Laboratoire d'Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact: Ga€l de Chalendar
Gael.de-chalendar@cea.fr
+33 (0)1 69 08 01 50

STAGE 2021 Réf : LVIC-2021-XX

Modèles Deep Learning de Traitement Automatique des Langues pour l'Ingénierie Logicielle

Présentation du laboratoire d'accueil

Basé à Paris-Saclay, le CEA List, membre de l'Université Paris Saclay, est l'un des quatre instituts de recherche technologique de CEA Tech, direction de la recherche technologique du CEA. Dédié aux systèmes numériques intelligents, il contribue au développement de la compétitivité des entreprises par le développement et le transfert de technologies.

L'expertise et les compétences développées par les 800 ingénieurs-chercheurs et techniciens du CEA List permettent à l'Institut d'accompagner chaque année plus de 200 entreprises françaises et étrangères sur des projets de recherche appliquée s'appuyant sur 4 programmes et 9 plateformes technologiques. 21 start-ups ont été créées depuis 2003. Labellisé Institut Carnot depuis 2006, le CEA List est aujourd'hui l'institut Carnot Technologies Numériques.

Le Laboratoire d'Analyse Sémantique des Textes et des Images (LASTI) est une équipe de 25 personnes (chercheurs, ingénieurs, doctorants) menant des travaux de recherche sur les technologies de description et de compréhension du contenu multimédia (image, texte, parole) et des documents multilingues, en particulier à grande échelle. Les enjeux scientifiques sont :

- développer des algorithmes efficaces et robustes pour l'analyse et l'extraction de contenu multimédia, leur classification et analyse sémantique ;
- reconstitution ou fusion de données hétérogènes pour interpréter des scènes ou documents ;
- développer des méthodes et des outils pour la construction, la formalisation et l'organisation des ressources et connaissances nécessaires au fonctionnement de ces algorithmes ;
- intégrer plusieurs de ces briques technologiques afin d'accéder à l'information et répondre à un besoin utilisateur (moteurs de recherche, agents conversationnels, rapports synthétiques de veille)

Description du stage

Le laboratoire LASTI participe au projet Européen H2020 Decoder visant entre autres à exploiter les technologies du traitement automatique des langues dans le cadre de l'ingénierie logicielle. En effet, l'information textuelle est partout dans ce cadre : exigences, spécifications, commentaires du code, documentations utilisateur, forums (stackoverflow…), gestionnaires de tickets, etc. De plus la quantité de texte et de code correspondant disponibles en ligne permettent d'utiliser efficacement les techniques d'apprentissage automatique. Les applications peuvent aller de la simple extraction d'information (entités nommées, semantic role labeling…) pour mettre des éléments en évidence dans les interfaces utilisateurs, jusqu'à la conversion automatique de texte en code source en utilisant des techniques issues de la traduction automatique, en passant par l'aide à la traçabilité pour repérer par exemple des commentaires ou du code qui violeraient des exigences.

Nous avons jusqu'à présent reproduit un certain nombre de modèles et collecté des données. Nous avons aussi développé les outils logiciels permettant de mettre nos résultats à la disposition des partenaires du projet. Nous avons enfin spécifié un certain nombre d'améliorations que nous comptons apporter aux modèles pour aller au-delà des résultats de l'état de l'art. Le travail du ou de la stagiaire consistera à participer à l'implémentation de ces améliorations et à leur évaluation sur des données génériques permettant la comparaison avec l'état de l'art ainsi que sur les données du projet. Les résultats seront soumis pour publication dans des conférences internationales. Les modèles sont implémentés en python avec les frameworks de deep learning PyTorch et TensorFlow.

Les modèles concernés sont ceux de [Strubell et al., 2018] pour le semantic role labeling ; [Thang et al., 2015] et [Iver et al., 2018] pour le semantic parsing (traduction de spécifications de haut niveau en

Laboratoire d'Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact: Gaël de Chalendar
Gael.de-chalendar@cea.fr
+33 (0)1 69 08 01 50

spécifications formelles ou en code) ; [Guo et al., 2017], [Narayanan, 2019] et [Seki, 2018, 2019] pour la traçabilité horizontale et verticale. Le ou la stagiaire pourra être amené.e à travailler sur plusieurs de ces modèles en fonction de l'avancement des travaux lors de son arrivée et de ses progrès pendant son stage.

Le ou la stagiaire utilisera les clusters de calcul du laboratoire. Ceux-ci incluent plusieurs dizaines de nœuds GPU régulièrement mis à jour. Si jamais le confinement devait se poursuivre, il ou elle aura accès au réseau CEA et aux clusters par VPN, permettant une poursuite du stage dans les meilleures conditions possibles.

Mots-clés :

Traitement automatique des langues, deep learning, ingénierie logicielle.

Références

[Guo et al., 2017] Guo, J., J. Cheng, et J. Cleland-Huang. « Semantically Enhanced Software Traceability Using Deep Learning Techniques ». In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 3-14, 2017.

[Iver et al., 2018] Iyer, Srinivasan, Ioannis Konstas, Alvin Cheung, et Luke Zettlemoyer. « Mapping Language to Code in Programmatic Context ». In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1643–1652. Brussels, Belgium: Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/D18-1192.

[Narayanan, 2019] Narayanan, Siddharth. « Semantic Similarity in Sentences and BERT ». Medium, 27 septembre 2019. https://medium.com/analytics-vidhya/semantic-similarity-in-sentences-and-bert-e8d34f5a4677. (Last accessed, 07/08/2020).

[Seki, 2018] Seki, Kazuhiro. « Exploring Neural Translation Models for Cross-Lingual Text Similarity ». In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1591–1594. CIKM '18. Torino, Italy: Association for Computing Machinery, 2018. https://doi.org/10.1145/3269206.3269262.

[Seki, 2019] Seki, Kazuhiro. « On Cross-Lingual Text Similarity Using Neural Translation Models ». *Journal of Information Processing* 27, n° 0 (2019): 315-21. https://doi.org/10.2197/ipsjjip.27.315.

[Strubell et al., 2018] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-Informed Self-Attention for Semantic Role Labeling. Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, Belgium. October 2018.

[Thang et al., 2015] Luong, Thang, Hieu Pham, et Christopher D. Manning. « Effective Approaches to Attention-based Neural Machine Translation ». In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics, 2015. https://doi.org/10.18653/v1/D15-1166.

Profil du candidat/de la candidate

Niveau demandé :	Ingénieur, Master 2
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.
Compétences requises : - Natural Language processing - Deep Learning - Python	

Laboratoire d'Analyse Sémantique Texte et Image
CEA Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact: Ga€l de Chalendar
Gael.de-chalendar@cea.fr
+33 (0)1 69 08 01 50

- Good proficiency in Engli
- sh