Stage de Master Recherche 2016-2017 :
Détermination des itinéraires migratoires contextualisés à partir de récits de vies

Responsables de stage locaux (TETIS) : **Mathieu Roche**, **Maguelonne Teisseire**
Autre encadrant (MIGRINTER) : **Nelly Robin**

Localisation :
UMR TETIS (AgroParisTech, Cirad, Cnrs, Irstea)
500, rue J.F. Breton, 34093 Montpellier Cedex 5, France

Contact :
mathieu.roche@cirad.fr
maguelonne.teisseire@irstea.fr
nelly.robinsn@orange.fr

1 Contexte du projet QDoSSI

Les migrations internationales ont pris dans le monde contemporain une ampleur inédite. Cela pose de nouveaux défis à la communauté scientifique en terme d'analyse et de compréhension des phénomènes migratoires. Le premier est celui des données et de leurs qualités. En effet, nombreuses sont les bases de données statistiques sur les migrations internationales. Pour aller plus loin dans l'analyse de ce phénomène complexe et multidimentionnel, la mise en synergie d'autres types de données est nécessaire. Le projet *QDoSSI* [1] propose d'étudier les parcours migratoires d'un point de vue du migrant, considéré comme un acteur clé dont les droits doivent être respecter et la protection assurée en toutes circonstances

Notre champ d'analyse porte sur différents types de données collectées par le laboratoire MIGRINTER : (1) affaires judiciaires (100 000 enregistrements par an sur 10 ans), (2) corpus juridique des pays d'Afrique de l'Ouest et des Balkans, carrefours important des circulations migratoires vers l'Europe, (3) récits de vie (plus de 300), notamment des mineurs en mobilité et des migrants de Calais, (5) recensement effectué récemment auprès des personnes déplacées en Syrie (200 000/ 250 000 individus).

2 Contexte du stage et état de l'art

Dans le cadre du stage, le travail se concentrera sur l'identification automatique d'"itinéraires contexualisés" à partir des corpus de récits de vie par des méthodes de fouille de textes.

De nombreuses méthodes permettent de reconnaitre les Entités Nommées (EN) en général et les Entités Spatiales (ES) en particulier (Nadeau & Sekine, 2007). On trouve des approches statistiques consistant généralement à étudier les termes co-occurrents par analyse de leur distribution dans un corpus (Agirre *et al.*, 2000) ou par des mesures calculant la probabilité d'occurrence d'un ensemble de termes (Velardi *et al.*, 2001). On trouve également des méthodes de fouille de données fondées sur l'extraction de motifs. Ces derniers permettent de déterminer des règles de transduction utilisant des informations syntaxiques propres aux phrases pour repérer les ENs (Nouvel *et al.*, 2011). La plupart des méthodes d'extraction et de désambiguisation d'entités spatiales exploitent des méthodes mixtes (symboliques et statistiques) (Kergosien *et al.*, 2014).

Plusieurs travaux se sont intéressés à l'étude des trajectoires (Yuan & Raubal, 2012) mais peu se concentrent sur leur identification automatique à partir de données textuelles, tâche éminemment difficile. Un tel processus s'appuie sur l'iden-

1. Qualité des Données multi-Sources – Un double défi pour les sciences Sociales et les sciences de l'Informatique - Mastodons CNRS

tification de descripteurs linguistiques, en particulier les verbes (Talmy, 2000) et les indicateurs spatiaux (Zenasni *et al.*, 2015) et également l'utilisation de connaissances et ressources externes (gazetteers, ontologies, etc.) (Lieberman & Samet, 2012). Dans ce cadre, les travaux de (Moncla, 2015) utilisent ces différents éléments pour identifier les itinéraires à partir des textes. L'approche proposée consiste à identifier les informations qui décrivent les itinéraires dans les textes (entités spatiales, expressions de déplacement ou de perception) afin de les reconstruire automatiquement en exploitant des informations géographiques (latitude/longitude, altitude) et les informations contenues dans les textes (par exemple, l'ordre d'apparition des entités spatiales) (Moncla *et al.*, 2016).

3 Travail à réaliser

Le travail de stage qui sera effectué dans le cadre des projets *QDoSSI* et *Songes*[2] (Science des Données Hétérogènes) s'articulera autour des tâches suivantes :

1. Il s'agira, dans un premier temps, de compléter l'état de l'art des approches les plus récentes ayant adopté une démarche similaire.
2. Dans un deuxième temps, une évaluation des approches de l'état de l'art appliquées aux données des récits de vies sera conduite.
3. Puis des informations thématiques liées à chaque parcours (par exemple, "parcours dangereux", intervention d'une dimension familiale, financière, etc.) seront extraites à partir des textes et intégrés aux itinéraires.
4. Enfin, une représentation et une visualisation cartographique des "itinéraires contexualisés" sera alors mise en œuvre.

Références

AGIRRE E., ANSA O., HOVY E. H. & MARTÍNEZ D. (2000). Enriching very large ontologies using the www. In *ECAI Workshop on Ontology Learning*.

KERGOSIEN E., LAVAL B., ROCHE M. & TEISSEIRE M. (2014). Are opinions expressed in land-use planning documents ? *International Journal of Geographical Information Science*, **28**(4), 739–762.

LIEBERMAN M. D. & SAMET H. (2012). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, p. 731–740, New York, NY, USA : ACM.

MONCLA L. (2015). *Automatic reconstruction of itineraries from descriptive texts. (Reconstruction automatique d'itinéraires à partir de textes descriptifs)*. PhD thesis, University of Pau and Pays de l'Adour, France.

MONCLA L., GAIO M., NOGUERAS-ISO J. & MUSTIÈRE S. (2016). Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science*, **30**(6), 1137–1160.

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1), 3–26.

NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & SOULET A. (2011). Recognizing named entities using automatically extracted transduction rules. In *(LTC'2011)*.

TALMY L. (2000). *Toward a Cognitive Semantics*. Number vol. 1 in Bradford book. MIT Press.

VELARDI P., FABRIANI P. & MISSIKOFF M. (2001). Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, p. 270–284.

YUAN Y. & RAUBAL M. (2012). *Extracting Dynamic Urban Mobility Patterns from Mobile Phone Data*, In N. XIAO, M.-P. KWAN, M. F. GOODCHILD & S. SHEKHAR, Eds., *Geographic Information Science : 7th International Conference, GIScience 2012, Columbus, OH, USA, September 18-21, 2012. Proceedings*, p. 354–367. Springer Berlin Heidelberg : Berlin, Heidelberg.

ZENASNI S., KERGOSIEN E., ROCHE M. & TEISSEIRE M. (2015). Discovering types of spatial relations with a text mining approach. In *Foundations of Intelligent Systems - 22nd International Symposium, ISMIS 2015, Lyon, France, October 21-23, 2015, Proceedings*, p. 442–451.

2. http ://textmining.biz/Projects/Songes/