ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Virgile Reignier
licencié.e ès histoire
diplômé.e de master mondes médiévaux

Vers l'indexation automatique du Trésor des chartes

Constitution, alignement et utilisation de référentiels d'entités nommées au sein du projet Himanis

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2022

Résumé

Blablabla résumé du mémoire.

Mots-clés : TNAH, IRHT, Himanis, Trèsor des Chartes, Archives Nationales, OCR, Numérisation d'instruments de recherche, Alignement de référentiels, HTR, REN, Machine learning, Intelligence artificielle, Reconnaissance d'écriture manuscrite, identity linking.

Informations bibliographiques : Reignier Virgile, *Vers l'indexation automatique du Trésor des chartes. Constitution, alignement et utilisation de référentiels d'entités nommées au sein du projet Himanis.*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Dominique Stutzmann et Thibault Clérice, École nationale des chartes, 2022.

Remerciements

Blablabla remerciements...
Faire une liste des abréviations avec : REN, TAL, HTR, IRHT

Bibliographie

Agirre (Eneko), Barrena (Ander), Lacalle (Oier Lopez de), Soroa (Aitor), Fernando (Samuel) et Stevenson (Mark), "Matching Cultural Heritage items to Wikipedia", dans *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012, p. 1729-1735, URL : http://www.lrec-conf.org/proceedings/lrec2012/pdf/1021_Paper.pdf (visité le 21/07/2022).

BVMM, URL : https://bvmm.irht.cnrs.fr/ (visité le 16/07/2022).

Ehrmann (Maud), *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*, These de doctorat, Paris 7, 2008, URL : https://hal.archives-ouvertes.fr/tel-01639190/document (visité le 14/04/2022).

Himanis - Chancery Indexing and Search, URL : http://himanis.huma-num.fr/app// (visité le 16/07/2022).

Holtz (Louis), "Les premières années de l'Institut de recherche et d'histoire des textes", *La revue pour l'histoire du CNRS*–2 (5 mai 2000), ISBN : 9782271057082 Number : 2 Publisher : CNRS Éditions, DOI : 10.4000/histoire-cnrs.2742.

Linhares Pontes (Elvys), Moreno (Jose G.) et Doucet (Antoine), "Linking Named Entities across Languages using Multilingual Word Embeddings", dans *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, New York, NY, USA, 2020, p. 329-332, URL : https://doi.org/10.1145/3383583.3398597 (visité le 23/07/2022).

Potin (Yann), *La mise en archives du trésor des chartes (XIIIe-XIXe siècle)*, Positions de thèse pour le diplôme d'archiviste-paléographe, Paris, Ecole nationale des chartes, 2007, URL : http://theses.enc.sorbonne.fr/2007/potin (visité le 16/07/2022).

Rijhwani (Shruti), Xie (Jiateng), Neubig (Graham) et Carbonell (Jaime), "Zero-Shot Neural Transfer for Cross-Lingual Entity Linking", *Proceedings of the AAAI Conference on Artificial Intelligence*, 33–1 (17 juill. 2019), Number : 01, p. 6924-6931, DOI : 10.1609/aaai.v33i01.33016924.

Soudani (Aicha), Meherzi (Yosra), Bouhafs (Asma), Frontini (Francesca), Brando (Carmen), Dupont (Yoann) et Mélanie-Becquet (Frédérique), "Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle", dans *Atelier Humanités Nu-*

mériques Spatialisées (HumaNS'2018), Montpellier, France, 2018, URL : https://hal.archives-ouvertes.fr/hal-01925816 (visité le 21/07/2022).

STERN (Rosa), *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Thèse de doctorat, Université Paris-Diderot - Paris VII, 2013, URL : https://tel.archives-ouvertes.fr/tel-00939420 (visité le 28/03/2022).

STUTZMANN (Dominique), MOUFFLET (Jean-François) et HAMEL (Sébastien), "La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique", *Médiévales. Langues, Textes, Histoire*, 73–73 (15 déc. 2017), p. 67-96, DOI : 10.4000/medievales.8198.

TORRES AGUILAR (Sergio) et STUTZMANN (Dominique), "Named Entity Recognition for French medieval charters", dans *Workshop on Natural Language Processing for Digital Humanities*, Helsinki, Finland, 2021 (Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop), URL : https://hal.archives-ouvertes.fr/hal-03503055 (visité le 17/07/2022).

ZHOU (Shuyan), RIJHWANI (Shruti) et NEUBIG (Graham), "Towards Zero-resource Cross-lingual Entity Linking", dans *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Hong Kong, China, 2019, p. 243-252, DOI : 10.18653/v1/D19-6127.

Introduction

> A ce sujet papa avait une plaisanterie. (...) Il disait, quand il présentait maman, « je l'ai connue et épousée à Paris » et (...) il attendait avant de dire « Texas » que tout le monde ait cru, que tout le monde ait pensé qu'il parlait de Paris, France. Ça faisait tordre de rire toutes les fois.

Si le mot "Paris" évoque en premier lieu la capitale française, il désigne également d'autres villes à travers le monde. C'est en exploitant l'homonymie entre cette première et une ville du Texas que la citation ci-dessus, extraite du film "Paris, Texas" (1984) de Wim Wenders, construit la plaisanterie. L'information "Paris" ne suffit en effet pas à identifier le lieu où lesdits parents se sont rencontrés. Utilisé seul, le mot est naturellement associé à la France. C'est seulement en précisant l'État dans lequel elle se situe que l'on peut identifier le lieu exact où les protagonistes se sont rencontrés et mariés. Ce jeu d'ambiguïté manifeste ainsi d'une difficulté rencontrée dans le langage naturel : l'identification des références utilisées. La connaissance lexicale ne suffit en effet pas à elle seule pour comprendre un discours, il faut également que les références soit comprises et associées à une réalité clairement identifiée.

Cet enjeu est également présent au sein du Traitement Automatique des Langues à travers la notion d'Entité Nommée qui désigne une expression linguistique qui se réfère à une entité unique de façon autonome [1]. L'analyse du contenu textuel a ainsi largement progressé ces dernières années autour de cette notion par le développement de deux techniques : la REN (Reconnaissance d'Entités Nommées) qui consiste à repérer ces objets textuels et à leur attribuer une catégorie et le liage d'entités qui permet d'associer ces objets textuels à un élément décrit par une ressource référentielle. Si un grand nombre de ces travaux concernent des corpus contemporains, quelques chercheurs s'intéressent également à leur application pour la lecture des archives anciennes et rencontrent ainsi les recherches menées par les spécialistes de ces corpus.

1. Sur la définition des entités nommées, *cf.* Maud Ehrmann, *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*, These de doctorat, Paris 7, 2008, URL : `https://hal.archives-ouvertes.fr/tel-01639190/document` (visité le 14/04/2022), p. 167-170.

Contexte scientifique de travail

L'Institut de Recherche et d'Histoire des Textes (IRHT) est un laboratoire de recherche fondé en 1937 par Félix Grat et rattaché au CNRS dans le but de faciliter l'accès des chercheurs aux manuscrits et imprimés anciens [2]. Les recherches qui y sont menées portent également sur la transmission des textes et l'étude des écritures et connaissent à ce titre des développements récents à propos de la lecture automatique des documents anciens. Initiés par sa collaboration au sein du projet GRAPHEM, les travaux en "paléographie artificielle" développés par la section de paléographie latine sont menés conjointement avec des chercheurs en informatique spécialisés dans l'analyse de l'image. Les projets développés prennent deux directions principales : la caractérisations des écritures médiévales (Oriflamms, ECMEN, CrEMe) d'une part et la lecture automatique des archives (Himanis, HOME, HORAE) d'autre part. Pilotés par Dominique Stutzmann, ces recherches ont permis le développement d'outils informatiques et de modèles d'intelligence artificielle qui ont largement renouvelé l'accès aux textes anciens.

Parmi les corpus étudiés par ces travaux, le Trésor des Chartes occupe une place centrale puisqu'il constitue le matériel source du projet Himanis et participe à celui du projet HOME. Conservé au sein de la série JJ des Archives Nationales, ce fonds se compose d'une immense collection de titres rassemblée par les rois de France. Il se présente sous la forme de registres contenant des actes organisés de manière plus ou moins systématique et linéaire [3]. Le projet Himanis (HIstorical MANuscript Indexing for user-controlled Search) a ainsi permis de numériser les registres et de convertir les inventaires et éditions disponibles afin de les structurer en un format homogène et unique [4]. Ces éléments ont ensuite servi de base au développement d'un modèle d'indexation automatique des mots présents dans le corpus [5]. Par la suite, le projet HOME (History of Medieval Europe) s'est proposé d'amplifier et de généraliser ce travail en numérisant de nouveaux documents, en associant chaque texte aux données disponibles les concernant et en déposant les résultats dans une plateforme librement accessible [6].

2. Sur la fondation de l'IRHT, *cf.* Louis Holtz, "Les premières années de l'Institut de recherche et d'histoire des textes", *La revue pour l'histoire du CNRS*–2 (5 mai 2000), ISBN : 9782271057082 Number : 2 Publisher : CNRS Éditions, DOI : 10.4000/histoire-cnrs.2742.

3. Sur la constitution du trésor des chartes, *cf.* Yann Potin, *La mise en archives du trésor des chartes (XIIIe-XIXe siècle)*, Positions de thèse pour le diplôme d'archiviste-paléographe, Paris, Ecole nationale des chartes, 2007, URL : http://theses.enc.sorbonne.fr/2007/potin (visité le 16/07/2022)

4. Les registres numérisés ont été intégrés à la Bibliothèque Virtuelle des Manuscrits Médiévaux *BVMM*, URL : https://bvmm.irht.cnrs.fr/ (visité le 16/07/2022). Tous les fichiers issus de ces travaux sont disponibles ici : https://github.com/oriflamms/himanis.

5. Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, "La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique", *Médiévales. Langues, Textes, Histoire*, 73–73 (15 déc. 2017), p. 67-96, DOI : 10.4000/medievales.8198. Les résultats sont disponibles dans l'interface *Himanis - Chancery Indexing and Search*, URL : http://himanis.huma-num.fr/app// (visité le 16/07/2022).

6. https://github.com/oriflamms/Home

Problématique du stage

Ces différents travaux ont ainsi permis de diffuser largement les textes qui composent le Trésor des chartes et de progresser dans l'analyse automatique des écritures qu'ils contiennent. Il reste néanmoins une problématique à approfondir : l'identification des références utilisées au sein des documents. Si les travaux réalisés permettent de faciliter la lecture des textes, cette dernière se trouve encore freinée par la difficile compréhension des références utilisées. Après des travaux récents portant sur la REN dans les chartes médiévales [7], l'objectif poursuivi est de parvenir à développer un modèle de liage d'entités afin d'enrichir et de désambiguïser les entités nommées reconnues dans les textes.

C'est dans ce contexte que mon stage, effectué dans le cadre du Master 2 Archives - Technologies Numériques Appliquées à l'Histoire de l'Ecole Nationale des Chartes, s'est donné pour mission de rassembler les éléments disponibles au sein du corpus Himanis pour avancer sur la problématique de l'identification des entités nommées. A partir des inventaires déjà convertis, des registres numérisés et des travaux préliminaires en HTR et REN, nous avons ainsi travaillé sur la construction d'un référentiel et d'une méthode de travail pour lier les entités nommées reconnues en limitant au maximum les ambiguïtés possibles. Le présent mémoire se propose donc de décrire les travaux effectués et la manière dont ils s'insèrent dans un contexte de travail. Quels sont les apports des données fournies par les projets Himanis et HOME pour apprendre à désambiguïser automatiquement les entités nommées reconnues dans un texte médiéval ? Nous aborderons les différentes étapes de construction du référentiel ainsi que les difficultés rencontrées dans ce cadre et dans son utilisation.

Dans cet objectif, nous exposerons dans une première partie le matériel disponible pour mettre en œuvre ce projet. Nous proposerons ainsi un état des lieux sur les recherches en cours à propos du liage d'entités, puis nous décrirons plus précisément les avancées permises par le projet Himanis dans l'accès au corpus du Trésor des Chartes, enfin nous analyserons l'apport des instruments de recherches convertis sous format numérique. Notre deuxième partie sera consacrée à la formalisation du référentiel. Nous développerons pour cela les différents enjeux liés à l'utilisation d'un instrument papier, puis nous proposerons une analyse du lien entre les entités décrites, enfin nous décrirons l'insertion des éléments dans une base de données relationnelle. Notre troisième et dernière partie se portera sur les différents traitements mis en œuvre afin de compléter et diffuser ce référentiel. Nous décrirons ainsi l'enrichissement des données à partir de référentiels externes, puis la mise à disposition du référentiel et enfin les premiers pas de son utilisation.

7. Sergio Torres Aguilar et Dominique Stutzmann, "Named Entity Recognition for French medieval charters", dans *Workshop on Natural Language Processing for Digital Humanities*, Helsinki, Finland, 2021 (Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop), URL : `https://hal.archives-ouvertes.fr/hal-03503055` (visité le 17/07/2022).

Première partie

De la *legacy data* au liage d'entité : quel matériel disponible pour entraîner un modèle ?

Avant d'aborder plus précisément les actions menées au cours de ce stage, il convient d'exposer dans cette première partie les différents éléments contextuels dans lequel il s'inscrit. Nous consacrerons donc un premier chapitre à la description des enjeux scientifiques actuels autour de la problématique du liage d'entité afin de mieux appréhender les perspectives d'évolution. Un second chapitre permettra de résumer les différents résultats offerts par le projet Himanis et leur utilisation possible dans le cadre du stage. Enfin, un troisième chapitre permettra d'envisager les différents moyens possibles pour utiliser des instruments de recherches papier afin de construire un référentiel numérique.

Chapitre 1

État des lieux de la recherche sur le liage d'entités

Initiée par les *Message Understanding Conferences* qui se réunissent entre 1987 et 1998, la REN est directement associée aux techniques d'extractions d'informations. L'objectif est en effet d'automatiser la lecture des textes afin d'en comprendre au mieux la substance. Reconnaître et classifier les références utilisées prend donc dans ce contexte une place centrale qui se perpétue par la suite dans de nombreuses recherches [1]. Dans un objectif similaire, d'autres travaux portant sur l'annotation sémantique des textes, c'est à dire l'enrichissement des contenus textuels à partir de métadonnées, ont mis en valeur la nécessité de construire un lien entre les entités nommées reconnues dans le texte et un référentiel à disposition dans ce but [2].

C'est dans ce contexte qu'est née le principe du Liage d'entité. Il se définit comme une technique permettant d'associer chaque élément reconnu comme devant être expliqué à un nœud d'une base de connaissances permettant la génération de ladite explication. La conception de cette technique procède donc de deux éléments : la construction d'une base de connaissances utilisée comme référence et la reconnaissance des entités à mettre en lien avec cette base. Son enjeu principal est de permettre la résolution des ambiguïtés qui peuvent exister entre les entités, soit parce qu'un même mot peut renvoyer vers plusieurs entrées (polysémie), soit au contraire parce qu'une même entité peut s'exprimer de plusieurs façon différentes (synonymie) [3].

Nous tenterons donc dans ce chapitre d'exposer succinctement l'état de l'art autour des problématiques associées au liage d'entité. Pour cela, nous décrirons dans un premier temps son fonctionnement général puis son application aux sources historiques. Nous

1. Maud Ehrmann, *Les entités nommées, de la linguistique au TAL…*, p. 17-19

2. Sur les enjeux de l'Annotation Sémantique, lire Rosa Stern, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Thèse de doctorat, Université Paris-Diderot - Paris VII, 2013, URL : `https://tel.archives-ouvertes.fr/tel-00939420` (visité le 28/03/2022), p. 15-16. Sur sa mise en œuvre, *Ibid.*, p. 96-99.

3. *Ibid.*, p. 110-114.

proposerons une analyse des différents outils disponibles et de leurs apports. Enfin, nous décrirons les différents travaux en cours et les perspectives d'amélioration.

1.1 Mise en œuvre du liage d'entité

1.1.1 Méthodologie

Une méthode utilisée naturellement pour résoudre les ambiguïtés est de considérer que ces entités se rapportent a priori à leur sens par défaut, défini généralement en fonction de sa fréquence d'apparition. Si on en revient à l'exemple utilisé en introduction, le fait de savoir qu'il existe plusieurs "Paris" à travers le monde ne dispense pas de penser que la phrase "je l'ai connue et épousée à Paris" renvoi par défaut vers la capitale française, et ce même lorsqu'elle est prononcée dans un pays au sein duquel se situe un grand nombre de villes homonymes. Pourtant cette méthode paraît ici très insatisfaisante puisqu'elle échoue à lier correctement la mention "Paris" vers l'entité qui lui correspond, à savoir "Paris, Texas". Les chercheurs ont donc établis une chaîne de traitement plus complexe en générant et sélectionnant les candidats susceptibles de correspondre à l'entité recherchée [4].

La première étape consiste à construire un sous-ensemble de la base de connaissances composé des entités susceptibles de correspondre à la mention. Elle est nécessaire car elle permet d'éviter de travailler avec l'ensemble d'une base de connaissances qui peut compter plusieurs milliers ou millions d'entrées. Mais la sélection doit aussi être effectivement large pour s'assurer que l'entité recherchée est bien dans cette sous-base. Il faut donc établir des critères de sélection basés sur la relation supposée entre la mention et sa correspondance dans la base de connaissances. La méthode d'usage consiste à se baser sur les variantes lexicales des entités : est considéré comme candidat toute entité qui dispose d'une variante lexicale correspondante à la mention recherchée. Cette étape peut également s'accompagner d'un pré-ordonnancement *a priori* des candidats en fonction de critères comme la popularité par exemple. On peut ainsi considérer par défaut que la mention "Paris" a plus de chance d'être un renvoi vers l'entité "Paris, France" que vers "Paris, Texas".

Cet ordonnancement *a priori* ne peut cependant être considéré comme suffisant pour réaliser le liage. Pour être juste, il faut également comparer le contexte d'apparition de la mention avec les métadonnées associées à chaque entité candidate. L'objectif est d'ordonner les entités en fonction de leur proximité avec le contexte de la mention afin de sélectionner celle qui a le plus de chance d'être celle qui lui correspond. Cette proximité peut s'établir en fonction de plusieurs critères comme la co-occurrence de certaines entités par exemple. Il faut également envisager la possibilité que cette mention ne soit pas disponible au sein de la base de connaissances, ou parce que le référentiel est lacunaire

4. *Ibid.*, p. 117-125.

ou parce qu'il s'agit d'une variante lexicale qui n'a pas encore été référencée. Ces cas doivent être clairement identifiés car ils représentent autant de potentiels ajouts à la base de connaissances.

Cette base de connaissances constitue donc ici la clé du processus. Elle se présente comme un ensemble d'entrées associées à des informations dont la structure est systématisée. Similaire à une ontologie, elle peut comme cette dernière se construire de deux façons : elle peut répondre à une logique de mise en place d'un ensemble général de connaissances sur un domaine et se forger dans un contexte industrielle ou participatif. Elle peut au contraire être contextuelle au corpus et se nourrir d'un repérage préalable - manuel ou automatique - des concepts pertinents et des relations qui les caractérisent [5]. Dans les deux cas, cette base de connaissances peut être emmenée à évoluer au cour du travail de liage par l'intégration de nouvelles entités qui ne correspondent à aucune entité de la base de connaissances.

1.1.2 Un enjeu pour les sources historiques

Le développement des techniques de liage d'entité est apparu dans un contexte d'étude de textes contemporains, mais il rencontre aussi le contexte propre aux recherches historiques. L'appropriation des outils numériques par les acteurs de la recherche en histoire et du patrimoine a permis d'accroitre largement la disponibilité des textes et de faciliter l'extraction d'information via des techniques d'OCR ou HTR et d'études statistiques. L'accès au contenu des textes est cependant freinée par des problématiques propres à ces documents. Tout d'abord, le passage par un processus d'OCR peut altérer pour partie le texte. De plus, les conventions orthographiques peuvent varier largement eb fonction des lieux et époques, ce qui rend la reconnaissance de certains mots encore plus délicate.

Le cas de confusion le plus courant se place entre le *f* et le *s* long présent dans de nombreux textes manuscrits et imprimés. D'autres cas de confusion portent sur le mélange des langues (par exemple un nom de lieu en français dans un texte en latin) ou sur des variations orthographiques d'un même mot qui peuvent exister au sein d'un même document. Tous ces éléments rendent d'autant plus complexe la tâche de reconnaissance d'entités nommées et d'extraction pour liage avec une base de connaissances. Pourtant, cette tâche est particulièrement pertinente dans ce contexte ou de nombreuses ambiguïtés existent, notamment pour identifier les personnes et lieux qui sont mentionnés par les documents.

5. *Ibid.*, p. 33

1.2 Un axe de recherche en pleine évolution

1.2.1 Mise en application du liage d'entités à partir d'archives historiques

Plusieurs travaux de recherches ont donc été menés ces dernières années afin de pallier ces difficultés et améliorer les techniques de liage d'entités pour les adapter au contexte des documents historiques. Ces travaux se sont souvent nourris d'autres recherches parallèles portant sur des problématiques proches. C'est le cas par exemple des recherches sur le liage d'entités multi-langue, c'est à dire un modèle dans lequel la langue des données sources n'est pas la même que celle de la base de connaissances. Des chercheurs ont proposés des modèles spécifiques développés à partir de l'incorporation de mots étrangers dans le corpus [6] ou, s'il existe quelques éléments pour produire une base de connaissances dans la langue source, à partir du mélange entre ces derniers et un modèle de liage issu d'une langue disposant d'une base de connaissances plus large [7]. Une dernière méthode consiste à construire un modèle se passant de toute ressource bilingue par l'utilisation d'une langue pivot suffisamment proche pour qu'il soit pertinent de construire un modèle dessus puis de l'utiliser sur la source [8].

Une des problématiques rencontrées par les chercheurs est celle de la base de connaissances utilisée au moment du processus. Un certain nombre de travaux ont ainsi fait le choix de procéder au liage des entités nommées présents dans leur corpus avec des ontologies web généralistes. Celles-ci ont l'avantage d'être très fournies, ce qui est particulièrement utile dans le cadre de données qui n'ont pas de contexte chronologique ou géographique précis. Mais cette situation comporte aussi des inconvénients : ces ontologies sont porteuses de nombreuses ambiguïtés, notamment liées à un grand nombre d'homonymies. Ces caractéristiques ont par exemple été décrites pour Wikipedia au moment de la création d'un algorithme de liage d'entités depuis la base Europeana [9]. D'autres travaux se sont portés sur la comparaison entre les différentes ontologies disponibles pour lier les entités nommées d'un corpus précis [10].

6. Elvys Linhares Pontes, Jose G. Moreno et Antoine Doucet, "Linking Named Entities across Languages using Multilingual Word Embeddings", dans *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, New York, NY, USA, 2020, p. 329-332, URL : `https://doi.org/10.1145/3383583.3398597` (visité le 23/07/2022).

7. Shuyan Zhou, Shruti Rijhwani et Graham Neubig, "Towards Zero-resource Cross-lingual Entity Linking", dans *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Hong Kong, China, 2019, p. 243-252, DOI : `10.18653/v1/D19-6127`.

8. Shruti Rijhwani, Jiateng Xie, Graham Neubig et Jaime Carbonell, "Zero-Shot Neural Transfer for Cross-Lingual Entity Linking", *Proceedings of the AAAI Conference on Artificial Intelligence*, 33–1 (17 juill. 2019), Number : 01, p. 6924-6931, DOI : `10.1609/aaai.v33i01.33016924`.

9. Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando et Mark Stevenson, "Matching Cultural Heritage items to Wikipedia", dans *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012, p. 1729-1735, URL : `http://www.lrec-conf.org/proceedings/lrec2012/pdf/1021_Paper.pdf` (visité le 21/07/2022).

10. Aicha Soudani, Yosra Meherzi, Asma Bouhafs, Francesca Frontini, Carmen Brando, Yoann

Utilisation Kb - Wikipedia : http://www.lrec-conf.org/proceedings/lrec2012/pdf/1021_Paper.pdf En gros il a une giga base de connaissance qu'il a voulu lier avec les articles wikipedia correspondant

- DBpedia : https ://dl.acm.org/action/downloadSupplement ?doi=10.1145

Il compare ici les différents modèles de kb en ligne à utiliser

Les deux derniers à utiliser pour dire un peu où on en est avec ça + sans doute des parties à utiliser pour parler de résultats :

Desambiguisation : https ://aclanthology.org/K18-1050.pdf

Résumé de la vie = https ://hal.archives-ouvertes.fr/hal-03034492 Le problème de la compréhension des sources et de l'absence de base ?

1.2.2 Les perspectives actuelles (ou les applications possibles ?)

Sur le lien avec la TEI : https ://hal.archives-ouvertes.fr/hal-01363709/document

Sur l'état de l'art dans la recherche en France : https ://hal.inria.fr/hal-02617950v2/document et https ://hal.archives-ouvertes.fr/hal-01925816/document Dans l'article résumé super cool j'ai : "Regarding the application of end-to-end EL in Digital Humanities, some works have focused on using available EL approaches to analyse historical data [16,23,28]"

Sur la question spécifique des noms de lieux : https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.8475&rep=rep1&type=pdf http://link.springer.com/10.1007/3-540-44796-2_12

Sur l'utilisation dans les archives : https://hal.archives-ouvertes.fr/hal-03625734/file/Atelier_Culture_Inria_NER4Archives_slides.pdf

https ://hal.archives-ouvertes.fr/hal-02187283/document

Il y a deux articles de Caroline Parfait sur l'influence de l'OCR dans la NER

Article sur l'influence de la qualité d'OCR dans l'entity linking : https ://hal.archives-ouvertes.fr/hal-02557116/document"

Thèse Yoann Dupont parle de la structure des entités nommées et de du modèle en relation des entités (intéressant par rapport à ce qu'on a fait avec Heurist)

1.3 Les perspectives d'amélioration

Aussi : https://www.researchgate.net/profile/Kasra-Hosseini-3/publication/345849651_DeezyMatch_A_Flexible_Deep_Learning_Approach_to_Fuzzy_String_Matching/links/5fafc469299bf10c367c6b2b/DeezyMatch-A-Flexible-Deep-Learning-Approach-to-Fuzzy-pdf?origin=publication_detail

Dupont et Frédérique Mélanie-Becquet, "Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle", dans *Atelier Humanités Numériques Spatialisées (HumaNS'2018)*, Montpellier, France, 2018, URL : https://hal.archives-ouvertes.fr/hal-01925816 (visité le 21/07/2022).

https ://obtic.sorbonne-universite.fr/tanagra/home

https://bnf.hypotheses.org/files/2018/07/Reden-_-pr%C3%A9sentation-atelier-Corpus
pdf

Nom de personnes : https ://hal.archives-ouvertes.fr/hal-01203784 et https ://hal.sorbonne-universite.fr/hal-01396037

Des propositions de visualisation des entités de lieu : https ://hal.archives-ouvertes.fr/hal-01925816/document

1.3.1 Les outils disponibles

Parler de Spacy + autres ? Parler de dbpedia spotlight ! DizzyMatch : https ://github.com/Living-with-machines/DeezyMatch DBPedia Spotlight : https ://hal.archives-ouvertes.fr/hal-01915730

Un framework fait pour l'annotation de textes historiques : https://www.researchgate.net/profile/Rainer-Simon-4/publication/220727253_Augmenting_Europeana_content_with_linked_data_resources/links/00b495174f471206ee000000/Augmenting-Europeana-conten pdf?origin=publication_detail

1.3.2 Des résultats ?

Sur les données un peu complexes : https://helda.helsinki.fi/bitstream/handle/10138/310657/heino_et_al_nel_2017.pdf?sequence=1

Repartir de https ://hal.archives-ouvertes.fr/hal-03034492 + du bilan de la journée à la Sorbonne et dire ce qu'il reste à améliorer

Chapitre 2

Les avancées du projet Himanis

Article Teklia sur les outils de la NER-HTR : `https://teklia.com/publications/DAS2022_NER.pdf`

Chapitre 3

Legacy Metadata : Des instruments de recherche précis mais incomplets

Chapitre 4

Un autre chapitre

4.1 Structuration du mémoire

Le mémoire se structure en plusieurs parties :

1. tout d'abord, les pièces liminaires : page de titre, résumé, remerciements, bibliographie, introduction
2. ensuite, le corps du texte, suivi d'une conclusion
3. après, les annexes (documentation, extraits de code, etc.)
4. enfin, les pièces finales : index (si besoin), glossaire (si besoin) ; tables (des figures et des tables, si nécessaire) ; table des matières

Ce mémoire s'accompagne d'une autre partie très importante : les **données**.

4.2 Les données

Elles constituent une partie primordiale du travail à rendre. Ce sont :

— les données traitées (textes, images, vidéos, BDD, etc.)

— les scripts de traitement

— la documentation associée aux données et au script

— tout autre document qui semble nécessaire au traitement du sujet.

Ces données doivent être ordonnées et accompagnées d'un fichier `lisezMoi` (format `.txt` ou `.md`), présent à la racine du dossier contenant les données. Ce fichier doit décrire l'arborescence des fichiers et dossiers et la fonction de chacun des fichiers.

Le principe important à retenir est celui de la **reproductibilité** du travail.

Deuxième partie

Une autre partie

Conclusion

Annexe A

Première annexe

Table des matières