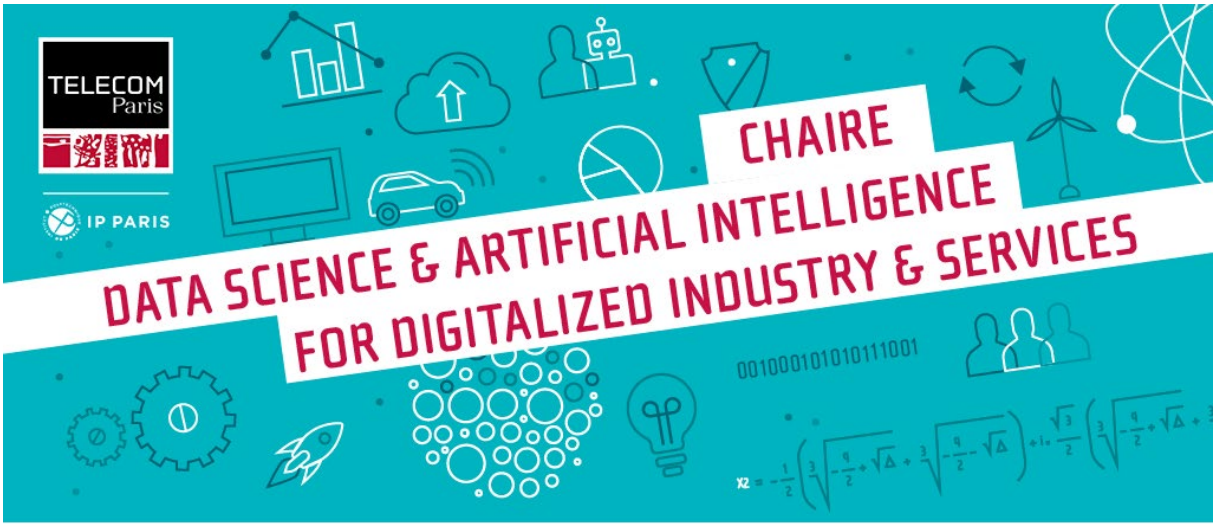

Offre de stage

Représentations de sortie pour la prédiction structurée de texte

Possibilité de poursuivre sur une thèse

A propos de la chaire

La Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services, portée par Florence d'Alché-Buc, enseignante-chercheuse dans le département Image, Données, Signal de Télécom Paris, la chaire DSAIDIS réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Son objectif général est de développer, en liaison étroite avec les partenaires, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

En savoir plus sur la chaire : **www.telecom-paris.fr/dsaidis**

Description du stage

Encadrement

Matthieu Labeau, Florence d'Alché-Buc

Lieu et dates du stage

Adresse : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage :

Début 2020

Équipe(s) d'accueil de la thèse

Département IDS, équipe S2A, Chaire DSAIDIS

Mots clés

Traitement automatique des langues, représentation structurée, plongement de phrases

Sujet détaillé

La prédiction de sortie structurée a une importance particulière dans de nombreux domaines comme la bio-informatique, la vision artificielle, ou le traitement automatique du langage. Cependant, elle pose de nombreuses difficultés, liées à la difficulté d'apprentissage et à la complexité des données, qui sont souvent rares. Plusieurs familles de modèles permettent de telles prédictions : par exemple, les energy-based models, qui sont très coûteux à entraîner et utiliser [7] et les approches neuronales end-to-end également coûteuses à entraîner [1]. Une autre ligne de recherche consiste à plonger les sorties à prédire dans un espace vectoriel et à transformer ainsi le problème en celui de la régression d'un ensemble de caractéristiques de sortie: par exemple, dans [3], on utilise une transformation linéaire du score de Fisher de la sortie. Dans [2], on définit un noyau de sortie ce qui revient à plonger les données dans un espace de Hilbert de dimension infinie. Enfin dans [8], on plonge les données dans un espace sémantique en utilisant GloVe ou Word2Vec.

L'objectif du stage est l'application de telles méthodes aux données textuelles, et notamment à la génération de texte. La génération de texte comprend un nombre important de tâches potentiellement très différentes (Abstractive Summarization, Question Answering, Automatic Image Description, Machine Translation…) qu'un grand nombre de travaux récents aborde à l'aide des modèles sequence-to-sequence, utilisant différentes méthodes selon la tâche et la supervision (auto-encodeurs variationnels [4], autoencoderus kernelisés, méthodes adversariales [6]).

On pourra dans un premier temps s'intéresser aux tâches traditionnelles de prédiction structurée en traitement automatique du langage (dont un exemple relativement simple est la prédiction d'une séquence de tags correspondant à une phrase d'entrée). On pourra aussi s'intéresser à la récente utilisation du score de Fisher pour la génération d'images [4].

Profil du candidat

Etudiant titulaire d'un master 2 recherche avec de solides connaissances en :

- Apprentissage statistique / reconnaissance des formes
- Traitement du langage naturel
- Programmation (Java, C/C++, Python)
- Niveau d'anglais

Candidatures

A envoyer à matthieu.labeau@telecom-paris.fr, florence.dalche@telecom-paris.fr

- Curriculum Vitae
- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)
- Relevés de notes des années précédentes
- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Références

[1] [Belanger et al. 2017] D. Belanger, B.Yang, A. McCallum: End-to-End Learning for Structured Prediction

Energy Networks. ICML 2017, 429-439, 2017.

[2] C. Brouard, M. Szafranski, F. d'Alché-Buc, Supervised and semi-Supervised Input Output

Kernel Regression, JMLR oct, 2016.

[3] Moussab Djerrad, Alexandre Garcia, Maxime Sangnier, Florence d'Alché-Buc.Output Fisher embedding regression. Machine Learning, 2018.

[4] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Bengio.Generating Sentences from a Continuous Space. Conference on Computational Natural Language Learning, 2016.

[5] P. Laforgue, S. Clémençon, F. d'Alché-Buc, Autoencoding any data with a kernel auto-encoder, AISTAT 2019.

[6] Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, Hongyan Li. Generative Adversarial Network for Abstractive Text Summarization. AAAI Conference on Artificial Intelligence, 2018.

[7], S. Nowozin, C. Lampert, Structured Learning and Prediction in Computer Vision. Foundations and Trends in Computer Graphics and Vision 6(3-4): 185-365 (2011).

[8] Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, Tom Mitchzll, Zero-shot learning with semantic output codes, NIPS 2009.

[9] Yang Song, Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. Yang Song, Stefano Ermon. Conference on Neural Information Processing Systems, 2019.