Offre de stage « Classification automatique de textes traduits en français et textes non traduits »

Équipes
Au sein du LIG, l'équipe GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) est composée d'une cinquantaine de personnes (permanent-e-s, doctorante-e-s, master-e-s). Issue de l'union vertueuse de chercheurs en traitement de l'écrit et de la parole, le GETALP est une équipe pluridisciplinaire (informaticiens, linguistes, phonéticiens, traducteurs et traiteurs de signaux, …) dont l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale).

Au sein de l'UMR Litt&Arts, le stage se déroulera en lien avec l'équipe ELAN (Élan, Littératures, Arts et Numériques), chargée des traitements et analyses des données, de la conception et du développement d'outils d'exploitation et de visualisation des données.

Le projet PSTT (Poétique et stylistique du texte traduit), qui étudie le texte littéraire traduit en français dans son autonomie (indépendamment du texte source) et dans ses rapports avec la langue et les formes littéraires française (en diachronie et en synchronie).
Ce stage permettra d'acquérir des compétences techniques en en apprentissage automatique appliqué au texte et des connaissances générales sur les enjeux de la traduction littéraire.

Objectifs
L'objectif du stage sera d'apprendre un système qui reconnaisse ces deux types de textes, à l'aide d'un corpus constitué de romans écrits en français et de romans traduits, voire d'apprendre un système qui puisse modifier les textes traduits de façon à ce qu'ils ne soient plus distinguables de textes non traduits. L'hypothèse que ce stage permettra d'explorer est qu'il existe des caractéristiques linguistiques et stylistiques différentielles permettant de distinguer les textes traduits en français et les textes originellement écrits en français. Le stage aura donc une application directe tant dans le domaine du TAL qu'en Littérature.
Le corpus utilisé dans cet objectif sera constitué de romans contemporains (publiés après 1980) déjà disponibles sous forme de texte brut dans le corpus Emolex du LIDILEM (https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/emolex) et/ou à partir d'autres sources.

Tâches à réaliser
Formaliser la ou les tâches. Une première tâche directe serait un classifieur discriminant les textes traduits des textes non traduits.
Étudier les modèles de classifications possibles et les ressources sur lesquels ils peuvent se baser. Proposer des implémentations. On pourra viser des classifications basés par exemple sur :
- les approches sac de mots
- les méthodes vectorielles statiques (word2vec, glove, fasttext…) – Mikolov, 2013
- les méthodes de comptages (Baroni, 2014)
- les modèles de langues contextuels à la BERT (Devlin et al, 2019 ; Le et al., 2020)

Analyser les résultats et proposer des améliorations sur la tâche

Compétences requises :
- Connaissances de bases en python
- Un intérêt pour le domaine de la littérature et/ou de la traduction sera apprécié

Lieux :

Maison de la création et de l'innovation (MACI), Université Grenoble Alpes, 339 Av. Centrale, 38400 Saint-Martin-d'Hères.

ET

Bâtiment IMAG - Université Grenoble Alpes - 700 avenue Centrale - Domaine Universitaire, 38401 St Martin d'Hères – France
Les deux bâtiments sont proches (400 mètres).

Période du stage : à partir de février 2024

Durée : 5-6 mois

Encadrement :
La personne recrutée sera co-encadrée par :
- Didier Schwab au LIG, enseignant-chercheur en informatique au LIG
- Emmanuelle Esperança-Rodier, enseignante-chercheure au LIG
- Anne Garcia-Fernandez, ingénieure de recherche CNRS à Litt&Arts
- Pascale Roux, enseignante-chercheure à Passages XX-XXI (Lyon 2), porteuse du projet PSTT de Litt&Arts

Comment candidater ?
Les candidatures devront comporter un CV, une lettre de motivation et les relevés de notes dans le supérieur. Elles doivent être adressées avant le 22/11/2023 aux 4 adresses suivantes :

`didier.schwab@univ-grenoble-alpes.fr,`
`emmanuelle.esperanca-rodier@univ-grenoble-alpes.fr`
`pascale.roux@univ-lyon2.fr`
`annegf@univ-grenoble-alpes.fr`

avec comme objet `[Stage PSTT] Candidature de Prénom Nom`

Biographie
Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey, *Efficient Estimation of Word Representations in Vector Space*, 2013, https://arxiv.org/abs/1301.3781
Marco Baroni, Georgiana Dinu, and Germa´n Kruszewski. 2014. *Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors*. In 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), pages 238–247, Baltimore, Maryland.
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. *FlauBERT: Unsupervised Language Model Pre-training for French*. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 2479–2490, Marseille, France. European Language Resources Association.