Démonstration de l'API de NLGbAse

François-Xavier Desmarais, Éric Charton

École Polytechnique de Montréal, 2900 boul. Edouard Montpetit, Montréal, Canada H3T 1J4
{francois-xavier.desmarais, eric.charton}@polymtl.ca

Description

Le système NLGbAse transforme des contenus encyclopédiques en métadonnées. Il utilise ensuite ces métadonnées pour entrainer et faire fonctionner des systèmes d'étiquetage de textes [1]. NLGbAse permet d'étiqueter les entités nommées (EN) d'un texte en reprenant la taxonomie ESTER [2]. Il peut ensuite établir un lien sémantique entre l'EN identifiée et sa représentation sur le web sémantique, notamment son point d'entrée dans le réseau LinkedData [3]. Le système NLGbAse est multilingue et peut étiqueter un texte en français, anglais ou espagnol. Il peut être utilisé en utilisant soit une interface en ligne, soit un API. Notre démonstration consiste à présenter ses fonctionnalités.

Dans un texte, chaque entité nommée est associée à son étiquette de classe et à des liens. Les liens relient l'EN avec sa page descriptive de Wikipedia *(*libellé *wp)* et à un point d'entrée sur le réseau LinkedData, au format « Resource Description Framework » (libellé *rdf*).

L'utilisation de l'interface en ligne n'est appropriée que pour l'étiquetage de courts textes et n'autorise qu'une étude rudimentaire. Pour l'obtention d'un corpus plus large et étiqueté, une API est disponible. Cette API permet d'envoyer une séquence de texte depuis un programme PERL et de recevoir en retour une sortie étiquetée par NLGbAse, fournie sous la forme de lignes composées de cinq colonnes, séparées par des tabulations.

Le mot (ou ponctuation), la nature du mot (POS), son étiquette taxonomique (EN), son lien avec sa métadonnées représentative de NLGbAse et un lien vers Dbpedia (qui permet de collecter tous les points d'entrées disponibles pour un terme sur le réseau LinkedData).

L'API de NLGbAse est disponible gratuitement pour étiqueter, par période de 24 heures, un maximum de 100 documents d'au plus 10000 caractères chacun. Des volumes plus importants peuvent être alloués sur demande pour des recherches académiques.

Bibliographie :

[1] Charton, E. & Torres-Moreno, J. (2010). NLGbAse: a free linguistic resource for Natural Language Processing systems. (Eds.)*English*, (1), 2621-2625. Proceedings of LREC 2010.

[2] Charton, E. & Torres-Moreno, J. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. Dans Taln 2009 , volume 1, pages 24–26. TALN.

[3] Charton, E., Gagnon, M., & Ozell, B. (2010). Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique. *Actes de TALN 2010*, *1*(1), 19-23.