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 portuguais 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 multilangue, 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 language 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 système 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 performences. Le premier système d'apprentissage permanent décrit dans la literature 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.

2 NELL, Never-Ending Language Learning

NELL est un système qui capitalise sur ses apprentissages pour apprendre de mieux en mieux chaque jour. Il implémente une ontologie initiale en entrée, composée de catégories et de relations, et il produit en sortie une base de connaissances qui augmente et s'améliore chaque jour.

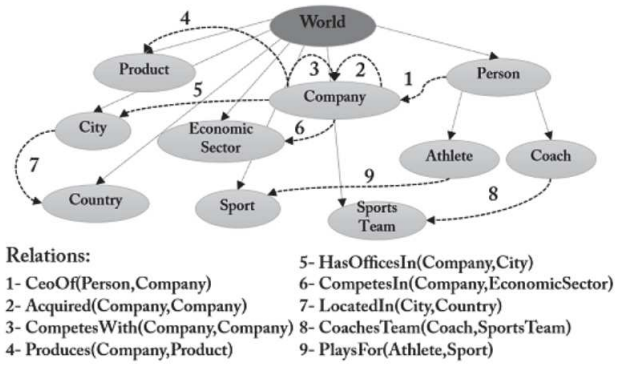

FIG. 1 – *Un extrait de l'ontologie de NELL, tiré de (Duarte et Hruschka, 2014)*

La figure 1 présente un exemple illustrant l'ontologie de NELL. Les connaissances sont décrites en termes de catégories et de relations. L'ontologie utilisée possède 276 catégories et 319 relations. A titre d'exemple, City, Person, Country, Company, etc. sont des catégories ; et CeoOf(Person, Company), LocatedIn(City, Country), etc. sont des relations. Les données qui peuplent les catégories sont par exemple : City(Saint-Etienne), Person(Barack Obama), etc. ; et les relations sont LocatedIn(Paris,France), CeoOf(Sundar Pichai, Google), etc.

NELL utilise divers composants (ou sous-systèmes) tels que CPL (Coupled Pattern Learning), SEAL (Coupled SEAL), PRA (Path Ranking Algorithm), Human Advice et ConceptResolver, etc. qui sont décrits dans (Mitchell et al., 2015).

Le processus général de NELL consiste à réaliser des itérations qui exécutent en séquence : lire le web, extraire les connaissances, et calculer la confiance pour éventuellement introduire de nouvelles connaissances dans l'ontologie. Dans le but d'optimiser le processus de lecture et d'extraction de connaissances, le système d'apprentissage de NELL utilise une base de motifs pré-calculés appelée *all-pairs-data*. Cette base des *all-pairs-data* est générée par une chaîne TLN prenant en entrée le corpus Clueweb (Callan et al., 2009).

3 Développement d'une chaîne TLN multilingue et correspondances d'ontologies anglais-français

Le développement d'une chaîne TLN multilingue et la mise en place des correspondances d'ontologies sont deux étapes indépendantes nécessaires à la création de la nouvelle instance

de NELL. Alors que les instances précédentes de NELL utilisaient une chaîne TLN dédié au language ciblé (anglais, portuguais), nous avons implémenté une nouvelle chaîne TLN qui est indépendante du language et qui peut être utilisée pour toutes les langues. Nous l'avons utilisée ensuite pour le français. Actuellement, la chaîne s'exécute pour créer la base des *all-pairs-data* en français en lisant le corpus Clueweb avec environ 50 million pages Web en français. Cette chaîne NLP comprend les étapes suivantes : 1) Lire une phrase dans Clueweb, introduire des balises, sauver le texte balisé ; et 2) Lire le texte balisé et extraire les occurrences et les co-occurences de catégories et de relations. Pour l'étape 1, nous avons intégré Wikifier (Carreras et al., 2014) qui permet d'une part de baliser les pages web et d'autre part d'identifier les entités nommées. Wikifier peut être utilisé pour 17 langues. Pour l'étape 2, nous avons adopté une implémentation Hadoop afin de calculer les occurrences et co-occurrences d'entités nommées et de motifs textuels, pour les catégories comme pour les relations. Ce processus mène à la création des *all-pairs-data*.

Parallèlement, nous avons réalisé une mise en correspondance d'ontologies entre l'Anglais et le Français. Il en résulte une ontologie en Français qui suit l'ontologie en Anglais [1], et qui est composée de catégories, de relations ainsi que d'exemples de ces éléments. L'ontologie en Français a été introduite dans le système NELL pour initier la nouvelle base de connaissances NELL Français (figure 2) et elle est utilisée avec les *all-pairs-data* pour le processus d'apprentissage du système.

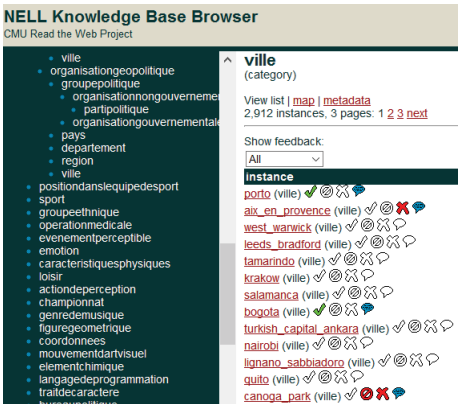

FIG. 2 – *Ontologie en Français*

La figure 2 présente la visualisation l'ontologie NELL Français qui sert aussi d'interface de supervision humaine de cette ontologie. La supervision humaine consiste à ce qu'une personne vérifie les connaissances apprises par le système et fasse un retour afin d'assister le processus d'apprentissage. Les valeurs possibles des retours sont : correct, faux, contre-exemple.

Le système est totalement configuré. Il sera présenté en même temps que le processus mené pour le mettre en place. Il reste à le mettre en production, dès que l'étape de création des *all-pairs-data* avec Clueweb sera terminée. Nous serons alors à même de lancer l'apprentissage et la supervision sur la base de connaissance NELL Français.

1. http ://rtw.ml.cmu.edu/rtw/kbbrowser

4 Conclusion

Dans cet article, nous avons présenté les étapes menées pour la création d'une nouvelle instance de NELL, en français. Nous pourrons présenter ces étapes ainsi que le procéssus général de production des connaissances NELL : textes sources, balisage, *all-pairs-data*, catégories et relations en français, visualisation et supervision humaine. Les impacts de ce travail sont nombreux pour l'extraction ininterrompue de connaissances en français sur le Web et leur exploitation.

Références

Callan, J., M. Hoy, C. Yoo, et L. Zhao (2009). Clueweb09 data set.

Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., et T. M. Mitchell (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Carreras, X., L. Padró, L. Zhang, A. Rettinger, Z. Li, E. García Cuesta, Z. Agic, B. Bekavac, B. Fortuna, et T. Stajner (2014). Xlike project language analysis services. In *EACL 2014 : 14th Conference of the European Chapter of the Association for Computational Linguistics : Gothenburg, Sweden : April, 26-30, 2014 : proceedings of the conference*, pp. 9–12. Association for Computational Linguistics.

Clark, P., P. Harrison, J. A. Thompson, R. Wojcik, T. Jenkins, et D. J. Israel (2007). Reading to learn : An investigation into language understanding. In *AAAI Spring Symposium : Machine Reading*, pp. 29–35.

Duarte, M. C. et E. R. Hruschka (2014). How to read the web in portuguese using the never-ending language learner's principles. In *14th International Conference on Intelligent Systems Design and Applications*, pp. 162–167.

Etzioni, O., M. Banko, et M. J. Cafarella (2006). Machine reading.

Mitchell, T., W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. D. Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, et J. Welling (2015). Never-ending learning.

Summary

We present and will demonstrate the steps for the creation of a new instance of NELL: NELL French. We have improved the NLP pipeline to make it multilingual, and we have mapped the NELL ontology to French. We will present the whole initiation process and show how the production process works, leading to the user interface to visualize and supervise the knowledge base.