1. NAME

TXMreannotate.pl - Packages Lingua::TXMkit ()

2. VERSION

Version 1.0

3. DESCRIPTION

Le développement s'est déroulé sous windows XP, mais un effort particulier a été fait pour essayer d'assurer la portabilité vers d'autres systèmes sans pouvoir asseoir cette garantie, faute de tests sur des machines dédiées.

Ce script est destiné à corriger **partiellement** les erreurs de notation rencontrées dans un étiqueteur morphosyntaxique, en l'occurrence TreeTagger dans le cas présent.

Le réétiquetage se fait à partir d'un fichier XML produit à partir d'une importation dans TXM 0.6.

4. ORGANISATION

Afin de respecter la logique organisationnelle du CPAN, ce package fait/fera partie d'un groupe orienté linguistique "Lingua".

De ce fait, la structure hiérarchique des modules en découle et par exemple le package Linguawrap.pm devra être dans le chemin:

    perl-lib-site-path/Lingua/TXMkit/Linguawrap.pm

etc.

5. INSTALLATION

Ce programme requiert principalement Perl 5.12.3 (ou plus) dont le package MOOSE, le logiciel MongoDB (dans le cas d'utilisation des dictionnaires associés à ce logiciel **http://search.cpan.org/dist/MongoDB/**) et le package FLEMMv3.1.

L'installation usuelle se fait sous $some_path/Perl/site/lib en créant éventuellement le chemin $some_path/Perl/site/lib/Lingua/TXMkit où seront placés les packages (.pm). Le script TXMreannotate.pl s'utilise comme un programme Perl habituel.
En cas de configuration non standard, il conviendra de positionner le chemin dans l'instruction (ligne 28) use lib 'your_path'.

6. SYNOPSIS

L'appel au programme est typiquement le suivant :

où l'option txmtdir spécifie le répertoire source choisi pour faire la réannotation et donc contenant les fichiers XML
produits par TXM 0.6.

Les résultats sont stockés sous ce même répertoire dans MULTEXT pour les fichiers retranscrits avec les annotations de ce type et dans REANNO pour les fichiers seulement corrigés. (Le répertoire TMP contient des fichiers temporaires).

Un autre répertoire cible peut être choisi en utilisant l'option --working_dir.

7. STRUCTURE

7.1 Principes

Le script principal fait appel à plusieurs packages, globalement répartis en quatre groupes fonctionnels. Le premier groupe correspond aux fonctionnalités pour interroger divers dictionnaires locaux au format JSON (LinguaDico.pm). Le second groupe permet l'interrogation de dictionnaires distants, autrement dit sur Internet. Le troisième groupe facilite le traitement des données avec TreeTagger. Un quatrième groupe traite les appels au package FLEMM-v3.1. Disponible au départ sur le site **http://www.cnrtl.fr/outils/flemm/** mais dans lequel nous avons dû faire une mise à jour dans le module Flemm::TreeTagger.

Les requêtes sont le plus possible standardisées sous la forme suivante :

où le tableau associatif $hash contient les informations récupérables, à savoir au moins le lemme et le pos (dans le cas où les réponses sont multiples, on a alors un "array of hash" (tableau de tableaux associatifs).

Le module LinguaWrap.pm lie l'ensemble de ces groupes syntactiquement et contient par ailleurs la méthode consult dans laquelle s'effectue la résolution des tokens indéterminés de TreeTagger.

Le script TXMreannotate gère les appels aux différentes méthodes disponibles pour atteindre le but fixé.

7.2 Script TXMreannotate.pl

Les différentes opérations suivantes sont donc lancées dans ce script :

* Récupération des données d'entrée

* Lecture des fichiers XML et récupération des informations sur TreeTagger

* Récupération des Tokens

* Exécution de TT avec l'option "-proto", ce qui permet l'obtention des données d'étiquetage (f : found, c : found in lowercase, h : half hyphen word, s : suffix, p : pretagging, l : lex, <unknown)

* consultations des dictionnaires sur un nombre limité de données en vertu des résultats précédents, création d'un fichier correctif.

* Réexécution de TT avec l'option "-lex". Mise à jour possible des fichiers XML à ce stade.

* Ajout de l'annotation Multext qui complète les formes verbales en particulier. Réannotation des fichiers XML.

8. PACKAGES/METHODS

8.1 Package Lingua::TXMkit::LinguaDico.pm

Ce package contient les données d'interrogations pour accéder au dictionnaire local Prolex, contenant des noms propres.

Ce fichier téléchargeable à l'adresse :

http://www.cnrtl.fr/lexiques/prolex

comporte les éléments suivants :

Le fichier d'origine en XML a été traduit en JSON afin de permettre son accès par MongoDB.
La commande suivante, pour un système windows, permet de réaliser son importation :

```
c:/mongodb/bin/mongoimport -d dicos -c Prolex --dbpath G:/Dictionnaires/DBmg
G:/Dictionnaires/json/prolex.json
```

si la base de données des dictionnaires se trouve, par exemple, dans le répertoire G:/Dictionnaires/DBmg et que les fichiers d'origine sont dans le répertoire **G:/Dictionnaires/json/**.

Le synopsis d'une requête depuis la création de l'objet se déroule ainsi :

La méthode lookatdico retourne un pointeur de tableau contenant lemmes et pos.

8.2 LinguaDicoDelaf.pm

Ce dictionnaire de noms est téléchargeable sur le site :

http://cental.fltr.ucl.ac.be/projects/delaf/

Le Dictionnaire DELA fléchi (au format XML et encodé en UTF 8) du français comporte :

683 824 entrées simples pour 102 073 lemmes différents

108 436 entrées composées pour 83 604 lemmes différents

Comme précédemment, nous l'avons transcodé au format JSON.

La procédure d'accès, similaire à la précédente, retourne les données selon les cas (le temps, le mode, le genre et le nombre ...).

8.3 LinguaRoleDico.pm

Ce module permet de transformer les pos retournées par les différentes requêtes dans les dictionnaires selon les termes utilisés par TreeTagger.

8.4 LinguaWeb.pm

C'est le module de base pour interroger le site :

http://www.cnrtl.fr/lexicographie

8.5 LinguaWeblx3.pm

C'est le module de base pour interroger le site Lexique3:

http://www.lexique.org/moteur/simple.php?database=lexique3

8.6 LinguaRoleFlemm.pm

Ce module permet l'utilisation de Flemm3 (**http://www.cnrtl.fr/outils/flemm/**) afin d'annoter notamment les formes verbales, qui ne sont pas complètement disponibles dans les dictionnaires.

Pour des raisons d'évolution des logiciels, en l'occurrence de Perl, il est souhaitable de remplacer le module/package TreeTagger.pm d'origine par celui que nous avons modifié, afin d'éviter de nombreux messages d'alerte ("warnings").
A noter que ces modules sont codés en 'cp1252', tandis que le reste du projet l'est en 'UTF8'.

En définitive, nous avons complété la partie Multext (**http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX2_3.html**) afin d'obtenir une solution homogène autant que faire se peut dans le cadre présent.

8.7 LinguaRoleTT.pm

Ce module contient les éléments de récupération des données de TXM 0.6 **`sub get_seqw`** et la réannotation **`sub anno_txm`**. La routine run_treetagger permet de lancer TreeTagger avec les options choisies et de récupérer un pointeur de tableau contenant les résultats de l'exécution.

8.8 LinguaWrap.pm

Ce module permet de fédérer l'ensemble des packages afin de les rendre facilement accessibles à partir d'un programme principal.

De plus, il contient le sous-programme 'consult' qui sert à générer un premier fichier correctif pour TreeTagger. Le principe est simple, car le taux de réussite de TreeTagger, d'environ 95%, donne à penser que l'effort à fournir est faible. Les améliorations du résultat, a priori, dépendront surtout du type de documents étudié, en relation avec un lexique complémentaire adapté.

Ce sous-programme avec les mots sélectionnés, fait juste une recherche dans un dictionnaire de noms

propres (PROLEX) et puis en cas d'échec, interroge le dictionnaire Delaf.

Si un utilisateur souhaite modifier ce programme, celui-ci est facilement compréhensible et accessible.

9. AUTHOR

Charles Minc, **<charles.minc at wanadoo.fr>**

10. BUGS

11. SUPPORT

12. ACKNOWLEDGEMENTS

13. LICENSE AND COPYRIGHT