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

STAGE 2021 Réf : LVIC-2021-XX

Deep Learning Models of Automatic Language Processing for Software Engineering

Presentation of the host laboratory

Based in Paris-Saclay, CEA List, a member of Paris Saclay University, is one of the four technological research institutes of CEA Tech, the technological research department of CEA. Dedicated to intelligent digital systems, it contributes to the development of business competitiveness through technology development and transfer.

The expertise and skills developed by the 800 research engineers and technicians of CEA List enable the Institute to support more than 200 French and foreign companies each year on applied research projects based on 4 programs and 9 technological platforms. 21 start-ups have been created since 2003. Labeled Institut Carnot since 2006, CEA List is today the Carnot Digital Technologies Institute.

The Laboratory of Semantic Analysis of Texts and Images (LASTI) is a team of 25 people (researchers, engineers, PhD students) conducting research on technologies for describing and understanding multimedia content (image, text, speech) and multilingual documents, particularly on a large scale. The scientific stakes are:

- to develop efficient and robust algorithms for the analysis and extraction of multimedia content, their classification and semantic analysis;
- reconstructing or merging heterogeneous data to interpret scenes or documents;
- to develop methods and tools for the construction, formalization and organization of the resources and knowledge necessary for the operation of these algorithms;
- Integrate several of these technological building blocks in order to access information and meet a user need (search engines, conversational agents, synthetic watch reports).

Description of the internship

The LASTI laboratory participates in the H2020 European project Decoder aiming among other things to exploit natural language processing technologies in the framework of software engineering. Indeed, textual information is everywhere in this framework: requirements, specifications, code comments, user documentations, forums (stackoverflow…), ticket managers, etc. In addition, the amount of text and corresponding code available online allows the efficient use of machine learning techniques. Applications can range from simple extraction of information (named entities, semantic role labeling…) to highlight elements in user interfaces, to the automatic conversion of text into source code using techniques derived from machine translation, to traceability assistance to spot, for example, comments or code that violate requirements.

To date, we have replicated a number of models and collected data. We have also developed the software tools to make our results available to project partners. Finally, we have specified a number of improvements that we intend to make to the models to go beyond state of the art results. The work of the intern will consist of participating in the implementation of these improvements and their evaluation on generic data allowing comparison with the state of the art as well as on project data. The results will be submitted for publication in international conferences. The models are implemented in python with deep learning frameworks PyTorch and TensorFlow.

The models concerned are those of [Strubell et al., 2018] for semantic role labeling; [Thang et al., 2015] and [Iver et al., 2018] for semantic parsing (translation of high-level specifications into formal specifications or code); [Guo et al., 2017], [Narayanan, 2019] and [Seki, 2018, 2019] for horizontal and vertical traceability. The trainee may be required to work on several of these models depending on the progress of the work upon his/her arrival and the progress made during the internship.

The trainee will use the laboratory's computing clusters. These include several dozen GPU nodes that are regularly updated. Should the confinement period continue, he or she will have access to the CEA network and the clusters via VPN, enabling the internship to continue under the best possible conditions.

Keywords :
Natural language processing, deep learning, software engineering.

References

[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^{o} 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.

Candidate Profile

Required degree:	Engineer, Masters Degree
This internship must be mandatory to validate a degree and made with an agreement with a university	
Duration:	6 months
Remuneration:	between 700 € and 1300 € depending on the formation
Required skills: - Natural Language processing - Deep Learning - Python - Good proficiency in English	

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