Élaboration d'un protocole d'annotation et extraction d'information à partir de données textuelles : application au suivi de la sécurité alimentaire

Stage master 2 – 2024 (6 mois)

Contexte général et projet de recherche

Le stage s'inscrit dans le cadre d'activités interdisciplinaires de l'UMR TETIS concernant l'anticipation et le suivi des risques liés à la sécurité alimentaire. Les activités de recherche de l'UMR sur cette thématique sont centrées sur le cas de l'Afrique de l'Ouest, où les risques agricoles sont d'autant plus aigus que les services nationaux de surveillance et de suivi peuvent être défaillants faute de moyens techniques et financiers. Pour calculer des indices de risque de crises alimentaire, les modèles prédictifs utilisent et combinent des informations issues sources de données hétérogènes, incluant classiquement de l'imagerie satellitaire, des données climatiques, ou encore des séries temporelles de prix alimentaires.

L'analyse et l'interprétation de ces données et des indices prédits pourraient être facilitées par l'utilisation conjointe de données provenant de sources journalistiques, telles que les articles en ligne ou les transcriptions de journaux télévisés (Ba et al. 2022; Deléglise et al. 2022). Ces données textuelles pourraient par exemple permettre de localiser correctement les événement conjecturels et structurels de situations de crises à l'échelle régionale voire locale, en temps quasi-réel ou rétrospectivement. Dans ce contexte, il devient indispensable de déployer des méthodes d'apprentissage automatique permettant d'extraire les informations pertinentes à partir de vastes volumes de données textuelles.

Les approches automatiques de classification et d'extraction d'informations, ainsi que leur évaluation, reposent sur la disponibilité de corpus annotés, utilisés pour paramétrer et/ou fine-tuner les modèles. A notre connaissance, il n'existe pas de corpus annoté spécifique aux enjeux du suivi de la sécurité alimentaire. Une récente étude a mis en évidence des corrélation spatio-temporelles significatives entre les occurrences de mots-clés spécifiques et des variables classiquement utilisées pour le suivi de la sécurité alimentaire (e.g. NDVI, décès issus de conflits, etc.) (Balashankar, Subramanian, et Fraiberger 2023). Cependant, les mots-clés sont extraits indépendamment de leur contexte d'occurrence et des entités spatio-temporelles de l'article associé, apportant un risque de biais dans la localisation spatio-temporelle des risques. Nous proposons donc d'élaborer une approche d'annotation la plus générique possible permettant d'identifier les facteurs structurels et conjecturels de crises alimentaires, ainsi que leurs attributs spatio-temporels, dans les données textuelles issues de l'actualité. Le corpus à partir duquel seront sélectionnées les données à annoter ont déjà été collectées dans le cadre d'un précédent stage de Master 2 réalisé en 2022 (corpus en français, couvrant l'actualité du Sénégal, du Burkina Faso, du Bénin de 2012 à 2022).

Ba, Cheick Tidiane, Chloé Choquet, Roberto Interdonato, et Mathieu Roche. 2022. « Explaining food security warning signals with YouTube transcriptions and local news articles ». In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, 315-22. GoodIT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3524458.3547240.

Balashankar, Ananth, Lakshminarayanan Subramanian, et Samuel P. Fraiberger. 2023. « Predicting food crises using news streams ». *Science Advances* 9 (9): eabm3449. https://doi.org/10.1126/sciadv.abm3449.

Deléglise, Hugo, Agnès Bégué, Roberto Interdonato, Elodie Maître d'Hôtel, Mathieu Roche, et Maguelonne Teisseire. 2022. « Mining News Articles Dealing with Food Security ». In *Foundations of Intelligent*

Systems, édité par Michelangelo Ceci, Sergio Flesca, Elio Masciari, Giuseppe Manco, et Zbigniew W. Raś, 63-73. Cham: Springer International Publishing.

Objectifs du stage

En tant que stagiaire de Master 2, vos activités de recherche consisteront à :

- Réaliser une revue de littérature des approches d'annotation de corpus dans le cadre du suivi spatio-temporel d'événements et/ou facteurs de risque.
- Concevoir une méthode d'annotation dédiée à l'identification de facteurs de risque et événements en sécurité alimentaire dans les données textuelles, ainsi que leurs attributs spatio-temporels.
- Appliquer la méthodologie d'annotation sur un échantillon du corpus existant.
- Mettre en œuvre des techniques de traitement automatique du langage naturel (TALN) et d'apprentissage automatique pour extraire des informations annotées, avec un focus sur les attributs spatio-temporels. En particulier, l'adaptation (*fine-tuning*) de modèles de langue pré-entrainés de type CamemBERT sera évaluée.
- Collaborer avec notre équipe multidisciplinaire pour évaluer la performance des modèles développés et proposer des améliorations potentielles en fonction des résultats obtenus.
- En fonction du temps imparti, participer à la rédaction d'un « data paper » décrivant l'approche d'annotation et le corpus produit.

Organisation

Le stage rémunéré se déroulera sur une période de 6 mois, à compter de février 2024. L'étudiant·e sera accueilli·e au sein de l'UMR TETIS, à la Maison de la Télédétection (Montpellier) et sera encadré·e par Sarah Valentin, chercheuse en fouille de données à l'UMR TETIS (Cirad) et Maguelonne Teisseire, directrice de recherche à l'UMR TETIS (INRAE). Des réunions hebdomadaires sont prévues conjointement aux échanges informels en continu avec les encadrants du stage afin de discuter de l'avancée du travail et des éventuelles difficultés rencontrées.

Une poursuite en thèse dans la continuité de ces travaux est envisageable (demande de financement en cours).

Profil recherché

Le/la stagiaire aura un profil en informatique avec des connaissances en traitement automatique de la langue et/ou apprentissage automatique, avec un intérêt pour le travail interdisciplinaire. Une expérience dans le langage de programmation Python est un plus.

Candidature

Les candidatures (CV, lettre de motivation et relevé de notes M1 – ou 4ème année) sont à envoyer à maguelonne.teisseire@inrae.fr et sarah.valentin@cirad.fr. Date limite des candidatures : 11/11/2023.