

Vers une instance française de NELL : chaîne TLN multilingue et modélisation d'ontologie

Maisa Cristina Duarte*, Pierre Maret*

*Univ. Lyon, UJM-Saint-Etienne, CNRS
Laboratoire Hubert Curien UMR 5516
F-42023 Saint-Étienne, France

maisa.cristina.duarte@univ-st-etienne.fr, pierre.maret@univ-st-etienne.fr

Résumé. Nous présentons les étapes de préparation de la création d'une instance nouvelle de NELL dédiée au français. NELL est à la fois un processus de lecture et de compréhension automatique du Web et un ensemble de base de connaissances de faits en anglais, en portugais et très prochainement en français. Cette mise en place de la nouvelle instance de NELL a donné lieu à l'amélioration de la chaîne NLP en la généralisant au multilingue, ainsi qu'au développement d'une ontologie par correspondance avec l'ontologie en anglais. Nous présenterons le processus de mise en place et de lancement de la nouvelle instance NELL Français avec l'interface de visualisation et de supervision humaine des données collectées.

1 Introduction

La lecture par machine (Machine Reading, MR) est un domaine de recherche s'intéressant à la compréhension du langage naturel (Natural Language Understanding) et qui cherche à aller au delà du traitement du langage naturel. Selon les principes de (Etzioni et al., 2006) le but principal du MR est la *compréhension autonome du texte*. Considérant que la principale méthode d'apprentissage de l'humain passe par la lecture, divers projets de recherche sont dédiés à la conception de systèmes capables d'apprendre en lisant (Clark et al., 2007).

L'approche par apprentissage permanent (Never-Ending Learning, NEL) est une technique utilisée dans des systèmes de MR. Dans ce paradigme, l'apprenant évolue de façon autonome et permanente dans le temps, et surtout il apprend progressivement pour améliorer ses performances. Le premier système d'apprentissage permanent décrit dans la littérature est le système NELL, Never-Ending Language Learner (Carlson et al., 2010). NELL a démarré en janvier 2010 et lit et relit le web en anglais dans le but de collecter des faits et de peupler sa base de connaissances (KB). Une deuxième instance de NELL lit le Portugais (Duarte et Hruschka, 2014). Nous présentons dans cet article nos travaux pour la création de NELL Français et proposerons en démonstration de présenter le processus de mise en place et de visualisation des éléments collectés.