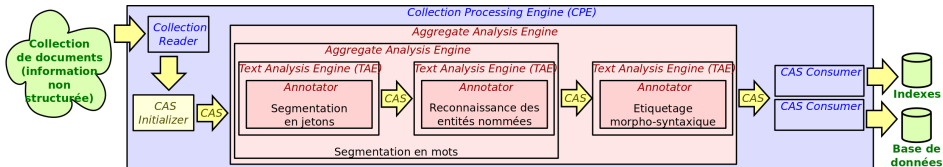


Collection Processing Engine (CPE)

Collection Processing Engine (CPE)

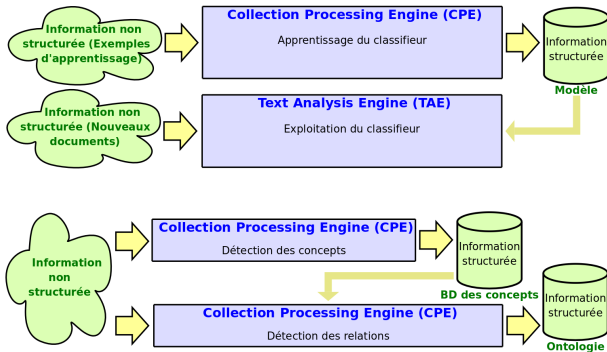
Composant complexe rassemblant tous les composants participant au traitement du *Collection Reader* jusqu'aux *CAS Consumer*.

- Contrôle le flux entre ses différents composants.



Utilisation de ressources d'informations structurées

- UIMA permet à un AE d'accéder à une ressources d'informations structurées.



Collection Processing Management (CPM)

Collection Processing Management (CPM)

Composant permettant de déployer et d'exécuter un CPE dans un environnement UIMA.

Le CPM permet :

- le démarrage, la pause et la reprise des traitements ;
- l'exécution d'un sous-ensemble des TAE en respectant les contraintes inhérentes aux méta-données du CAS ;
- la définition d'une stratégie concernant la gestion des documents provoquant des erreurs ;
- le monitoring des performances (temps, mémoire, ...) ;
- la parallélisation des traitements sur différents documents.

- D. Ferrucci, & A. Lally. (2004). Building an example application with the Unstructured Information Management Architecture. In *Unstructured Information Management* (Vol. 43, p. 455-475). IBM Systems Journal.
- Mustapha Es-salihe, & Stéphane Bond. (2006, 29 mars). *Étude des frameworks UIMA, Gate et OpenNLP*. Internet. (<http://www.crim.ca/fr/R-D/Technologies/Internet/documents/Etude-UIMA-GATE-OpenNLP.pdf>)
- Scott Piao, Sophia Ananiadou, & John McNaught. (2007, septembre). Integrating Annotation Tools into UIMA for Interoperability. In *Sixth UK e-Science All Hands Meeting (AHM2007)*.
- UIMA SDK User's Guide and Reference*. (s. d.). Internet. (http://dl.alphaworks.ibm.com/technologies/uima/UIMA_SDK_Users_Guide_Reference.pdf)