Titre:
Extraction d'information à partir de documents hétérogènes avec généralisation automatique multilingue et visualisation appliquées sur une base de données mondiale d'appel d'offres

Extraction of information from heterogeneous documents with automatic multilingual generalization and visualization applied to a global call for tenders database

Contexte
OctopusMind spécialiste de l'Open Data économique développe des services autour de l'analyse de données. Nous éditons, notamment, www.J360.info, plateforme mondiale de détection d'appels d'offres publics et privés, et de mise en relation professionnelle. L'entreprise d'une dizaine de personnes développe ses propres outils en mode agile à partir de technologies open source (utilisation et contribution) pour analyser une grande masse de données quotidiennement. Elle est engagée dans des projets innovants axés sur l'intelligence humaine et l'intelligence artificielle (Traitement automatique du langage et Deep Learning).

La plus-value de l'entreprise est, notamment, d'extraire des données décisionnelles au sein d'une base de données de marchés publics, de plus de 20 millions de documents, dans un format non structuré multilingue, contenant des informations pertinentes (informations stratégiques, zones géographiques, chiffre d'affaires, prix, nom commercial etc). L'entreprise dispose également d'une base de données [7] pour la classification de textes et la traduction automatique issue de la base TED[1].

Travaux envisagés
L'objectif est de proposer une méthodologie qui permettra de manière supervisée ou semi-supervisée d'extraire une ou plusieurs informations [5,6], à partir de documents non structurés multilingues.

Ces informations pourront être par exemple des entitées nommées et des relations entres ces dernières. La méthode devra résoudre un problème d'extraction d'informations multilingue à partir de données étiquetées dans une seule langue.

Aussi, il sera nécessaire de développer des méthodes de visualisation afin de permettre l'interprétation du modèle.

Plusieurs méthodes pourront être explorées, notamment:

- Utilisation de réseaux de neurones profonds (i.e. BERT [8]), afin de projeter les vecteurs de mots de différentes langues dans le même espace de représentation, le modèle sera entraîné sur une ou plusieurs langues et la prédiction sera multilingue.
- Traduction de dataset en utilisant des méthodes d'alignement statistique ou des réseaux de neurones profonds [8].
- Utilisation des réseaux GAN [2, 1] (Generative Adversarial Network) afin de générer des données synthétiques dans le but d'enrichir et/ou traduire des datasets (pouvoir créer des exemple synthétiques dans plusieurs langues).

1 https://ted.europa.eu/

Profil et compétences recherchées

Niveau master ou école d'ingénieur avec des compétences en informatique et traitement des données. Des connaissances en intelligence artificielle (machine learning), fouille de texte, TALN (Traitement Automatique du Langage Naturel) et traitement statistique des données seront appréciées.

La maîtrise de la langue française n'est pas indispensable : si ce critère est un plus, la qualité du dossier sera privilégié.

Conditions

La thèse sera effectuée en partie dans les locaux de l'entreprise OctopusMind à Nantes, et en partie dans les locaux de l'IRISA à Vannes ou Lorient.

- L'Entreprise OctopusMind[1], PME d'une 15aine de collaborateurs, développe des services et outils de traitements automatisés de l'information stratégique pour la veille d'opportunités commerciales depuis 2005.

- L'équipe EXPRESSION[2] du Laboratoire IRISA, site de Vannes apportera son savoir-faire en matière de recherche dans le domaine de la fouille de texte, du TALN et du machine learning. L'accès à des moyens de calcul distribués sera assuré par l'équipe et l'UMR IRISA.

- Démarrage de la thèse : dès que possible

Contacts :

OctopusMind : Oussama Ahmia : o.ahmia@octopusmind.info

IRISA : Nicolas Béchet, Pierre-François Marteau :

{nicolas.bechet, pierre-francois.marteau}@univ-ubs.fr

1 https://www.octopusmind.info/

2 https://www-expression.irisa.fr/

Bibliographie

1. RASHID, Ahmad, DO-OMRI, Alan, REZAGHOLIZADEH, Mehdi, et al. Bilingual-GAN: Neural Text Generation and Neural Machine Translation as Two Sides of the Same Coin. 2018.
2. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.
3. Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training gans." In Advances in neural information processing systems, pp. 2234-2242. 2016.
4. Lucic, Mario, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. "Are gans created equal? a large-scale study." In Advances in neural information processing systems, pp. 700-709. 2018.
5. Freitag, Dayne. "Machine learning for information extraction in informal domains." Machine learning 39, no. 2-3 (2000): 169-202.
6. Cowie, Jim, and Wendy Lehnert. "Information extraction." Communications of the ACM 39, no. 1 (1996): 80-91.
7. AHMIA, Oussama, BÉCHET, Nicolas, et MARTEAU, Pierre-François. Two Multilingual Corpora Extracted from the Tenders Electronic Daily for Machine Learning and Machine Translation Applications. In : Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
8. DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.