Nouveautés de l'analyseur linguistique LIMA

Gaël de Chalendar[1]
(1) CEA,LIST, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, F-91191
gael.de-chalendar@cea.fr

RÉSUMÉ

LIMA est un analyseur linguistique libre d'envergure industrielle. Nous présentons ici ses évolutions depuis la dernière publication en 2014.

ABSTRACT

What's New in the LIMA Language Analyzer.

LIMA is a free language analyzer of industrial scope. We present here its evolutions since the last publication in 2014.

MOTS-CLÉS : tokenisation, morphologie, étiquetage morphosyntaxique, analyse syntaxique, entités, relations, interface graphique.

KEYWORDS: tokenization, morphology, PoS tagging, parsing, entities, relations, GUI.

LIMA est l'analyseur linguistique du CEA LIST. Commencé en 2002, il a été présenté à la communauté en 2010 (Besançon *et al.*, 2010) puis placé sous licence libre en 2014 (de Chalendar, 2014) [1].

Depuis la dernière version numérotée 2.1 en 2015, plus de 1200 modifications ont été apportées. La plupart ne sont que des améliorations à la marge, corrections de bugs ou améliorations de l'infrastructure. Nous présentons dans les sections suivantes les changements les plus importants, mais commençons par résumer ci-dessous quelques autres évolutions.

De nouveaux tests unitaires ont été ajoutés. Le système d'intégration continue (IC) a été amélioré avec l'utilisation de conteneurs docker sur les plateformes Semaphore, Appveyor et Travis. Nous utilisation désormais le système des release github pour distribuer les paquets générés par l'IC. Nous avons aussi amélioré la construction multiplateforme en utilisant le système de construction Ninja sur l'ensemble d'entre elles. Concernant les aspects TAL, nous avons débuté la transition vers l'utilisation d'étiquettes issues du projet Universal Dependencies. Enfin, nous utilisons désormais SVMTool comme étiqueteur morphosyntaxique par défaut.

Support du portugais

Nous avons ajouté à LIMA le support de la langue portugaise en utilisant le dictionnairs Delaf PB (sous licence LGPL) et le corpus annoté Mac-Morpho (sous licence CC BY 4.0). Nos derniers résultats d'évaluation de l'étiquetage morphosyntaxique par validation croisée sur le corpus d'apprentissage donnent une précision de 96%, du niveau de l'état de l'art. Il nous reste désormais à ajouter une prise en compte des expressions idiomatiques, le traitement des entités nommées et des règles d'analyse syntaxique pour avoir les mêmes fonctionnalités que dans les autres langues.

1. https://github.com/aymara/lima

Entités nommées améliorées

Nous avons amélioré nos traitements des entités nommées selon deux axes. D'une part, nous avons ajouté un module permettant d'effectuer une recherche approximative, ce qui permet de repérer des noms malgré des formes variables. Quand un dictionnaire de référence existe, on admet des erreurs (suppression ou ajout de caractères) et on utilise des motifs de généralisation. On peut alors trouver les noms du dictionnaire avec une marge d'erreur spécifiée. D'autre part, nous avons ajouté un module de reconnaissance statistique des entités à base de CRF, fondé sur la bibliothèque Wapiti. Celui-ci nous a permis d'obtenir d'excellents résultats, sur certains types d'entités, dès lors qu'un corpus annoté est disponible.

Interface graphique

Notre ambition est de rendre LIMA accessible à tous, aussi bien des industiels désirant intégrer des traitements de TAL dans leurs applications que des étudiants en linguistique devant aborder le TAL. Pour ces derniers et tout autre utilisateur occasionnel, nous avons développé une interface graphique permettant d'accéder aux principales fonctionnalités de LIMA. Celle-ci ne permet pour le moment que de lancer une analyse et de consulter les résultats sous divers formats. Elle sera enrichie à l'avenir avec de nouveaux outils de visualsation et une interface de configuration.

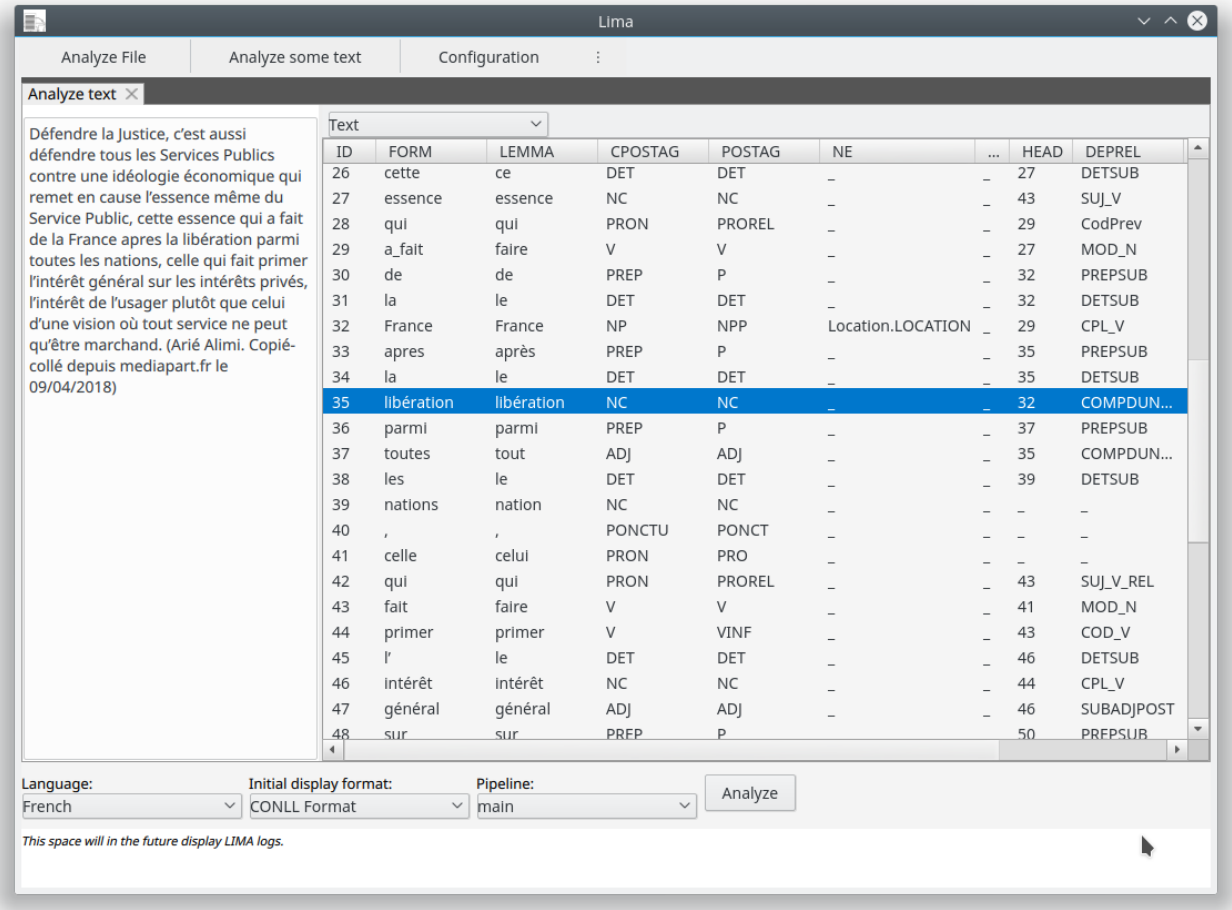

FIGURE 1 – L'interface graphique de LIMA

Conclusion

LIMA ne reste pas en marge de la vague de fond qui révolutionne le domaine du TAL, les approches neuronales. Déjà, un module neuronal d'extraction d'entités nommées a été ajouté et un module d'analyse syntaxique est en cours de finalisation de son intégration. Au delà de ces deux modules, le travail sur l'infrastructure nécessaire pour générer de bout en bout des analyseurs à partir de corpus annotés est en cours. Cela nous permettra de participer aux prochaines occurrences de la tâche partagée CoNLL. Pour autant, nous voulons conserver les fonctionnalités qui font la spécificité de LIMA : multiplateforme, performance, facilité d'usage et intégrabilité dans des outils industiels.

Références

BESANÇON R., CHALENDAR (DE) G., FERRET O., GARA F. & SEMMAR N. (2010). LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. In *Proceedings of Language Resources and Evaluation Conference, 2010*, Malta.

DE CHALENDAR G. (2014). The LIMA Multilingual Analyzer Made Free : FLOSS Resources Adaptation and Correction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, p. 2932–2937.