CasSys
Un système libre de cascades de transducteurs

Denis Maurel Nathalie Friburger
Université François Rabelais Tours
denis.maurel@univ-tours.fr nathalie.friburger@univ-tours.fr

RÉSUMÉ

CasSys est un système de création et de mise en œuvre de cascades de transducteurs intégré à la plateforme Unitex. Nous présentons dans cette démonstration la nouvelle version implantée fin 2012. En particulier ont été ajoutées une interface plus conviviale et la possibilité d'itérer un même transducteur jusqu'à ce qu'il n'ait plus d'influence sur le texte. Un premier exemple concernera le traitement de texte avec une gestion complexe de balises XML et un deuxième présentera la cascade CasEN de reconnaissance des entités nommées.

ABSTRACT

CasSys, a free transducer cascade system.

CasSys is a free toolkit integrated in the Unitex platform to create and use transducer cascades. We are presenting the new version implemented at the end of 2012. The system interface has been improved and the Kleen star operation has been added: this operation allows applying the same transducer until it no longer produces changes in the text. The first example deals with complex XML text parsing and the second with CasEN, a free cascade for French Named Entity Recognition.

MOTS-CLÉS : cascade de transducteurs, graphes Unitex, texte avec balises XML, reconnaissance d'entités nommées.
KEYWORDS : transducer cascade, Unitex graphs, XML text, French Named Entity Recognition.

1 Présentation de CasSys

CasSys est un système de création et de mise en œuvre de cascades de transducteurs (Friburger, Maurel, 2004), aujourd'hui intégré à la plateforme Unitex. Il s'agit donc en fait de cascades de graphes au sens Unitex, plus puissants que de simples transducteurs, puisqu'ils permettent l'utilisation de variables. Une nouvelle version a été implantée en décembre 2012. En particulier une importante fonctionnalité a été ajoutée : la possibilité d'itérer un même transducteur jusqu'à ce qu'il n'ait plus d'influence sur le texte.

Dans cette démonstration, nous proposerons un premier exemple concernant le traitement de texte avec une gestion complexe de balises XML (section 2) et un deuxième présentant la cascade CasEN de reconnaissance des entités nommées (version 1), réalisée en suivant les consignes de la campagne Ester (section 3). Cette cascade est, elle aussi, librement accessible et ses ressources sont ouvertes. La cascade réalisée pour la campagne Etape sera disponible aussi, dès que les résultats officiels seront parus.

2 Traitement du balisage XML

Dans le cadre du projet Région Centre Renom pour la recherche d'entités nommées dans des textes de la Renaissance[1], nous avons dû traiter des textes où l'ensemble de la mise en page était indiquée sous un format XML, rendant difficile l'accès à l'analyse du texte lui-même. Le texte final devait comporter à la fois les balises de mise en page et les balises désignant les entités nommées. L'utilisation d'une cascade permet à des non-informaticiens de faire des manipulations complexes sur le texte sans avoir à coder : par exemple, ignorer des balises lorsqu'elles ne sont pas nécessaires à l'analyse (lettrines, début de ligne…), rétablir les mots coupés par un saut de ligne ou bas de page, choisir la forme corrigée et non la forme originale lorsque des corrections sont ajoutées (en général, des apostrophes absentes du texte original). Par exemple :

Coupure en fin de ligne	 <hi rend="larger">E</hi>Nceste mesme heure Gargan <lb rend="hyphen"/>tua [...] <lb/> fut adverty [...] comment Picrocho- <lb rend="hyphen"/>le sestoit rempare a la Rocheclermaud [...] 	Pour la REN, on veut disposer de : Gargantua Picrochole la Rocheclermaud
Ajout d'apostrophe	[...] <lb/>[...] sainct <lb/>Thomas <choice><orig>Langloys</orig><reg>L'angloys</reg></choice> voulut bien pour <lb/>yceulx mourir, [...] 	et sainct Thomas L'angloys

3 La cascade CasEN, version 1

La cascade CasEN, réalisée pour la campagne Ester, dans le cadre le cadre du projet ANR Variling et du projet FEDER Région Centre Entités nommées et nommables, est disponible librement et en ressources ouvertes[2]. Elle permet la reconnaissance d'entités nommées. Cette cascade, décrite dans (Maurel et al., 2011), est composée de 56 graphes et sera commentée lors de la démonstration (en particulier l'ordre des graphes qui n'est pas anodin !) . Un exemple de reconnaissance est donné ci-dessous :

« Au pire de la crise, <ENT type="time.date.rel">à l'automne dernier</ENT>, nous avons détenu jusqu'à 20 % de liquidités dans notre portefeuille », indique <ENT type="pers.hum"><ENT type="pers.hum"><forename>Denis </forename> <surname>Remacle</surname>, <ENT type="job">gérant d'<ENT type="org.com">Amplitude Pacifique</ENT></ENT></ENT>, une sicav de <ENT type="org.com">La Poste</ENT>.

Références

FRIBURGER N., MAUREL D. (2004), Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1):69-96.

[1] Ici, *Gargantua* de Rabelais, dans sa version originale
[2] http://tln.li.univ-tours.fr/Tln_CasEN.html