Stage de M2 Informatique au sein du projet JOKER: Détection et interprétation des Jeux de Mot avec des méthodes d'Apprentissage Profond

Durée: 5-6 mois

Lieu du stage: Centre Européen de Réalité Virtuelle, (ENIB) 25 rue Claude Chappe, Plouzané, France / HCTI (UBO) / MSHB

Site web du projet: http://www.joker-project.com/

Contacts: Liana Ermakova, HCTI, UBO (Liana.Ermakova@univ-brest.fr), Anne-Gwenn Bosser , COMMEDIA, Lab-STICC, ENIB (bosser@enib.fr)

Pré-requis: M1 informatique ou équivalent, la maîtrise de la langue française et anglaise est souhaitable.

Contexte

Ce stage participe au projet JOKER qui vise à faire progresser l'automatisation de la traduction des jeux de mots en fournissant un corpus parallèle (c'est à dire multilingue) approprié [1].
Alors que la traduction moderne est fortement aidée par des outils technologiques, pratiquement aucun n'a de support spécifique pour les jeux de mots. En effet, la plupart des outils de traduction basés sur l'IA nécessitent une qualité et une quantité de données d'entraînement (par exemple, des corpus parallèles) qui ont toujours fait défaut pour les jeux de mots. L'objectif du projet JOKER est de construire automatiquement un corpus parallèle de jeux de mots en entraînant un classificateur basé sur l'IA à détecter automatiquement les instances de jeux de mots et à les aligner avec leurs traductions.

Objectifs du stage (liste indicative)

- Détection des jeux de mots: Une première étape de détection sera de traiter les données de sources électroniques (pdf, html etc) variées sélectionnées pour vous (romans, essais, etc.) pour augmenter les corpus disponibles, avec des méthodes d'intelligence artificielle.
- Alignement des traductions avec les algorithmes traditionnels ou avec les modèles neuronaux
- Interprétation des jeux de mots en français et en anglais
- Augmentation des corpus monolingues

Environnement technique (liste indicative)

- Python, Pandas, NLTK, expressions régulières
- bibliothèques pour traiter différents formats de fichiers (pdf, epub,...)
- De grands modèles pré-entraînés:
    - Google mT5 (https://github.com/google-research/multilingual-t5)
    - BLOOM (https://huggingface.co/bigscience/bloom)

Références

[1] L. Ermakova *et al.*, « Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop », in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Cham, 2022, p. 447‑469.

Version ouverte: https://ceur-ws.org/Vol-3180/