LOISEAU Mathieu
I110
Tél. : 04 76 82 43 93
Mél : mathieu.loiseau@u-grenoble3.fr
http://mathieu.loiseau.free.fr/bdtip

Proposition de projet : Système d'affichage de différentes facettes d'un document textuel.

Nombre de personnes : Ce sujet s'adresse à des groupes de 3 personnes (COPAL ou TALEP, si possible un groupe mixte).

Cadre du projet : Ce projet s'inscrit dans le cadre de mon travail de thèse sur l'indexation pédagogique de textes pour l'enseignement des langues. Si le module développé fonctionne, il sera intégré au prototype final.

Sujet :

Dans le cadre de ma thèse, je dois développer un prototype de base de textes indexée pédagogiquement pour l'enseignement des langues. Cette base de données permettra à des enseignants d'exécuter des requêtes relevant de la problématique de la didactique des langues (ex : textes en anglais pour l'introduction du présent simple). Une fois la requête effectuée, le système renverra un certain nombre de textes candidats. Pour aider l'utilisateur dans son choix entre ces différents candidats, il faudra lui proposer différentes facettes de chaque texte.

Facettes :

Dans le cadre de ce travail, une facette correspond à une représentation (parmi d'autres) d'un texte. Par exemple, pour « I try to sing along, I get it all wrong », une facette mettant en exergue tous les verbes au présent simple serait :

« I **TRY** to sing along, I **GET** it all wrong »

Mais on pourrait vouloir représenter cette facette d'une autre manière, inspirée des concordanciers par exemple :

I **TRY** to sing along
I **GET** it all wrong

Objectif :

Le but de ce projet est de créer un système en ligne de présentation des textes sous plusieurs facettes. À partir d'un texte brut, on veut pouvoir accéder à ses différentes facettes. Le système doit être aussi générique que possible : à terme le système doit intégrer divers outils d'annotation, qui ne proposeront pas nécessairement ni le même format d'annotation, ni les mêmes informations. Le système devra permettre l'analyse de n'importe quel document XML, mais sera testé à partir d'un étiquetage effectué par Tree Tagger transformé en XML.

Développement :

Le développement s'articulera autour de plusieurs modules :

Gestion des outils d'annotation

L'annotation du texte par Tree Tagger doit pouvoir être effectuée en ligne. À terme ce ne sera pas le seul outil d'annotation, la gestion de tous ces outils sera confiée à un module dédié. Ce module devra donc permettre le choix d'un ou plusieurs outils pour l'analyse d'un texte. Dans le cadre de ce projet, une version temporaire de ce module, ne fonctionnant que pour Tree Tagger est acceptable.

Générateur de documents présentables sous différentes facettes

Ce module doit permettre de récupérer un document XML et d'en présenter les différentes facettes. Les traitements pourront être décomposés en plusieurs sous-parties :

Analyse du document

Un module d'analyse des documents, qui va lister les différents contextes existants dans le document. Le contexte correspondra à une position dans l'arborescence du document XML. Ces contextes prendront en compte les attributs des balises XML.

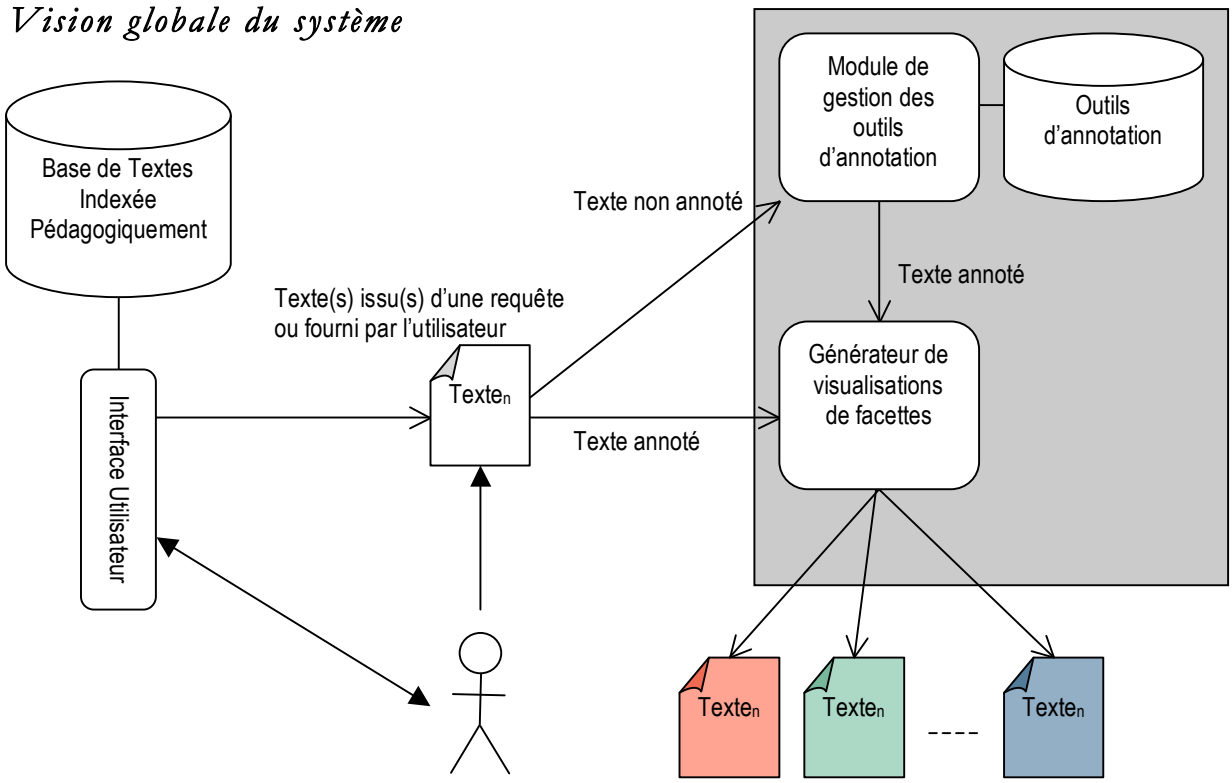
Paramétrage des facettes

Nommer les différentes facettes qui seront rendues disponibles et les associer à des contextes ou choisir parmi des paramétrages existants.

Affichage des facettes

En fonction de la définition des facettes, proposer un document HTML qui permette de visualiser les différentes facettes du document. Ce module pourra proposer plusieurs représentations différentes pour une même facette.

Vision globale du système



Principales difficultés

Extraction des contextes

- Choix d'une stratégie : ouverture du fichier comme du texte / choix + utilisation d'un parser

Paramétrage des facettes

C'est la plus grosse partie du travail. Plusieurs questions se posent :

- Qui définit les facettes ? (utilisateur lambda / concepteur)
- Selon quelle(s) stratégie(s) ? (en partant des contextes / du texte / de notion de didactique des langues et de linguistique)
- Comment associer une facette à un ou plusieurs contexte(s) ? (ER, liste exhaustive)
- Comment implémenter la stratégie choisie ? (interface graphique, lignes de commande, …)

Affichage des facettes

- Choix des moyens techniques : (XML + XSLT, XML + CSS, Javascript, XML + CSS + Javascript …)

Globalement

Le problème majeur est la gestion de la généricité du système. Sans prétendre à une réalisation complètement générique, son architecture devra rendre sa généralisation la plus aisée possible.