Extraction d'informations appliquée aux documents non-structurés pour la valorisation de périodiques historiques : application au patrimoine de la région Bourgogne Franche-Comté en France

Journée nationale sur la fouille de textes, 10/06/2024

Nicolas Gutehrlé, Iana Atanassova
C.R.I.T., Université de Franche-Comté

Sommaire

Introduction

Introduction

- Les campagnes de numérisation menées ces dernières années par les archives et les bibliothèques ont permis une meilleure préservation et un meilleur accès à leurs collections
- L'exploitation et la valorisation de ces collections restent des tâches difficiles en raison :
  - du manque du structure de leur contenu textuel
  - erreurs dans les transcriptions obtenues par OCR
  - manque d'outils et de ressources dédiées

Introduction

- Proposition d'une nouvelle approche pour la tâche d'Extraction Jointe d'Entités et de Relations (EJER) (*Joint Extraction of Relations and Entities*, JERE)
- Financé par la région Bourgogne Franche-Comté (2020-2023) dans le cadre du projet EMONTAL (Extraction et Modélisation ONTologique des Acteurs et Lieux pour la valorisation du patrimoine de Bourgogne Franche-Comté)

Données

Construction du corpus EMONTAL

- Périodiques imprimés d'origines diverses publiés aux 19ème et 20ème siècles en Bourgogne et en Franche-Comté
- Collectés à partir de Gallica (archives numériques de la Bibliothèque Nationale de France)
- Métadonnées des documents au format Dublin Core, contenu des documents au format XML ALTO

	Collections	Issues	Pages	textblocks	textlines	strings
Fond *Bourgogne*	113	5,738	637,407	8,117,055	24,183,489	189,266,170
Fond *Franche-Comté*	46	2,648	255,670	3,733,845	11,373,606	83,001,454
Total	159	8,386	893,077	11,850,900	35,557,095	272,267,624

Pré-traitement du corpus

- Approche à base de règles pour corriger les erreurs produites par l'OCR (césure, transcriptions erronées)
- Approche à base de règles pour déterminer la structure logique des documents (*Logical Layout Analysis*, LLA) (Gutehrlé et Atanassova 2022 ; Gutehrlé et Atanassova 2021)
- Conversion des documents dans un nouveau format XML (EMONTAL) décrivant la structure logique des documents

Méthodologie

Approche ELIJERE

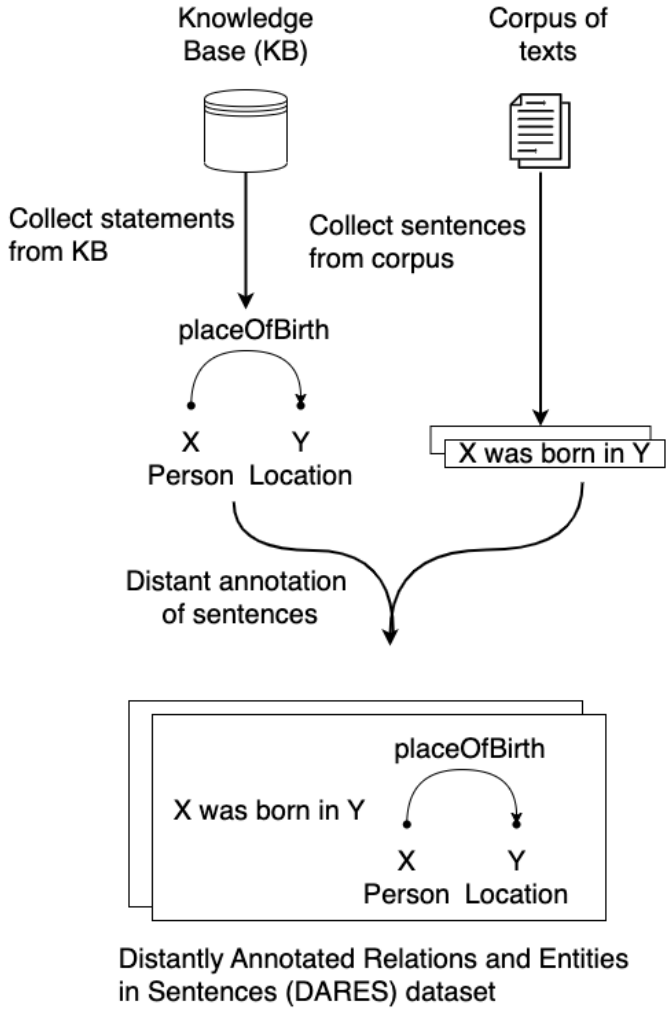
L'approche Extensible, Lightweight and Interpretable Joint Extraction of Relations and Entities repose sur deux ressources linguistiques pour extraire et catégoriser les mentions d'entités impliquées dans des relations depuis les phrases :

- un Index Syntaxique, qui décrit comment une relation est exprimée syntaxiquement, ainsi que le type d'entités impliquées dans la relation
- un Index Lexical, qui décrit comment une relation est exprimée lexicalement

Les ressources linguistiques sont construites à partir de patrons lexico-syntaxiques exprimant une relation entre des entités

Méthodologie

- Annotation faible de phrases via la méthode de supervision distante (*distant supervision*) (Mintz et al. 2009)
- Permet de rendre notre approche extensible



Distribution des relations dans le jeu de données DARES

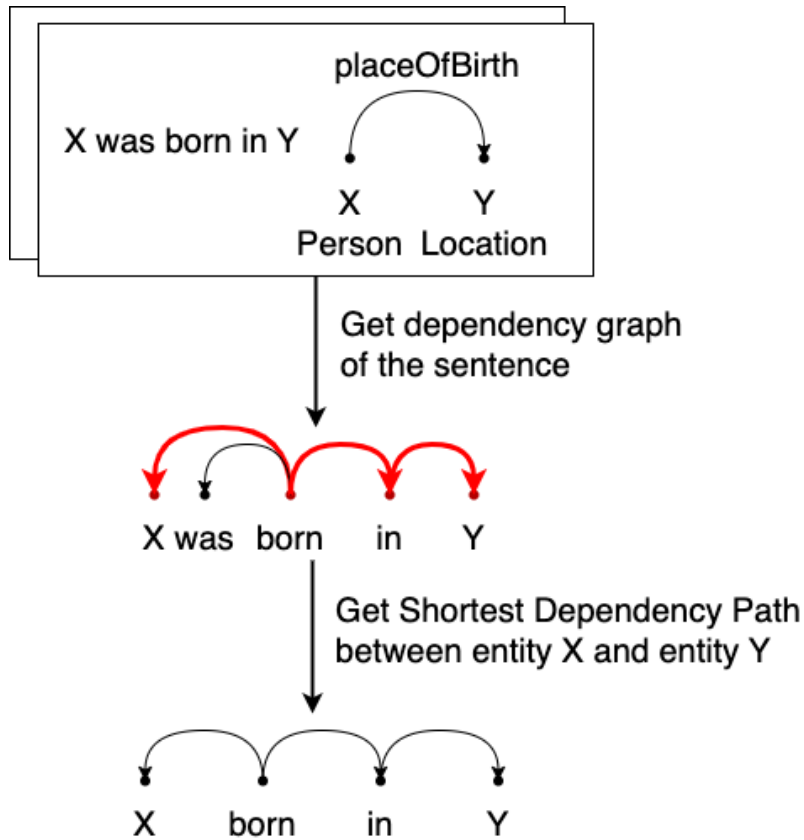
	Train set		Test set	
Label	Count	Proportion	Count	Proportion
capitalOf	303	4.227 %	140	11.914 %
country	303	4.227 %	183	15.574 %
dateOfBirth	303	4.227 %	71	6.042 %
dateOfDeath	303	4.227 %	39	3.319 %
educatedAt	303	4.227 %	13	1.106 %
headOfGoverment	303	4.227 %	15	1.276 %
inception	303	4.227 %	12	1.021 %
memberOf	303	4.227 %	53	4.510 %
nextInBodyWater	303	4.227 %	47	4.000 %
occupation	303	4.227 %	125	10.638 %
Other	0	0 %	213	18.127 %
placeOfBirth	303	4.227 %	67	5.702 %
sharesBordersWith	303	4.227 %	174	14.808 %
spouse	303	4.227 %	23	1.957 %
Total	7,167	100 %	1,175	100 %

Distribution des entités dans le jeu de données DARES

	Train set		Test set	
Label	Count	Proportion	Count	Proportion
Person	2,424	30.769 %	391	19.628 %
Location	4,242	53.846 %	1,337	67.118 %
Time	909	11.538 %	128	6.425 %
Misc	303	3.846 %	136	6.827 %
Total	7,878	100 %	1,992	100 %

Collecte des patrons lexico-syntaxiques

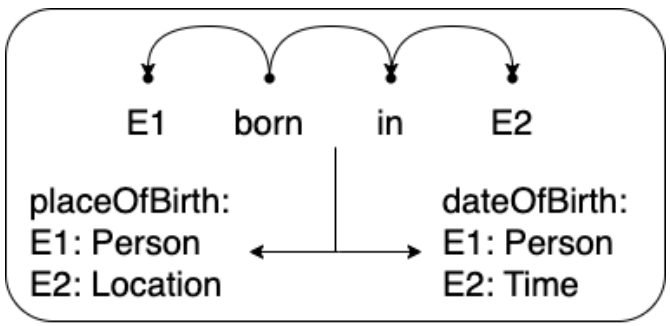
- Extraction du plus court chemin de dépendance (Shortest Dependency Path (SDP)) (Bunescu et Mooney 2005) entre les entités pour collecter les patrons lexico-syntaxiques



Construction de l'Index Syntaxique

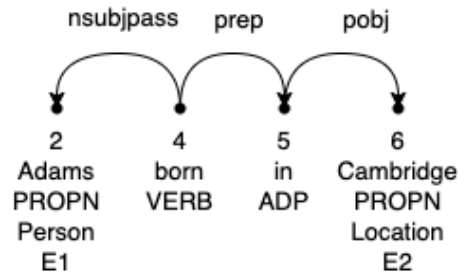
Identification des classes de patrons :

- structure syntaxique identique
- parties du discours identiques
- prédicat identique (noeud qui gouverne tous les autres noeuds)



Syntactic Index

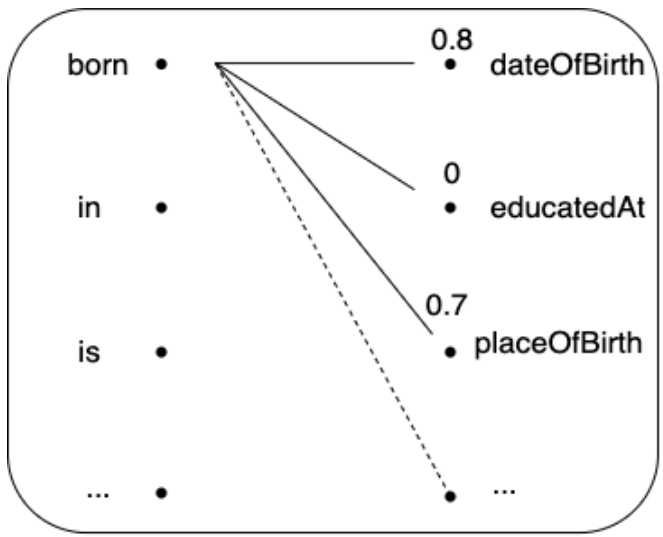
Exemple de l'Index Syntaxique




```
{  "born_VERB" :  [
        {  "graph" :
            "size" : 3,
            "relations" : ["placeOfBirth"],
            "ambiguous" : False,
            "support" : {"placeOfBirth": 100},
            "entities" : {
                "placeOfBirth" : {
                    "E1" : "Person",
                    "E2" : "Location"
                }
             "source_node" : 2,
             "target_node" : 6,
        }
```

Construction de l'Index Lexical

- Mesure d'un score d'association TF-IDF entre les vocabulaires des patrons lexico-syntaxiques et les relations qu'ils expriment
- inspiré par l'index inversé pondéré de la méthode *Explicit Semantic Analysis* (ESA) (Gabrilovich et Markovitch 2007)



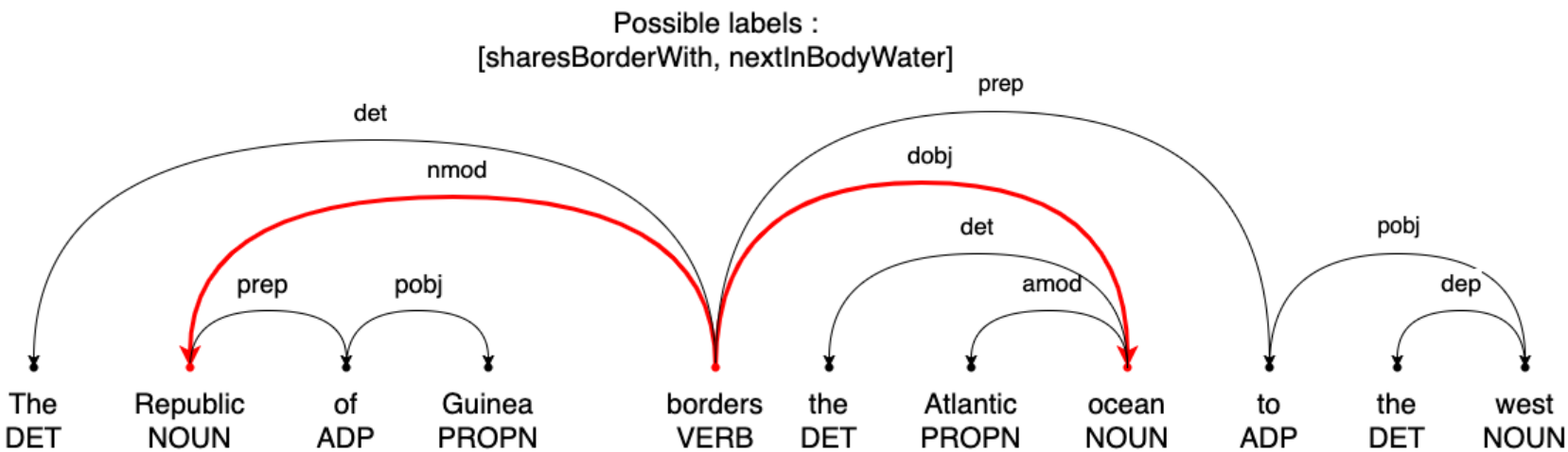
Lexical Index

Exemple de l'Index Lexical

Words	occupation	nextInBodyWater	dateOfBirth	educatedAt
river_NOUN	0.000	1.000	0.000	0.000
university_NOUN	0.000	0.000	0.000	0.996
death_NOUN	0.219	0.000	0.000	0.000
...	...	...	...	...
autor_NOUN	0.998	0.000	0.033	0.000
born_VERB	0.084	0.000	0.770	0.008

Extraction Jointe d'Entités et de Relations

- Application des patrons stockés dans l'Index Syntaxique pour extraire des candidats dans le graphe de dépendance de la phrase
- Chaque candidat est associé avec un ensemble d'étiquettes de relations possibles



Extraction Jointe d'Entités et de Relations

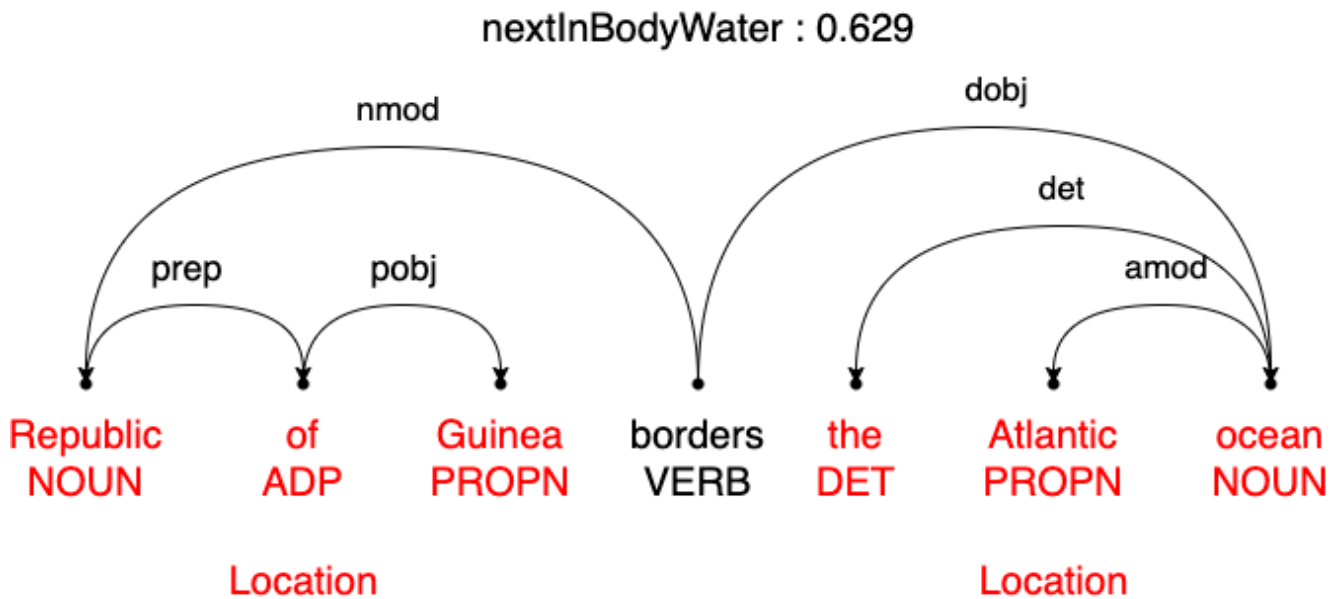
- Sélection de la relation ayant le score d'association le plus élevé, calculé comme la moyenne harmonique des scores d'association du vocabulaire du candidat
- Seuil sémantique pour garantir un score suffisamment élevé

Possible labels:
[nextInBodyWater, sharesBorderWith]

	nextInBodyWater	sharesBorderWith
Republic_NOUN	0.032	0.923
borders_VERB	0.856	0.887
ocean_NOUN	1.000	0.000
Harmonic mean	0.629	0.603

Extraction Jointe d'Entités et de Relations

- Le type des entités impliquées dans la relation est déterminé par le patron ayant matché et la relation prédite
- Les frontières des entités sont déterminées par des règles conçues manuellement exploitant les parties du discours et rôles syntaxiques des noeuds environnants



Evaluation

Evaluation

Nous évaluons deux implémentations de l'approche ELIJERE :

base ELIJERE model : s'appuie sur l'Index Syntaxique pour extraire les candidats et sur l'Index Lexical pour classer ces candidats en fonction de leur vocabulaire

hybrid ELIJERE model : s'appuie sur l'Index Syntaxique pour extraire les candidats et sur un modèle XGBoost pour catégoriser les candidats, sur la base de leur vocabulaire sous forme de word embeddings (ConceptNet, (Speer, Chin et Havasi 2018))

Evaluation

- Evaluation sur le jeu de données DARES, puis sur le corpus EMONTAL
- Dix évaluations pour chaque modèle avec différentes valeurs de seuil sémantique entre 0 et 0,9
- Evaluation sur les tâches d'Extraction de Relation (ER) et de Reconnaissance d'Entité Nommées (REN) séparément
- Evaluation sur la tâche de REN selon la méthode SemEval (Segura-Bedmar, Martınez Fernández et Herrero Zazo 2013)

Moyenne des scores par seuil sur la tâche d'ER

	DARES			EMONTAL		
	P	R	F1	P	R	F1
base ELIJERE model	0.518	0.120	0.189	0.352	0.154	0.199
hybrid ELIJERE model	**0.675**	**0.137**	**0.218**	**0.469**	**0.166**	**0.218**

Moyenne des types d'erreur par seuil sur la tâche d'ER

	Model	Predicate not found	Pattern not found	Wrong label	Semantic Score too weak
DARES	*base ELIJERE model*	**56.572** %	**36.232** %	**2.000** %	5.193 %
	hybrid ELIJERE model	57.630 %	36.910 %	2.965 %	**2.491** %
EMONTAL	*base ELIJERE model*	**52.909** %	**34.722** %	**2.746** %	9.622 %
	hybrid ELIJERE model	54.572 %	35.813 %	3.458 %	**6.154** %

Moyenne des scores par seuil sur la tâche de REN

		DARES			EMONTAL		
Model	Setting	P	R	F1	P	R	F1
base	type	**0.982**	0.111	0.199	**0.856**	0.133	0.225
ELIJERE	partial	**0.829**	0.095	0.169	**0.800**	0.132	0.220
model	strict	**0.649**	0.074	0.133	**0.488**	0.076	0.129
	exact	**0.659**	0.076	0.135	**0.600**	0.099	0.167
hybrid	type	**0.980**	0.129	0.226	**0.827**	0.160	0.265
ELIJERE	partial	**0.846**	0.111	0.195	**0.818**	0.160	0.264
model	strict	**0.681**	0.090	0.157	**0.506**	0.097	0.161
	exact	**0.693**	0.091	0.160	**0.637**	0.124	0.205

Discussion

- Manque de diversité des patrons lexico-syntaxiques
  - corpus hétérogène, Grand Modèles de Langues, patrons de surface
- Annotations incorrectes des phrases par supervision distante
  - clustering
- Difficulté à identifier les frontières des entités nommées
  - apprentissage des frontières avec les patrons
- Impact des erreurs de transcription OCR
  - améliorer post-traitements, recherche approximative (*fuzzy matching*)
- Difficulté à s'adapter aux différents styles d'écriture du corpus cible
  - *bootstrapping* (ex : Snowball (Agichtein et Gravano 2000))

Conclusion

Conclusion

Proposition d'une nouvelle approche pour la tâche d'Extraction Jointe d'Entités et de Relation

- Extensible, puisque repose sur la supervision distante pour collecter et annoter les données
- Interprétable, puisqu'elle repose sur des ressources linguistiques explicites
- Légère, puisqu'elle nécessite peu de ressources informatiques et peut fonctionner sur des processeurs

Application de notre méthode à un corpus de périodiques français du 19ème et 20ème siècle

Conclusion

Les évaluations montrent que notre approche est valide, mais doit être améliorées sur plusieurs aspects :

- Diversité des patrons
- Erreurs d'annotation dues à la supervision distante
- Adaptation au corpus cible et au bruit
- Identification des frontières des entités

Travaux futurs

- Mise en place des améliorations suggérées
- Application de notre approche à des corpus d'autres natures (ex : scientifiques, journalistiques)
- Evaluation de l'approche ELIJERE sur des jeux de données standards (ex : New York Times, TACRED) afin de la comparer aux autres approches, notamment en *deep-learning*

Bibliographie

Agichtein, Eugene et Luis Gravano (2000). “Snowball : Extracting Relations from Large Plain-Text Collections”. In : *Proceedings of the Fifth ACM Conference on Digital Libraries*. DL ’00. San Antonio, Texas, USA : Association for Computing Machinery, p. 85-94. isbn : 158113231X. doi : 10.1145/336597.336644. url : https://doi.org/10.1145/336597.336644.

Bunescu, Razvan et Raymond Mooney (oct. 2005). “A Shortest Path Dependency Kernel for Relation Extraction”. In : *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada : Association for Computational Linguistics, p. 724-731. url : https://aclanthology.org/H05-1091.

Gabrilovich, Evgeniy et Shaul Markovitch (2007). “Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis”. In : *International Joint Conference on Artificial Intelligence*. url : https://api.semanticscholar.org/CorpusID:5291693.

Gutehrlé, Nicolas et Iana Atanassova (2021). “Dataset for Logical-layout analysis on French historical newspapers”. In.

— (2022). “Processing the structure of documents : Logical Layout Analysis of historical newspapers in French”. In : *Journal of Data Mining & Digital Humanities*.

Mintz, Mike et al. (août 2009). “Distant supervision for relation extraction without labeled data”. In : *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore : Association for Computational Linguistics, p. 1003-1011. url : https://aclanthology.org/P09-1113.

Segura-Bedmar, Isabel, Paloma Martınez Fernández et Marıa Herrero Zazo (2013). “Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)”. In : Association for Computational Linguistics.

Speer, Robyn, Joshua Chin et Catherine Havasi (2018). *ConceptNet 5.5 : An Open Multilingual Graph of General Knowledge*. arXiv : 1612.03975 [cs.CL].

Merci pour votre attention !

Extraction d'informations appliquée aux documents non-structurés pour la valorisation de périodiques historiques :
application au patrimoine de la région Bourgogne Franche-Comté en France

Nicolas Gutehrlé
nicolas.gutehrle@univ-fcomte.fr