TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

NLP ET LLM

FFCNCERCERXBD10	PRIX : 2 580 €	DURÉE : 3 JOURS

Pauses et déjeuners offerts

AVANCÉ

ATELIER

RÉALISABLE EN ANGLAIS

FAISABLE À DISTANCE

PRÉSENTATION

Les données linguistiques possèdent une structure profonde mais implicite, qui se base sur la connaissance d'une (ou plusieurs) langue(s) donnée(s). Elles sont ubiquitaires (sur le Web, dans des documents, dans les emails, etc.), mais ne se prêtent pas à des analyses automatiques.

Le traitement automatique de langue et la fouille de texte (Text Mining) ont pour but de permettre l'extraction d'informations et de connaissances de ces données. Elles sont donc d'importance capitale pour les entreprises qui manipulent des données textuelles (Web, échanges avec les clients, rapports, documentation, etc.).

Dans ce contexte, les transformers ont révolutionné le domaine du traitement automatique du langage naturel (NLP) : ils utilisent des mécanismes d'attention permettant de traiter efficacement les dépendances à long terme dans le texte. Des modèles basés sur cette architecture, comme BERT, GPT et leurs variantes, ont montré des performances remarquables sur la traduction automatique, la génération de texte, la reconnaissance d'entités nommées, et bien d'autres applications.

OBJECTIFS

- Identifier les outils de traitement de langue, qu'ils soient basés sur des méthodes statistiques ou sur de méthodes symboliques
- Expliquer le fonctionnement et identifier les atouts et les limites des grands modèles de langage LLM comme GPT-4
- Évaluer les techniques et les adapter à chaque type de problème
- Comparer et combiner les approches statistiques et symboliques

PROGRAMME

Introduction à la linguistique

Approches neuronales

Approches statistiques
- Désambiguïsation de mot
- Classification supervisée de texte
- Similarité et parenté sémantiques
- Pré-traitement du texte
- Modèles fréquentistes : Représentation Bag-of-words, modèles de langue n-gram et dérivés.
- Deep learning et modèles de langue neuronaux
- Modèles séquentiels et Mécanisme d'attention
- Transformers
- Représentations contextuelles
- Apprentissage par transfert
- Utilisation de réseaux de neurones sur le même corpus de textes)

Travaux pratiques

Approches symboliques
- Langages formels Graphes conceptuels / ontologies / bases de connaissances
- Extraction d'informations
- Désambiguïsation
- Détection d'entités

Travaux Pratiques

Synthèse et conclusion

PUBLIC/PRÉREQUIS

Ingénieurs, chefs de projets, data scientists devant traiter, générer ou intégrer des fonctionnalités avec des données textuelles et du langage naturel.

Des connaissances en langage Python sont nécessaires afin de tirer pleinement profit de cette formation.

RESPONSABLE(S)

Fabian SUCHANEK
Enseignant-chercheur à Télécom Paris. Il a fait ses recherches à l'Institut Max Planck en Allemagne, chez Microsoft Research Cambridge/UK, chez Microsoft Research Silicon Valley/USA, et à l'INRIA Saclay. Il est l'auteur principal de YAGO, une des plus grandes bases de connaissances publiques dans le monde.

Matthieu LABEAU
Maître de conférences à Télécom Paris. Il a effectué son doctorat à l'Université Paris-Saclay et s'est intéressé aux problèmes posés par les grands vocabulaires en modélisation du langage, il a ensuite été chercheur postdoctoral à l'Université d'Édimbourg. Ses domaines de recherche, au sein du traitement automatique du langage, concernent principalement l'apprentissage de représentations et la modélisation du langage.

MODALITÉS PÉDAGOGIQUES

La formation comprend des travaux pratiques qui permettent d'appliquer les notions théoriques abordées.