Master Informatique

Prédiction structurée pour le traitement auto. de la langue

Informations

Composante : Faculté des Sciences

Responsable

Carlos RAMISCH (Enseignant)

Langue(s) d'enseignement

Anglais

Contenu

Ce cours introduit des modèles, algorithmes, ressources et outils pour la résolution de problèmes structurés en traitement automatique des langues (TAL), en particulier à l'aide de méthodes d'apprentissage profond (deep learning). Le cours est composé de 6 blocs thématiques, chacun avec 2h de cours/TD et 2h de TP avec implémentation du modèle étudié en python avec la bibliothèque pytorch.

1. Étiquetage en parties du discours avec réseaux de neurones récurrents (RNN)

- Modèle neuronal de classification de textes

- Réseaux de neurones récurrents, p.ex. GRU

2. Reconnaissance d'entités nommées avec modèles de Markov cachés (HMM)

- HMM : estimation de paramètres et inférence (algorithme de Viterbi)

- Entités nommées et encodage begin-inside-outside

3. Prédiction de traits morphologiques avec modèle sous-lexical et multi-tâches

- Modèles sous-lexicaux (RNN ou convolution 1D sur les caractères)

- Apprentissage joint multi-tâches avec partage de paramètres

4. Modèle BERT: principes et pré-entraînement

- Prédiction de super-senses avec embeddings contextuels

- Apprentissage par transfert : affinage (fine-tuning)

5. Analyse syntaxique en dépendances par transitions

- Représentation d'arbres comme séquences d'actions (oracle)

- Modèle d'analyse par transitions : pile, buffer, configuration, action

6. Analyse syntaxique en dépendances par graphe

- Architecture analyseur bi-affine

- Arbre couvrant maximal (Algorithme Chu-Liu-Edmonds)

Compétences à acquérir

- Concevoir et mettre en oeuvre une solution logicielle pour un problème spécialisé aux données textuelles

- Appliquer des notions théoriques de statistiques et de science des données à un contexte spécialisé aux données textuelles

- Élaborer une démarche scientifique appliquée à l'analyse et génération des données textuelles

- Présenter oralement et à l'écrit les enjeux, capacités et limitations d'une solution proposée

Modalités d'organisation

Cours magistraux (9h) intégrés aux travaux dirigés (9h), et travaux pratiques (TP, 9h).

Modalités de contrôle des compétences et connaissances :

NoteFinale = 1/2 * Projet + 1/2 * ExamenTerminal Projet = 1/2 * Rapport + 1/2 * Soutenance La note de projet est composée d'une note de rapports de TP et d'une note de présentation orale.

Les TP notés seront à choisir parmi les TP proposés, la présentation orale portera sur des extensions à partir des TP proposés.

Bibliographie, lectures recommandées

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft online) - https://web.stanford.edu/~jurafsky/slp3/

- Yoav Goldberg. Neural network methods for NLP - https://www.morganclaypool.com/doi/abs/10.2200/S00762ED1V01Y201703HLT037

- Mohammad Taher Pilehvar and Jose Camacho-Collados. Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning - https://www.morganclaypool.com/doi/abs/10.2200/S01057ED1V01Y202009HLT047

Pré-requis obligatoires

apprentissage automatique, deep learning, python

Prérequis recommandés

Modèles de langage

VOLUME HORAIRE

- Volume total: 27 heures
- Cours magistraux: 9 heures
- Travaux dirigés: 9 heures
- Travaux pratiques: 9 heures

Codes Apogée

- SINCC8HJ [ELP]

Pour plus d'informations

Aller sur le site de l'offre de formation...

Dernière modification le 13/11/2024