(English version below)

Offre de thèse CIFRE :
Apprentissage profond pour l'extraction d'entités et de relations dans le domaine scientifique

Laboratoire :
Le/la doctorant.e intégrera l'équipe de Représentations des Connaissances et Langage Naturel (RCLN; https://lipn.univ-paris13.fr/accueil/equipe/rcln/) du Laboratoire d'Informatique de Paris Nord (LIPN), UMR CNRS 7030 attaché à l'Université Sorbonne Paris Nord. L'équipe RCLN est membre du laboratoire d'excellence EFL (Empirical Foundations of Linguistics; http://www.labex-efl.com).

Société :
Un groupe international de conseil dans le domaine de la Recherche, du Développement et de l'Innovation, intervenant sur les aspects organisationnels, structurels, méthodologiques, scientifiques et financiers. La société conseille les entreprises les plus à la pointe dans leurs secteurs, celles qui innovent et fondent leur avance stratégique sur des travaux de recherche expérimentaux et fondamentaux, avec une expérience de plus de 20 ans et plusieurs milliers de collaborations à travers le monde, sur tous les continents. Au sein du groupe, la Direction Scientifique a pour mission de coordonner l'ensemble des actions scientifiques, tant sur les plans opérationnels, que méthodologiques et conceptuels. Elle s'appuie, en France, sur les ressources de plus de 60 docteurs de toutes disciplines. Au cœur de cette Direction Scientifique, le doctorant intégrera l'équipe du Research Lab, le département R&D interne du groupe basé à la Défense à Paris.

Sujet
Le/la doctorant.e travaillera sur la conception et la mise en oeuvre d'un système d'extraction jointe d'entités et de relations sémantiques à partir de textes écrits par des experts dans des différents domaines techniques (plus de détails ci-dessous).
https://lipn.univ-paris13.fr/~tomeh/public/uploads/offers/phd-cifre-relation-extraction.pdf

Mots clés
- Traitement automatique des langues (natural language processing) ;
- Apprentissage profond (deep learning) ;
- Extraction d'entités et de relations (entity and relation extraction) ;
- Analyse de dépendances syntaxiques (dependency parsing) ;
- Apprentissage automatique (machine learning) ;
- Representations des connaissances (knowledge representation).

Porfil recherché :
- Master 2 (ou équivalent) en informatique ou mathématiques appliqués ;

- Spécialisation en traitement automatique des langues (TAL) ou en apprentissage automatique (machine learning) ;
- Connaissances en réseaux de neurones et apprentissage profond (deep learning) ;
- Des connaissances en représentations des connaissances seraient appréciées ;
- Bonne maîtrise des langages python et C++ ;
- Bon niveau d'anglais ;
- Bon niveau de français.

Pour postuler
Envoyer les documents suivants à Nadi Tomeh (tomeh@lipn.fr)

- CV ;
- Diplômes et relevés des notes ;
- Lettre de motivation ;
- Lettres de recommandations ;
- Mémoire du master.

Communication scientifique
Le/la doctorant.e sera encouragé.e à publier ses résultats dans les grandes conférences du domaine (par exemple ACL, EMNLP, NAACL, COLING, CoNLL, etc).

Salaire
Environ 1,800 euros net par mois.

Contexte scientifique
L'extraction des entités et des relations est une tâche importante de l'extraction d'informations qui vise à identifier les entitées mentionnées dans un texte ainsi que les relations sémantiques entre elle. La découverte de ces relations pourrait être bénéfique pour de nombreuses applications du traitement du langage naturel (TAL), notamment le peuplement de bases de connaissances (Zhang et al., 2017) et les systèmes de questions-réponses (Yu et al., 2017).

Traditionnellement, une approche pipeline est utilisée pour extraire d'abord les mentions d'entités, puis pour prédire les relations entre chaque paire de mentions d'entités extraites (Chan et Roth, 2011). Des modèles joints d'extraction d'entités et de relations ont été proposés (Miwa et Sasaki, 2014 ; Ren et al., 2017) ont été construits pour tirer profit de l'interaction étroite entre ces deux tâches. Tout en montrant les avantages de la modélisation conjointe, ces méthodes d'apprentissage structurés reposent fortement sur l'ingénierie des fonctions caractéristiques (features).

Avec le succès des réseaux neuronaux profonds, des méthodes d'apprentissage de représentations utilisant les CNNs, LSTMs, ou Tree-LSTMs sur la séquence de mots entre deux mentions d'entités ont été utilisées pour encoder les informations pertinentes pour chaque paire de mentions d'entités (Zeng et al., 2014 ; dos Santos et al., 2015). Toutefois, ces méthodes

supposent que les mentions d'entités sont données et leurs performances devraient se dégrader de manière significative lorsque l'extraction d'entités est nécessaire dans le pipeline.

Un défi pour l'extraction des relations est donc de prendre en compte l'interaction entre les relations, ce qui est particulièrement important pour les relations qui se partagent des mentions d'entités communes. Des approches basés sur les réseaux convolutionnels de graphes (Fu et al., 2019), sur la recherche en faisceau (beam search) (Lin et al., 2020) ou sur une modélisation seq2seq (Zeng et al., 2018 ; Cui et al., 2018) ont été proposé pour modéliser ces interactions.

La dépendance sur un analyseur syntaxique externe dans la plupart de ces modèles est souvent un problème (Socher et al. 2012 ; Xu et al. 2015 ; Li et al., 2015 ; Zhang, Qi, et Manning 2018). Dans ces travaux, les arbres syntaxiques sont utilisés pour structurer les graphes de calcul des architectures neuronaux. Cependant, un parseur indépendant n'est pas garanti de produire les arbres les plus adaptés pour l'ER. Une modélisation jointe de la structure syntaxique et des relations sémantiques pourrait produire des meilleurs résultats en encourageant la compatibilité entre les représentations syntaxiques et sémantiques (Veyesh et al., 2020).

La plupart des systèmes d'ER supervisés existants nécessitent une grande quantité de données étiquetée en relations spécifiques, ce qui est coûteux et demande beaucoup de travail. Les bases de connaissances existantes pourraient alors être utilisées comme source de supervision faible (Lin et al., 2016). Des approches complètement non-supervisées ont été proposées utilisant les bases de connaissances uniquement pour localiser les entitées et une fonction objective exploitant la perte de reconstruction pour l'apprentissage (Marcheggiani and Titov, 2016), ou une fonction objective mieux adaptée à l'apprentissage non-supervisé (Simon et al., 2019). Cependant, ces systèmes ne modélisent pas les entitées et les relations conjointement.

Proposition

Comme les travaux précédents (Ammar et al. 2018 ; Buscaldi et al., 2019), nous nous concentrons sur l'extraction de relations dans un domaine technique. Suivant Kuiper et al. (2020) nous visons à combiner l'extraction d'informations ouverte (Open IE) et celle basée sur des schémas prédéfinis. Notre objectif est de construire un modèle neuronal joint pour l'extraction d'entités et de relations. Nous proposons d'étendre au système joint le modèle de Wang et al. (2019) qui utilise des transformeurs de type BERT pré-entraînés pour l'encodage de l'entrée. L'apprentissage structuré neuronal basé sur les graphes (Wu et al., 2020) sera utilisé pour modéliser l'interaction entre les relations dans la sortie. Nous expérimentons également la modélisation des dépendances syntaxiques en marginalisant sur l'ensemble d'arbres de dépendance avec attention structurée (Kim et al., 2017) ou avec représentation explicite (Jin et al., 2020) pour guider l'extraction. Nous utiliserons des méthodes d'apprentissage non supervisé ou avec supervision faible pour choisir les paramètres du modèle.

Open PhD position (CIFRE):
Deep learning for entity and relation extraction in the scientific domain

Scientific context:
The natural language and knowledge representation team (RCLN; https://lipn.univ-paris13.fr/accueil/equipe/rcln/), a member of the computer science laboratory (LIPN) of the University Sorbonne Paris Nord (Paris 13; https://www.univ-paris13.fr/). RCLN is a member of the Laboratory of Excellence "Empirical Foundations of Linguistics" (LabexEFL; http://www.labex-efl.com).

Industrial context:
An international consulting group in the field of Research, Development and Innovation, working on organisational, structural, methodological, scientific and financial aspects. The company advises the most advanced companies in their sectors, those that innovate and base their strategic advance on experimental and fundamental research, with over 20 years of experience and several thousand collaborations throughout the world, on all continents. Within the group, the Scientific Department is responsible for coordinating all scientific actions, both at the operational, methodological and conceptual levels. In France, it relies on the resources of more than 60 PhDs from all disciplines. At the heart of the Scientific Department, the PhD student will join the Research Lab team, the group's internal R&D department based at La Défense in Paris.

PhD Subject
The selected candidate will work on joint extraction of relevant entities and their semantic relations in technical text, edited by experts in various technical and scientific domains (details below;
https://lipn.univ-paris13.fr/~tomeh/public/uploads/offers/phd-cifre-relation-extraction.pdf).

Keywords
- Natural language processing;
- Deep Learning;
- Entity and relation extraction;
- Dependency parsing;
- Knowledge representation.

Candidate profile
- Masters (or equivalent) in Computer Science;
- Specialisation in natural language processing or machine learning with focus on deep learning;
- Experience in knowledge representation is appreciated;
- Good programming skills in Python and/or C++;
- Proficiency in French and English.

Application

To apply, send the following documents to Nadi Tomeh (tomeh@lipn.fr)

- CV
- Academic transcripts
- Master thesis or report (if available)
- Motivation letter adapted to the context
- Recommendation letters

Scientific communication
Publication of results in top conferences in the domain will be encouraged.

Salary
1,800 euros/month net income

Detailed subject
Entity and relationship extraction is an important task of information retrieval that aims at identifying the relevant entities mentioned in a text as well as the semantic relationships between them. The discovery of these relationships is beneficial for many Natural Language Processing (NLP) applications, including knowledge base populating (Zhang et al., 2017) and question-answering (Yu et al., 2017).

Traditionally, a pipeline approach is used to first extract entity mentions and then to predict the relationships between each pair of extracted entity mentions (Chan and Roth, 2011). Joint models of entity and relation extraction have also been proposed (Miwa and Sasaki, 2014; Ren et al., 2017) to take advantage of the close interaction between these two tasks. While demonstrating the advantages of joint modeling, these structured learning methods rely heavily on feature engineering.

With the success of deep neural networks, representation learning methods using CNNs, LSTMs, or Tree-LSTMs on the word sequence between two entities have been used to encode the relevant information for each pair of entity mentions (Zeng et al., 2014; dos Santos et al., 2015). However, these methods assume that the entity mentions are given and their performance is expected to degrade significantly when entity extraction is required in the pipeline.

A challenge for relationship extraction is therefore to take into account the interaction between relationships, which is particularly important for relationships that share common entity mentions. Approaches based on convolutional graph networks (Fu et al., 2019), beam search (Lin et al., 2020) or seq2seq modelling (Zeng et al., 2018; Cui et al., 2018) have been proposed to model these interactions.

Dependence on an external parser in most of these models is often a problem (Socher et al. 2012; Xu et al. 2015; Li et al., 2015; Zhang, Qi, and Manning 2018). In this work, syntax trees are used to structure the computational graphs of neural architectures. However, an

independent parser is not guaranteed to produce the most suitable trees for RE. Joint modeling of syntactic structure and semantic relations could produce better results by encouraging compatibility between syntactic and semantic representations (Veyesh et al., 2020).

Most existing supervised RE systems require a large amount of data labelled with specific relationships, which is costly and labour-intensive. Existing knowledge bases could then be used as a source of weak supervision (Lin et al., 2016). Completely unsupervised approaches have been proposed using knowledge bases only to locate entities and an objective function exploiting reconstruction loss for learning (Marcheggiani and Titov, 2016), or an objective function better suited to unsupervised learning (Simon et al., 2019). However, these systems do not model entities and relationships jointly.

Proposition

As in previous work (Ammar et al. 2018; Buscaldi et al., 2019), we focus on extracting relationships in a technical field. Similar to Kuiper et al. (2020) we aim to combine open information extraction (Open IE) and extraction based on predefined schemas. Our goal is to build a joint neural model for entity and relationship extraction. We propose to extend the model of Wang et al (2019) which uses pre-trained BERT transformers for input encoding to the joint system. Graph-based neural structured learning (Wu et al., 2020) can be used to model the interaction between relations in the output. We propose to model syntactic dependencies by marginalizing over the set of dependency trees with structured attention (Kim et al., 2017) or with explicit representation (Jin et al., 2020) to guide extraction.

References

- Yee-Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extrac- tion. In Proceedings of ACL.
- Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table repre- sentation. In Proceedings of EMNLP.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In Proceedings of COLING.
- Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In Proceedings of ACL.
- Jiwei Li, Minh-Thang Luong, Dan Jurafsky, and Eu- dard Hovy. 2015. When are tree structures neces- sary for deep learning of representations? In Proceedings of EMNLP.
- Diego Marcheggiani and Ivan Titov. 2016. Discrete- state variational autoencoders for joint discovery and factorization of relations. Transactions of ACL.
- Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In Proceedings of WWW.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural Relation Extraction with Selective Attention over Instances. In Proceedings of ACL.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In Proc. of EMNLP.
- Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In Proceedings of ACL.
- Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cíero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In Proc. of ACL.
- Yoon Kim, Carl Denton, Luong Hoang, Alexander M. Rush. 2017. In proceedings of ICLR.
- Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In Proceedings of ACL.
- Lei Cui, Furu Wei, Ming Zhou. 2018. Neural Open Information Extraction. In proceedings of ACL.
- Tsu-Jui Fu, Peng-Hsuan Li, Wei-Yun Ma. 2019. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. In proceedings of ACL.
- Étienne Simon, Vincent Guigue, Benjamin Piwowarski. 2019. Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses. In proceedings of ACL.
- Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine

van Zuylen, Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In proceedings of NAACL-HLT.

- Davide Buscaldi, Danilo Dessì, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero. 2019. Mining Scholarly Publications for Scientific Knowledge Graph Construction. In Extended Semantic Web Conference.
- Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, Saloni Potdar. 2019. Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers. In proceedings of ACL.
- Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu. 2020. A Comprehensive Survey on Graph Neural Networks. In IEEE Transactions on Neural Networks and Learning Systems.
- Lifeng Jin, Linfeng Song, Yue Zhang, Kun Xu, Wei-yun Ma, Dong Yu. 2020. Relation Extraction Exploiting Full Dependency Forests. In proceedings of AAAI.
- Ying Lin and Heng Ji and Fei Huang and Lingfei Wu. 2020. A Joint Neural Model for Information Extraction with Global Features. In proceedings of ACL.
- Ruben Kruiper, Julian F.V. Vincent, Jessica Chen-Burger, Marc P.Y. Desmulliez, Ioannis Konstas. 2020. In Laymans Terms: Semi-Open Relation Extraction from Scientific Texts. In proceedings of ACL.