Géosciences pour une Terre durable
brgm

LaSTI
Lab on Geographic Information Science for sustainable development and smart cities

Annotation sémantique pour la géolocalisation d'entités spatiales dans des tweets

G. Caillaut, C. Gracianne, S. Auclair, N. Abadie et G. Touya

Juin 2022

Introduction

Contexte
ANR RéSoCIO

Développer des outils pour exploiter automatiquement des données issues des réseaux sociaux lors de catastrophes naturelles à cinétique rapide :

- Séismes
- Crues éclairs

Contexte
ANR RéSoCIO

Développer des outils pour exploiter automatiquement des données issues des réseaux sociaux lors de catastrophes naturelles à cinétique rapide :

- Séismes
- Crues éclairs

Dans le cadre de ce travail, on cherche à géolocaliser ces événements

- le plus vite possible
- le plus précisément possible

Contexte

Les réseaux sociaux

Des millions de personnes utilisent les réseaux sociaux et produisent un flux d’informations en temps réel.

Contexte

Les réseaux sociaux

Des millions de personnes utilisent les réseaux sociaux et produisent un flux d'informations en temps réel.

Lors de catastrophes naturelles, les gens sont plus enclins à fournir des informations spatiales (Grace 2021).

- Les témoins peuvent partager des adresses, des noms de routes ou des sorties d'autoroute (Hu et Wang 2020).
- Les autres peuvent relayer des informations plus grossières, comme des départements ou des villes.

Contexte

Les réseaux sociaux

Des millions de personnes utilisent les réseaux sociaux et produisent un flux d'informations en temps réel.

Lors de catastrophes naturelles, les gens sont plus enclins à fournir des informations spatiales (Grace 2021).

- Les témoins peuvent partager des adresses, des noms de routes ou des sorties d'autoroute (Hu et Wang 2020).
- Les autres peuvent relayer des informations plus grossières, comme des départements ou des villes.

Mais **moins de 1 %** partagent leur géolocalisation !

Context
Les objectifs

Automatiser les processus

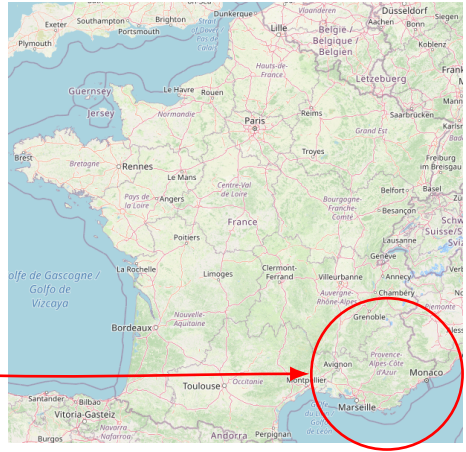
Beaucoup d'actions peuvent (doivent) être automatisée :

- surveillance des réseaux sociaux (en particulier Twitter)
- identification des messages pertinents (tremblements de terre ou innondations)
- extraction d'informations (coordonnées GPS entre autres)
- partage des informations aux équipes de secours et aux autorités compétentes

Pour le moment, on ne s'intéresse qu'aux évènement ayant lieu en France.

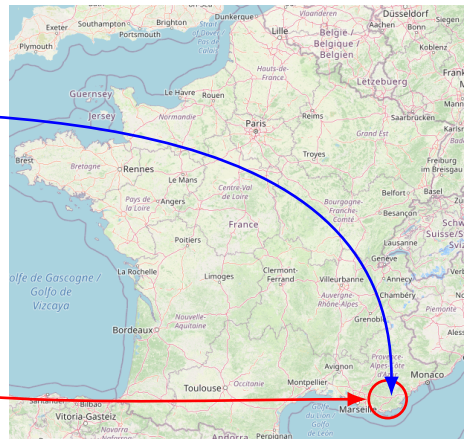
Cas d’utilisation et verrous

Exemple 1



Cas d’utilisation et verrous

Exemple 2



Cas d’utilisation et verrous

Exemple 3

🥳 Séisme dans l’est de la Vendée

⚠️ Équateur

⚠️ « crue » : verbe ou nom ?

🥳 Innondations…

⚠️ …mais en Thaïlande

L’approche proposée

Représentation des entités spatiales

Les systèmes de géolocalisation automatique reposent sur deux étapes :

1. Détection des mentions d'entités spatiales dans le texte
2. Géolocalisation de ces entités

Régression prédiction de coordonnées GPS

Classification discrétisation de la zone étudiée puis classification

Représentation des entités spatiales

Les systèmes de géolocalisation automatique reposent sur deux étapes :

1. Détection des mentions d'entités spatiales dans le texte
2. Géolocalisation de ces entités

   Régression prédiction de coordonnées GPS

   Classification discrétisation de la zone étudiée puis classification

Mais il existe différents type d'entités spatiales :

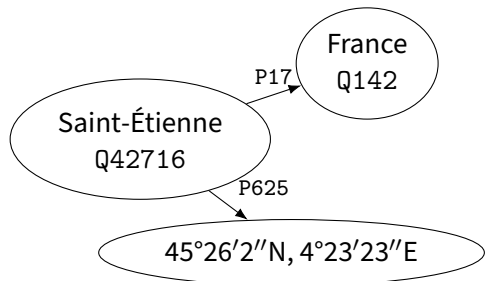
- adresses, bâtiments : point
- départements, villes : polygone
- routes, cours d'eau : ligne

Uniformiser les représentations des entités spatiales

De nombreuses entités spatiales sont référencées dans des bases de connaissances, comme **Wikidata**.

Wikidata structure les connaissances sous la forme d'une ontologie, dans laquelle chaque entité :

- est référencée par un identifiant, le *QID*.
  - Le QID de Saint-Étienne est `Q42716`
- correspond à un article Wikipédia (mais pas forcément en français)
- contient une liste de propriétés.
  - `P625` est la propriété *coordonnées géographiques*



Extractions d'informations géographiques depuis Wikidata

Notre proposition :

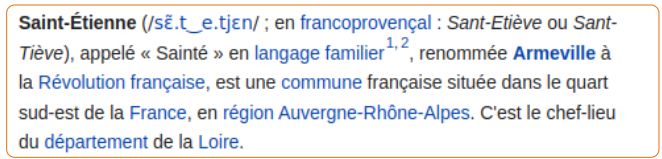
1. Détecter les mentions d'entités spatiales dans un texte
2. Récupérer les entités Wikidata correspondantes
3. Extraire les coordonnées GPS depuis les entités Wikidata

On applique la tâche d'Annotation Sémantique (*Entity Linking*) au problème de géolocalisation d'entités spatiales dans un texte.

Combiner Wikipédia et Wikidata

Wikipédia est très largement utilisé pour entraîner des systèmes d'Annotation Sémantique.

- Corpus massif, multilingue et de « qualité »
- Les mentions d'entités sont annotées par des liens
- En évolution constante
- « Quasi-bijection » entre Wikidata et Wikipedia


Saint-Étienne (/sɛ̃.t_e.tjɛn/ ; en francoprovençal : *Sant-Etiève* ou *Sant-Tiève*), appelé « Sainté » en langage familier[1, 2], renommée **Armeville** à la Révolution française, est une commune française située dans le quart sud-est de la France, en région Auvergne-Rhône-Alpes. C'est le chef-lieu du département de la Loire.

Wikipédia est un (le ?) corpus idéal pour l'entraînement de systèmes d'Annotation Sémantique.

Le jeu de données

Construire un corpus Wikipédia français

Wikipédia maintiens une liste de *bons articles* et d'*articles de qualité*.

Construction du corpus :

1. Initaliser le mécanisme de *scrapping* en récupérant les *bons articles* et les *articles de qualité*
2. Télécharger les articles référencés dans dans les *bons articles*
3. Nettoyer les pages HTML
    - suppression de ce qui est « hors contenu »
    - suppression des tableaux, images
    - suppression de certaines sections (références, voir aussi, …)

Scripts disponibles sur `https://github.com/GaaH/frwiki_good_pages_el`.
Jeu de données disponible sur `https://huggingface.co/datasets/gcaillaut/frwiki_good_pages_el`.

Description du jeu de données

Chaque document est défini par :

title Le titre de la page Wikipédia

qid L'identifiant Wikidata

words Les mots

labels Annotation `BIO`

qids Identifiants Wikidata correspondant aux entités référencées

words	Saint-Étienne	est	une	commune	de	la	région	Auvergne	Rhône	Alpes
labels	B	O	O	B	O	O	B	B	I	I
labels	Q42716			Q484170			Q36784	Q18338206		

Nous avons extraits un second jeu de données centré sur les entités :

qid L'identifiant Wikidata

description Le premier paragraphe de la page Wikipédia correspondante

Résumé du jeu de données

Documents 6023

Entités distinctes 304 826

Mentions d'entités 1 619 961

Le système

L’architecture à double encodeur

On s’inspire de l’architecture à double encodeur proposée par BOTHA, SHAN et GILLICK (2020).

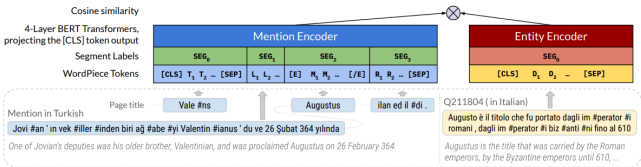


Figure 1 : L’architecture à double encodeur.

L’architecture à double encodeur

On s’inspire de l’architecture à double encodeur proposée par BOTHA, SHAN et GILLICK (2020).

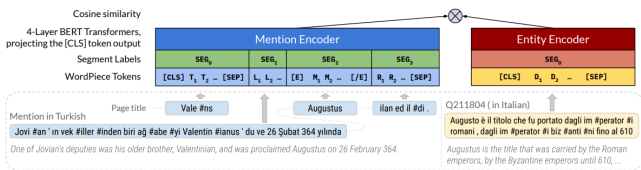


Figure 1 : L’architecture à double encodeur.

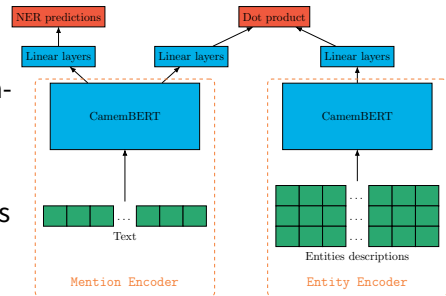
Inconvénients

- Les entités doivent être annotées
- Les entités sont traitées une par une
- Nécessite 2 modèles BERT lors de l’entraînement (voire 3, pour l’annotation des mentions)

Notre nouvelle architecture à double encodeur

On modifie le *Mention Encoder* de manière à, simultanément :

- Détecter **toutes** les mentions d'entités
- Calculer les représentations de **toutes** ces entités



Notre système détecte **toutes** les entités présentes dans le texte et en calcule des représentations en **une seule étape**.

Utilisation du système pour la prédiction

Entity Encoder :

- Calcul des représentations pour toutes les entités cibles

Utilisation du système pour la prédiction

Entity Encoder :

- Calcul des représentations pour toutes les entités cibles

Mention Encoder :

- Détection des entités
- Calcul de représentations pour ces entités
- Comparaison des représentations avec celles calculées par l'Entity Encoder

Résultats préliminaires

Tâches et jeu de données

Deux tâches :

Détection des Mentions (MD) Détecter les mentions d'entités dans un texte

Entity Linking (EL) Lier les mentions aux entités d'une base de connaissances (Wikidata)

Deux jeux de données :

`frwiki EL` Notre jeu de données issu de Wikipédia. Dédié aux tâches MD et EL.

`CAp + Wikiner` Union des corpus CAp 2017 et WikinerFR. Dédié uniquement à la tâche MD.

- CAp 2017 est un corpus de tweets français (Lopez et al. 2017)
- Wikiner provient de Wikipédia (Nothman et al. 2013)

Training

Trois systèmes :

- **MD** Transformer spécialisé sur `frwiki EL` sur la tâche MD
- **EL** Double Encodeur « nature » entraîné sur `frwiki EL` sur la tâche EL
- **MD + EL** Notre Double Encodeur modifié entraîné sur `frwiki EL` sur la double tâche MD + EL

Évaluation

Tâche détection de mentions

Systèmes	Micro Fscore
MD	0,77
MD + EL	0,79

Labels	MD	MD + EL
B	0,65	0,66
I	0,77	0,75
O	0,78	0,81
Global	0,77	0,79

- L'entraînement sur la tâche EL semble avoir peu d'impact sur les performances en MD

Evaluation
Entity Linking task

Systems	R@1	R@5	R@10	R@100
EL	0,31	0,56	0,67	0,89
MD + EL	0,84	0,91	0,93	0,97

- L'entraînement sur la tâche MD semble mieux orienter le modèle sur la tâche EL
- Mais les scores ne sont calculés que sur les mentions détectées ! ($\approx$66 % des mentions)

Conclusion

Conclusion
Jeu de données

Notre contribution :

- Un corpus français dédié à la double tâche MD et EL
- Un ensemble de scripts pour construire et mettre à jour le corpus

Perspectives :

- Agrandir le corpus
- S'appuyer sur les dumps XML au lieu des pages HTML
- Typer les mentions en s'appuyant sur les propriétés Wikidata ou les descriptions Wikipédia

Conclusion
Modèle

Notre contribution :

- Une amélioration du système à double encodeur permettant d'accélérer la prise de décision
  - détection de **toutes** les mentions d'entités
  - liaison de **toutes** les mentions sur Wikidata

} Simultanément

Perspectives

- Améliorer les performances EL en filtrant/ordonnant les candidats potentiels
- Spécialiser le modèle sur des données issues de réseaux sociaux
  - possible collaboration avec le SDIS 06 (Alpes-Maritimes) pour une collecte de données
  - Campagne d'annotations de tweets

Bibliographie I

- Botha, Jan A, Zifei Shan et Daniel Gillick (2020). « Entity linking in 100 languages ». In : *arXiv preprint arXiv:2011.02690*.
- Grace, Rob (2021). « Toponym usage in social media in emergencies ». In : *International Journal of Disaster Risk Reduction* 52, p. 101923.
- Hu, Yingjie et Jimin Wang (2020). « How do people describe locations during a natural disaster : an analysis of tweets from Hurricane Harvey ». In : *arXiv preprint arXiv:2009.12914*.
- Lopez, Cédric et al. (2017). « Cap 2017 challenge : Twitter named entity recognition ». In : *arXiv preprint arXiv:1707.07568*.
- Nothman, Joel et al. (2013). « Learning multilingual named entity recognition from Wikipedia ». In : *Artificial Intelligence* 194, p. 151-175.