Université Claude Bernard Lyon 1

Projet de recherche

Reconnaissance d'entités géographiques dans des documents textuels

Cahier des charges

Réalisé par :
BROU Jonathan 11413381

Encadrant :
DUCHATEAU Fabien

1. Présentation du projet

Ce projet s'inscrit dans le cadre de l'UE Mif20 « Projet de recherche », et porte sur la reconnaissance d'entités géographiques dans des documents textuels. Le Web est devenu une source riche en informations, seulement, celles-ci ne sont souvent pas structurées et sont uniquement présentes sous forme textuelle. Ces documents textuels sont alors difficilement exploitables et nous avons besoin de la reconnaissance d'entités (Entity Linking, également appelé Named Entity Disambiguation (NED)). Celle-ci consiste à lier une mention dans un document textuel à une entité dans un référentiel. Dans ce projet nous nous intéresserons uniquement à la reconnaissance d'entités géographiques à partir de référentiels comme Geonames [5] et DBPedia [6]. Ainsi, la phrase « la basilique de Fourvière se trouve à Lyon » va créer une relation entre les entités dont les mentions sont « basilique de Fourvière » et « Lyon ». Une difficulté du projet sera de savoir reconnaître la bonne entité, car une même mention peut correspondre à plusieurs entités mais également que plusieurs mentions peuvent correspondre à une seule entité. Par exemple, la mention « Michael Jordan » peut correspondre à plusieurs entités (le joueur de la NBA, le footballeur, le scientifique…) mais également les mentions « Hewlett-Packard » ou « HP » peuvent correspondre à la même entité (la société américaine). Il conviendra alors d'analyser le contexte à l'aide de différentes approches et de trier les entités candidates selon leur pertinence.

Exemple de fonctionnement de l'Entity Linking

Nous voulons reconnaître les entités spatiales à l'intérieur de ce texte :

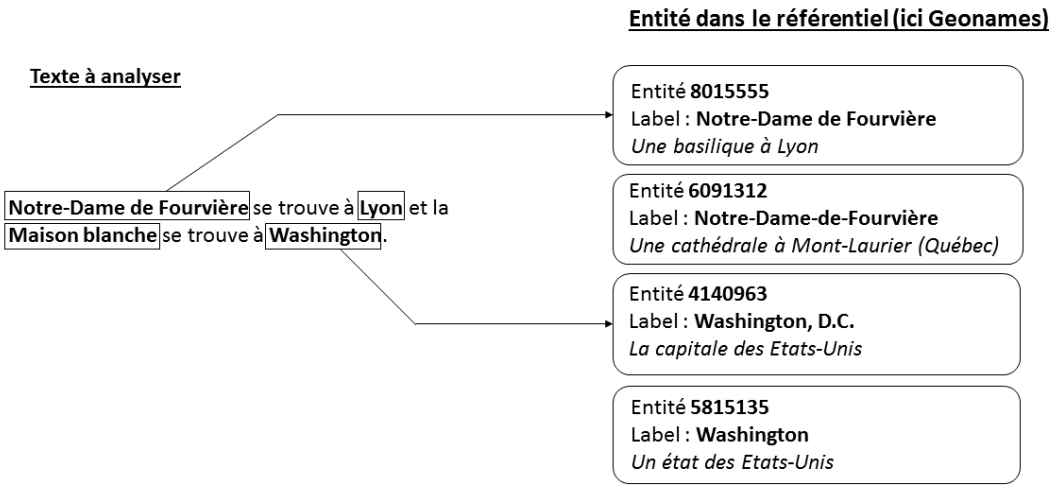

Figure 1 : Processus d'Entity Linking

Le processus relie les mentions reconnues dans le texte à leur entité correspondante dans le référentiel. Dans cet exemple, les mentions « Notre-Dame de Fourvière » et « Washington » correspondent à 2 entités différentes dans le référentiel. Le but sera de retrouver celle qui correspond le mieux en fonction du contexte. Ici, les mentions « Lyon » et « maison blanche », peuvent nous aider à déduire la mention qui correspond à « Notre-Dame de Fourvière » de Lyon et à la ville de Wahington.

2. Réalisation du projet et technologies utilisées

Dans ce projet nous utiliserons la base de données Open Source Geonames. Cette base est constituée d'informations sur des lieux comme son nom, sa position géographique… mais certaines données comme la ville peuvent manquer, le projet aura pour but de déduire ces données manquantes à partir d'un texte.

Nous créerons alors une interface Web en utilisant un serveur Tomcat et peut-être nodejs, mais son apprentissage devra être considéré. Le projet sera développé en Java côté serveur, et en HTML5, CSS3 et Javascript côté client. Nous utiliserons l'outil Apache Maven pour la gestion de la production du projet.

Nous utiliserons une base de données relationnelle en locale (PostreSQL, SQLite ou MySQL) dans laquelle nous aurons importé les dumps de Geonames disponible au format CSV.

Nous devrons implémenter au moins 2 approches qui sont décrites dans un article sur Entity Linking [1].

Notre interface devra permettre de comparer nos différentes approches avec d'autres outils déjà existants tels qu'Alchemy API [2], TextRazor [3], GéoPot ou DBPedia Spotlight [5]. Ils fournissent tous une API qui nous permettra de les interroger pour récupérer les entités reconnues.

Voici une maquette de l'interface :

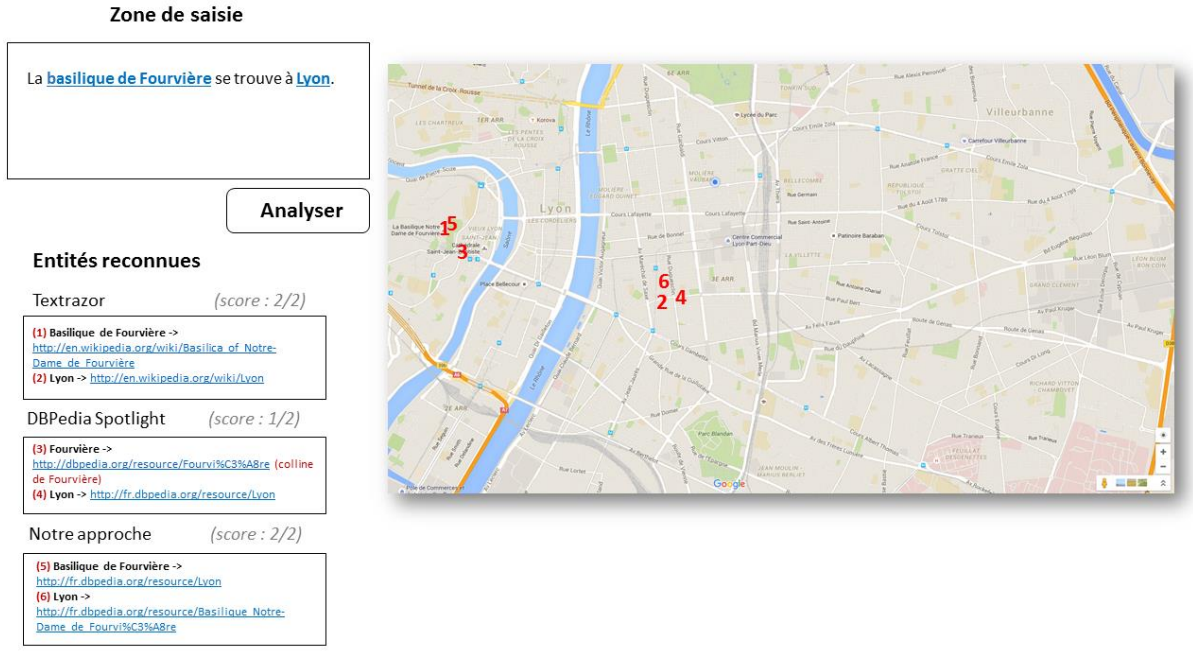

Figure 2 : Maquette de l'interface

L'interface nous permettra de saisir un texte et de l'analyser pour lier chaque mention à son entité correspondante dans le référentiel. Le résultat sera alors une liste d'entités reconnues à partir des différents outils et de notre approche. Chaque entité spatiale reconnus pourra

s'afficher sur une carte Google Maps. L'utilisateur pourra alors comparer les différents outils avec un score correspondant à la qualité des résultats obtenus par rapport à ceux attendus. Ici, DBPedia Spotlight n'obtient pas le meilleur score car il n'a pas reconnu l'entité « basilique de Fourvière » mais la colline de Fourvière. Notre but sera de développer une approche qui obtienne de meilleurs résultats que les outils existants. On pourra ajouter une fonctionnalité d'export des entités reconnus en JSON.

3. Livrables

- Le présent cahier des charges.
- L'implémentation des différentes approches.
- Une interface Web permettant à l'utilisateur de saisir un texte pour y visualiser les entités sur une carte ainsi qu'un lien vers leur entité sur Geonames[5] ou DBPedia[6].
- Une comparaison des différentes approches et des outils existants.
- Le rapport final détaillant le travail réalisé pendant ce TER.
- Création d'un benchmark (datasets + métriques) pour évaluer/comparer les outils existants et les approches implémentées.

4. Déroulement du projet

Dates	**Tâches à effectuer**
18-20 Janvier	Rédaction du cahier des charges
21-31 Janvier	Etude des outils existants (TextRazor, Spotlight…) et lecture de l'article « survey » sur l'Entity Linking [1]
1-14 Février	Implémentation de 2 approches du survey
15-17 Février	Réalisation d'une interface graphique pour qu'un utilisateur puisse tester les différents outils.
18-25 Février	Création d'un benchmark (datasets + métriques) pour évaluer/comparer les outils existants et les approches implémentées.
26 Février-2 Mars	Préparation des livrables et de la soutenance
3-4 Mars	Soutenance

5. Références

[1] W. Shen, H. Wang et J.Han. "Entity linking with a knowledge base: Issues, techniques, and solutions."
[2] http://www.alchemyapi.com/
[3] https://www.textrazor.com/
[4] https://dbpedia-spotlight.github.io/demo/
[5] http://www.geonames.org/
[6] http://fr.dbpedia.org/