Assises MAGIS - AP HNS - 24 juin

Programme et supports de présentations :
https://projet.liris.cnrs.fr/aphns-magis/AtelierAssisesMAGIS2020.html

Participants : environ 50 personnes connectés

Ludovic Moncla, Carmen Brando, Bertrand Dumenieu, Jean Pierre Girard, Christian Sallaberry, Liset Diaz , Christine Plumejeaud, Matthieu Noucher, Yoann Dupont, Tian Tian, Gaël Lejeune, Sarah Kalinowski, Katie McDonough, Julie Aucagne, Thérèse Libourel, Anne Laurent, Géraldine Del Mondo, Margot Ferrand, Philippe Gambette, Frédérique Mélanie, Vincent Jolivet, Vincent Razanajao, Alberto Dalla Rosa, Motasem Alrahabi, Sylvie Servigne, Alexis Grillon, Thierry Joliveau, Nathalie Abadie, Marie Gradeler, Julien Velcin, Iris Eshkol-Taravella, Gabriela Elgarrista, Maguelonne Teisseire, Mathieu Roche, Laurence Jolivet, Catherine Dominguès, Claudia Marinica, ...

Session 1 - Ressources

- Emmanuelle Perrin. Extraire et structurer des informations géographiques sur les lieux de fouille archéologiques : le référentiel du site de Bibracte (mont Beuvray, Morvan)

Les relations spatiales sont exprimées sous formes de relations associatives (skos) dans le thesaurus.
Questions:
Q1. Vous avez fait thésaurus, n'auriez-vous pas besoin de certaines propriétés des ontologies ?
A1. thésaurus plus facile à utiliser et fait partie d'un projet plus large (HyperThésau) ; recherches continuent dans le cadre d'un autre projet (HisArc-RDF) qui, lui, associe thésaurus et ontologie(s).
Q2. Travail avec PACTOLS
A2. Pas pour l'instant car thésaurus forgé beaucoup plus précis que PACTOLS ; les deux ont vocation à se rejoindre ultérieurement
Q3. Pas d'autre référentiel utilisable ?
A3. Voies de travail : positionner dans GeoNames + lien avec autorités géographiques d'IdRef (accessibles par exemple sur https://data.idref.fr/yasgui.html avec la requête PREFIX foaf: <http://xmlns.com/foaf/0.1/> PREFIX dbpedia-owl: <http://dbpedia.org/ontology/> select distinct ?p ?nom where {?p a dbpedia-owl:Place; foaf:name ?nom.} LIMIT 100)

- Gabriela Elgarrista. Annuaires de propriétaires de Paris : une analyse socio-économique et spatiale de la population parisienne en 1898

Source : annuaire des propriétaires de Paris et de la Seine (publication ; 1894-1937)
Deux listes : propriétaires (alphabétique) - propriétés classées par rues
Objectif :
- numériser et publier en XML-TEI
- géolocaliser : https://geo.api.gouv.fr/adresse
- réaliser des analyses à partir de ce corpus normalisé
Comparaison de la formalisation des informations : 1898-1903-1913-1923
Transcription avec Transkribus (export XML) + corrections dans Oxygen + réimport dans Transkribus pour

synchronisation avec corrections (optimisation de process : 0,27 de taux d'erreur)
Test sur 150 pages de l'édition 1898

Session 2 - TAL

- Catherine Dominguès. Annotation en lieux et sentiments de récits de vie transcrits

Mémoire collective (récits de vie) de Républicains espagnols exilés en France entre 1936 et 1939
Annotations des lieux cités/évoqués et des sentiments exprimés (lexique généraliste + lexique spécifique au corpus)
Entités nommées Lieu : noms propres et expressions plus vagues (camp de …, banlieue de …).
Plus de précisions sur ces travaux : https://journals.openedition.org/rfsic/7228?lang=en .
Problèmes de segmentation du texte oral transcrit, pas de ponctuation.
Envie de travailler sur la combinaison d'outils (symbolique et apprentissage).

- Gaël Lejeune et Yoann Dupont. Comparaisons et combinaisons d'extracteurs d'entités spatiales sur un corpus multilingue

Travail présenté ici : https://impresso.github.io/CLEF-HIPE-2020/ .
Problématique de combiner des outils de TAL sur des textes en HN qui ne sont nativement numériques, bruités,anciens.. Problèmes d'OCR, de segmentation (en pages, phrases decoupées), de tokenisation (problèmes d'espaces, tiret, ..). Les systèmes NER pour repérer les lieux doivent être adaptés.
Documents à traiter : journaux anciens numérisés.
Extracteurs évalués : Spacy, StanfordNLP, SEM.
Spacy très bon pour l'anglais mais les autres modèles sont de mauvaise qualité (et plutôt les corpus d'entraînement).
Textes de CLEF HIPE : 1798 - 2018.
Pas besoin d'entraîner, car les modèles existants devraient avoir a priori les connaissances de qu'est-ce que c'est une EN dans les corpus journalistique, mais possibilité d'adapter les modèles à moindre coût,
Problème de lire et produire les formats CLEF HIPE.
SEM : https://github.com/YoannDupont/SEM
SPACY + modèle NER allemand : https://spacy.io/models/de/
StanfordNER : https://nlp.stanford.edu/software/CRF-NER.html
4e outil en compétition, adapté par les auteurs : architecture Bi-LSTM-CRF, un réseau de neurones (LSTM) pour accumuler contexte gauche et droite (bidirectionnel).
Résultats généraux (il manque pour les lieux). De manière général, RN meilleurs que CRF (métriques), RN plus facilement adaptables que CRF. Adapter une modèle pré-entraîné améliore les résultats.
Q. adapter un modèle spacy ?
pas plus que mettre à jour les poids.
Adaptation de modèles neuronaux vers du biomédical : https://videos.univ-lorraine.fr/index.php?act=view&id=9732
Le sujet de combinaison d'outils est à suivre. Création d'un groupe de travail des membres de l'AP intéressés par la question.

Session 3 - Démos

- Mathilde Labbé. Litep et le plugin Itemrelationsnetwork : visualisation de réseaux littéraires

Visualisation de réseau littéraire : François Mauriac
Relier des monuments comme traces de réseaux de sociabilité (commémoration de la littérature dans l'espace public)
Base de données et graphe déduit (généré avec un plug-in Omeka Classic dispo dans la forge Huma-Num : Item Relation Network, basé sur le plugin Item Relations, à redévelopper pour Omeka S – voir https://daniel-km.github.io/UpgradeToOmekaS/)
Objectifs :

- - lien avec personnalités politiques
- - mettre en évidence des phénomènes d'héritages esthétiques et politiques croisés
- - étude des commanditaires de monuments

Relation entre monde littéraire et environnement social
Relation entre monument(s) et adresse administrative
Données temporelles : tous les événements de la vie d'un monument (y compris destruction éventuelle) + événements commémorés par le(s) monuments + "moment" de l'inauguration
Graphe montre que quelque(s) sculpteur (ex : Rodin) joue un rôle de "metteur en scène" d'écrivains dans l'espace public
Parti pris de représentation de l'absence de monument pour un.e écrivain.e donné.e
Beaucoup de photos et de publications de listes d'inauguration = communication publique, consentement à représentation collective pour les personnes concernées + traces de communication informelle entre ces personnes – ex : comités d'inauguration des monuments à Voltaire – mise en apparence de l'influence (dynamique sous-jacente) d'un groupe de promotion de la libre-pensée.
L'outil est un plugin Omeka et disponible dans la forge huma-num. Lien à venir

- Alberto Dalla Rosa et Vincent Razanajao. Reconstruire la géographie du patrimoine des empereurs romains : édition, annotation et exploration des données spatiales dans la plateforme PATRIMONIVM

PATRIMONIVM : localisation du patrimoine des empereurs romains (ERC)
Info: https://patrimonium.huma-num.fr
Sources littéraires + numismatiques + archéologiques
Sujet de recherche : patrimoine immobilier, mais aussi esclaves, argent, etc. (patrimoine mobilier) + colons, etc. = sujets liés à la propriété des empereurs
Question spécifique pour la géolocalisation : absente de références de type cadastral
ex. de source archéo à transcrire : bornes de domaines ou de district, inscription funéraire
Sur ces sources la question se pose de traduire de localisations relatives décrites en langu naturel en coordonnées (au moins de manière approximative), aussi c'est le cas pour les travaux d'Emmanuelle Perrin.
Pour le moment, ces informations sont encodées manuellement dans les sources numériques.
Cartographie inclut une représentation symbolique de l'incertitude relative d'une localisation – l'incertitude est formalisée, quand c'est nécessaire, en rattachant l'entité à un lieu plus vaste, connu et certain (par ex. une

province, ou au pire "l'empire romain")

Outils ouverts mais pas encore publiés (contacter les porteurs de projet)

- Ludovic Moncla et Katherine McDonough. Reconnaissance d’entités nommées et géoparsing appliqués à l’Encyclopédie

Geoparsing sur articles de l'Encyclopédie
Accès au tutoriel : Notebook - https://github.com/GEODE-project/perdido-geoparsing-notebook
Utilisation des services web (API REST) de l'outil d'annotation PERDIDO (http://erig.univ-pau.fr/PERDIDO/) pour le prétraitement (annotation morpho-syntaxique avec Treetagger).
Utilisation d'une version customisée de PERDIDO pour l'Encyclopédie
Chargement de données issues de fichiers TEI vers un dataframe Python, visualisation graphique des informations géo-sémantiques annotés en TEI grâce à la librairie displaCy.
Geocoding et affichage des entités de lieux sur une carte
Illustration des problèmes de geotagging et geocoding lié à l'analyse de documents historiques.

Discussions

Problématiques communes (?)

- Combinaisons des outils TAL, symboliques et apprentissage (RN, CRF), notamment pour le repérage de noms de lieux mais aussi d'autres tâches TAL associées.
- Adaptation, prise en compte de la difficulté des textes en HN (oraux, bruités par l'océrisation ,..).
- Difficulté de désambiguïser les noms de lieux, manque de référentiels (historiques).
- Traduire de localisations relatives et spatiales décrites dans les sources en langue naturel en coordonnées, au moins de manière approximative
- Visualisation de grands masses de données littéraires, historiques, (réseaux, diagrammes, SIG)
- ...

Prochaines actions

- Organisation de l'édition 2020 de l'atelier Géospatial Humanities associé à la conférence ACM SIGSPATIAL (3 novembre 2020 à Seattle pas d'info liées au Covid pour le moment) : https://ludovicmoncla.github.io/sigspatial-geohumanities-2020/
- Organisation d'un atelier Humanités Numériques Spatialisées lors de la conférence SAGEO 2021 (5-7 mai 2021) à La Rochelle
- ...