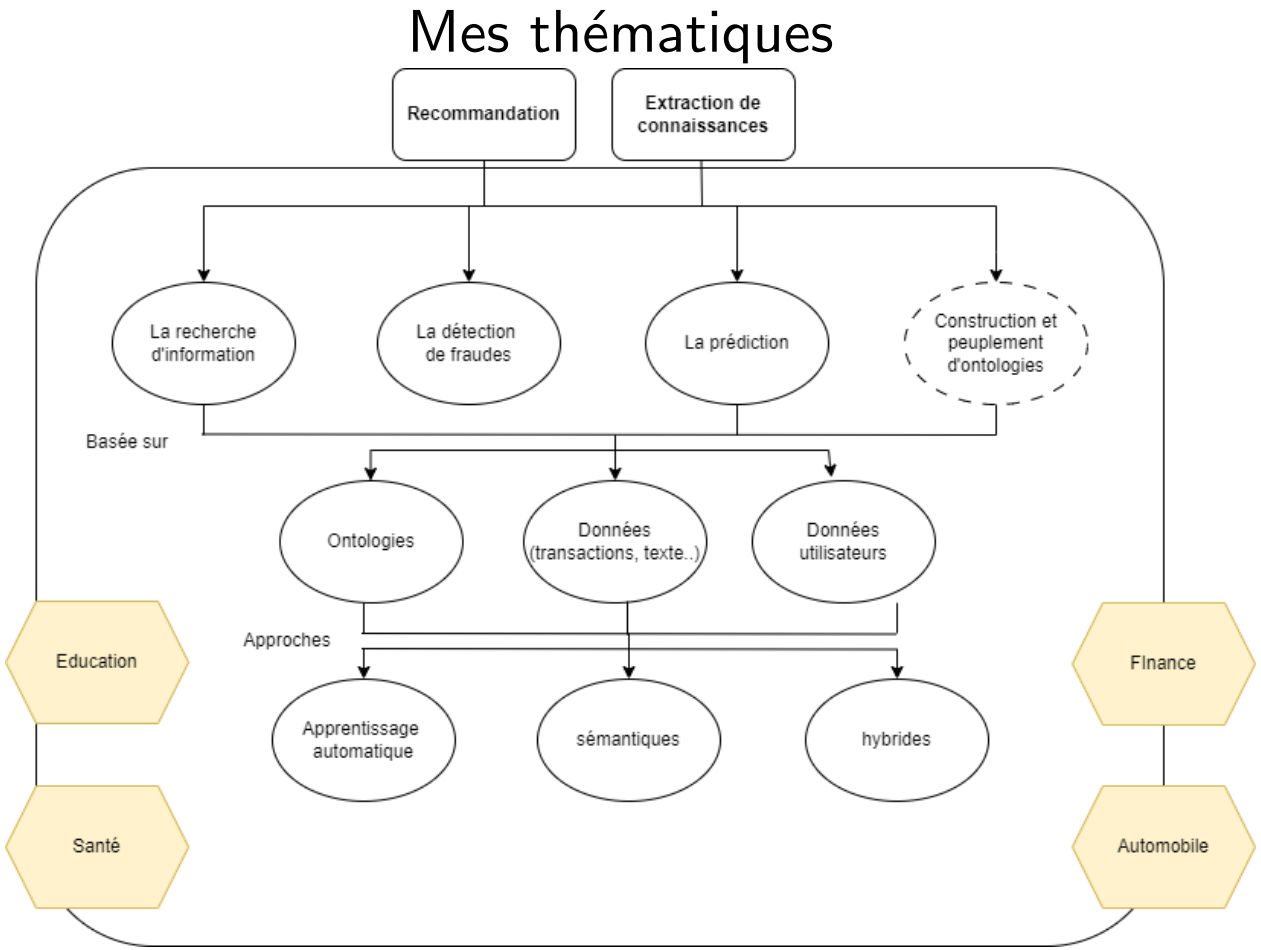
Extraction d'informations à partir de rapports automobiles pour le peuplement d'ontologies

Lylia Abrouk[1]

[1]LIB Laboratory, University of Burgundy
lylia.abrouk@u-bourgogne.fr

13 novembre 2023

Thématiques de recherche



Thématiques de recherche

Proposition de méthodes et d'approches pour l'analyse et l'exploitation des données dans différents domaines.

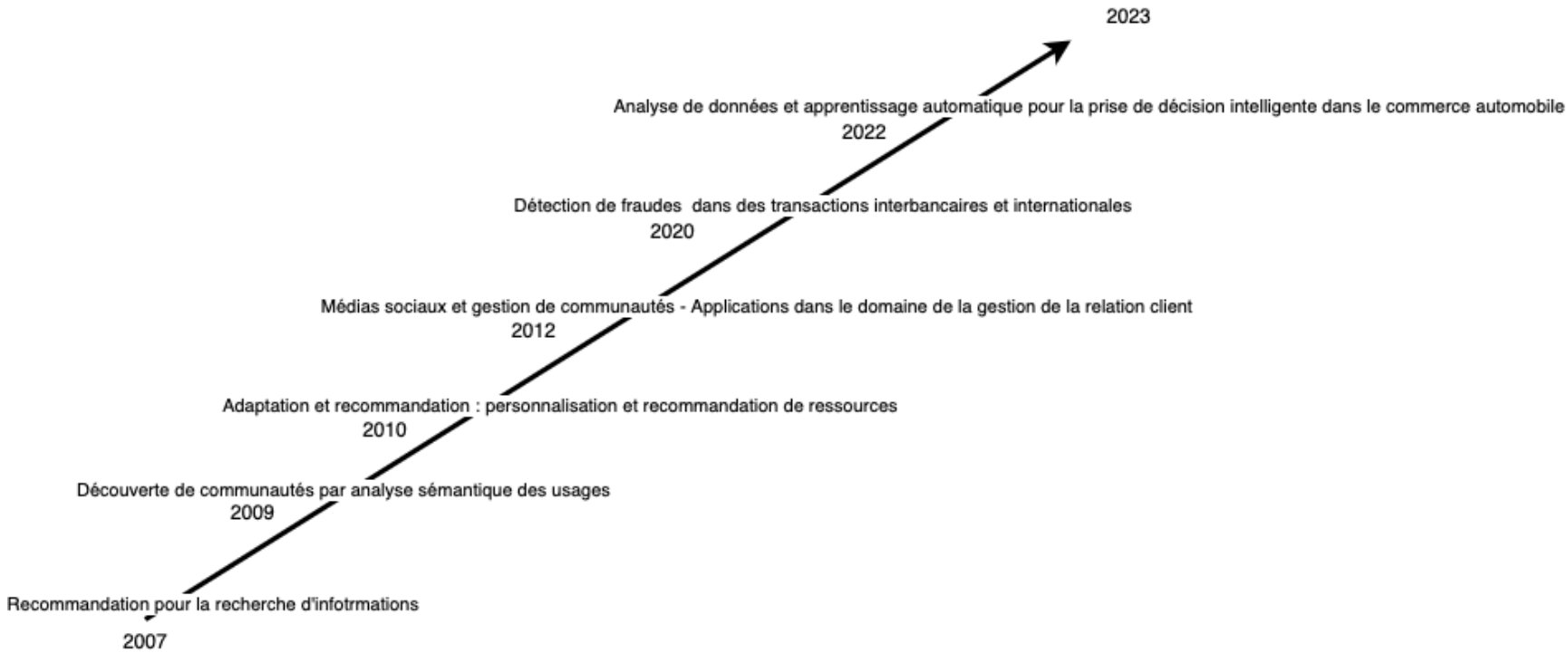


Figure: Travaux de recherche

Plan

Contexte

- La gestion des transports de voiture dans la vente d'automobile est complexe : nécessité de détecter les dommages qui varient en taille et en nature.
- Divers processus mis en place : photos, rapports...
- Processus manuel coûteux par les experts.
- Approches d'extraction d'informations et modélisation.

Contexte

- Objectif : Modéliser les dommages sur les voitures à l'aide d'une ontologie.
- Utilisation de l'ontologie par les compagnies d'assurance avec des données réelles.
- Idée : Évaluer les dommages et déterminer le coût de réparation basé sur les images ou la description textuelle.
- Utilisation de techniques de traitement du langage naturel et de reconnaissance d'entités nommées pour évaluer les dommages à partir du texte.
- Plan :
  - Une ontologie pour la modélisation des dommages (OCD).
  - Méthodologie : La construction de l'ontologie et l'extraction d'informations pour alimenter l'ontologie.

Les ontologies

- Utilisation répandue des ontologies pour organiser la connaissance.
- Importance dans la modélisation des dommages dans divers secteurs.
- Domaine automobile : utilisées pour modéliser les accidents de la circulation en décrivant les circonstances, l'emplacement, les causes et les effets de l'accident.
- Pas de travaux sur modélisation des dommages subis par le véhicule.

Les ontologies existantes

- Everett et al. (2002) - Ontologie pour dommages de copieurs.
- Rachman et Ratnayake (2018) - Ontologies pour équipements de traitement.
- Hamdan et al. (2019) - Ontologie pour dommages de bâtiment.
- Barrachina et al. (2012), Dardailler (2012) - Modélisation des accidents de trafic.
- Klotz et al. (2018), Feld et Müller (2011) - Ontologies pour informations de véhicules.
- Hepp (2010) - Ontologie pour la vente de véhicules.

Notre Ontologie

- Développement d'une ontologie spécifique à l'évaluation des dommages automobiles.
- Capture des types et niveaux de gravité des dommages.
- Modélisation de la voiture et de ses composants.

Comparaison des Ontologies Automobiles

Table: Comparaison des Ontologies Automobiles.

CritèresTravaux	Barrachina et al (2012)	Feld et Muller (2011)	Hepp (2010)	Klotz et al. (2018)	Notre ontologie (OCD)
Modélisation des Dommages de Voiture	×	×	×	×	✓
Modélisation des Informations de Voiture	✓	✓	✓	✓	✓
Modélisation des Pièces	×	✓	×	✓	✓
Support Multilingue	×	×	×	×	✓
Accès Public	×	×	✓	✓	✓
Capacités d'Inférence	×	✓	×	×	✓
Objectif	Modélisation des Accidents Routiers	Partage de Connaissances sur les Véhicules	Modélisation des Véhicules pour E-commerce	Modélisation des Signaux de Véhicules	Modélisation des Dommages de Voiture

Extraction d'Informations

- Processus automatique d'extraction de données pertinentes de sources diverses.
- Utilisation de techniques de NLP et de ML.
- Informations allant de faits simples à des éléments complexes.
- Domaine de la santé, Domaine de l'industrie de la construction

Extraction d'information

- Domaine automobile :
  - Rubens et Agarwal (2002) : une combinaison d'algorithmes de classification TALN et d'apprentissage automatique pour extraire des attributs à partir d'annonces automobiles en ligne. Pour automatiser la recherche.
  - Bhatia et al. (2008) : basé sur la REN et des règles manuelles.
  - Jalal (2020) : basé sur les expressions régulières sur des annonces automobiles en ligne.

Reconnaissance d'entités nommées

Plusieurs types d'approches :

- Basées sur des dictionnaires.
- Basées sur des règles.
- Basées sur un corpus annoté.
- Basées sur l'apprentissage actif (Chen et al. (2015) et Tran et al. (2017)).
- Basées sur les réseaux de neurones (Huang et al. (2015) et Lopez et Kalita (2017)) : apprend à partir de schémas complexes.
- Approches hybrides : combinent plusieurs méthodes pour améliorer la performance (Thomasand Sangeetha (2019)).

Extraction de Relations

- Sous-domaine de l'extraction d'informations identifiant les relations sémantiques entre entités textuelles.
- Représentation des phrases par séquences d'embeddings de mots: $S = \{w_1, w_2, \ldots, w_n\}$.
- Représentation des entités par embeddings d'entités: $E = \{e_1, e_2, \ldots, e_m\}$.

Objectif de l'extraction de relations

- Identifier les relations R existantes entre les entités dans E.
- Définition des relations: $R = \{r_1, r_2, \ldots, r_k\}$.
- Chaque relation r_i est une fonction des embeddings d'entités indiquant le type de relation.
- Fonction d'extraction de relations F :

$$F(S, E) = \{(e_i, e_j, r_k) \mid e_i, e_j \in E, r_k \in R\}$$

Méthodes d'Extraction de Relations

- Méthodes basées sur des règles.
- Approches supervisées.
- Apprentissage semi-supervisé.
- Méthodes non supervisées.

Extraction de relations

- Approches basées sur des règles : utilisent des règles ou des motifs pour définir les relations entre les entités. Peut être complexe.
- Approches supervisées : prédire la relation entre deux entités. Nécessitent de grands ensembles de données annotées pour l'entraînement.
- Approches semi-supervisées : elles combinent des données étiquetées et non étiquetées.
- Approches non supervisées : reconnaître les paires d'entités qui apparaissent fréquemment dans la même phrase ou le même document.

Synthèse

- Pas de travaux sur la modélisation des dommages des véhicules
- Approches basées sur les règles : limités selon le contexte et le texte.
- Les approches d'apprentissage profond produisent de bons résultats
- Approche : REN et ER basées sur l'apprentissage profond.

Approche

- Deux étapes principales :
  - La construction d'une ontologie : les concepts pertinents et les relations dans le domaine de l'évaluation des dommages
  - L'extraction d'informations : extraction des informations du texte et mise en correspondance avec l'ontologie pour la peupler
- Construction de l'ontologie OCD
- Améliorer le processus d'évaluation des dommages

Objectifs de l'Ontologie

- Créer une représentation complète et sémantiquement riche du domaine des dommages automobiles.
- Capturer les relations entre différentes pièces de voiture, types de dommages, niveaux de gravité et autres entités pertinentes.
- Améliorer l'interopérabilité des données et automatiser l'extraction d'informations à partir de rapports de dommages non structurés.
- Établir l'interopérabilité sémantique en définissant des concepts et relations partagés.

Questions de Compétence

- **Q1:** Quels sont les différents types de dommages pouvant survenir sur une voiture ?
- **Q2:** Quelles sont les différentes parties de la voiture pouvant être affectées par des dommages ?
- **Q3:** Quel est le degré de gravité de chaque type de dommage ?
- **Q4:** Quelles sont les relations possibles entre les pièces de voiture et les dommages ?

Questions de Compétence (suite)

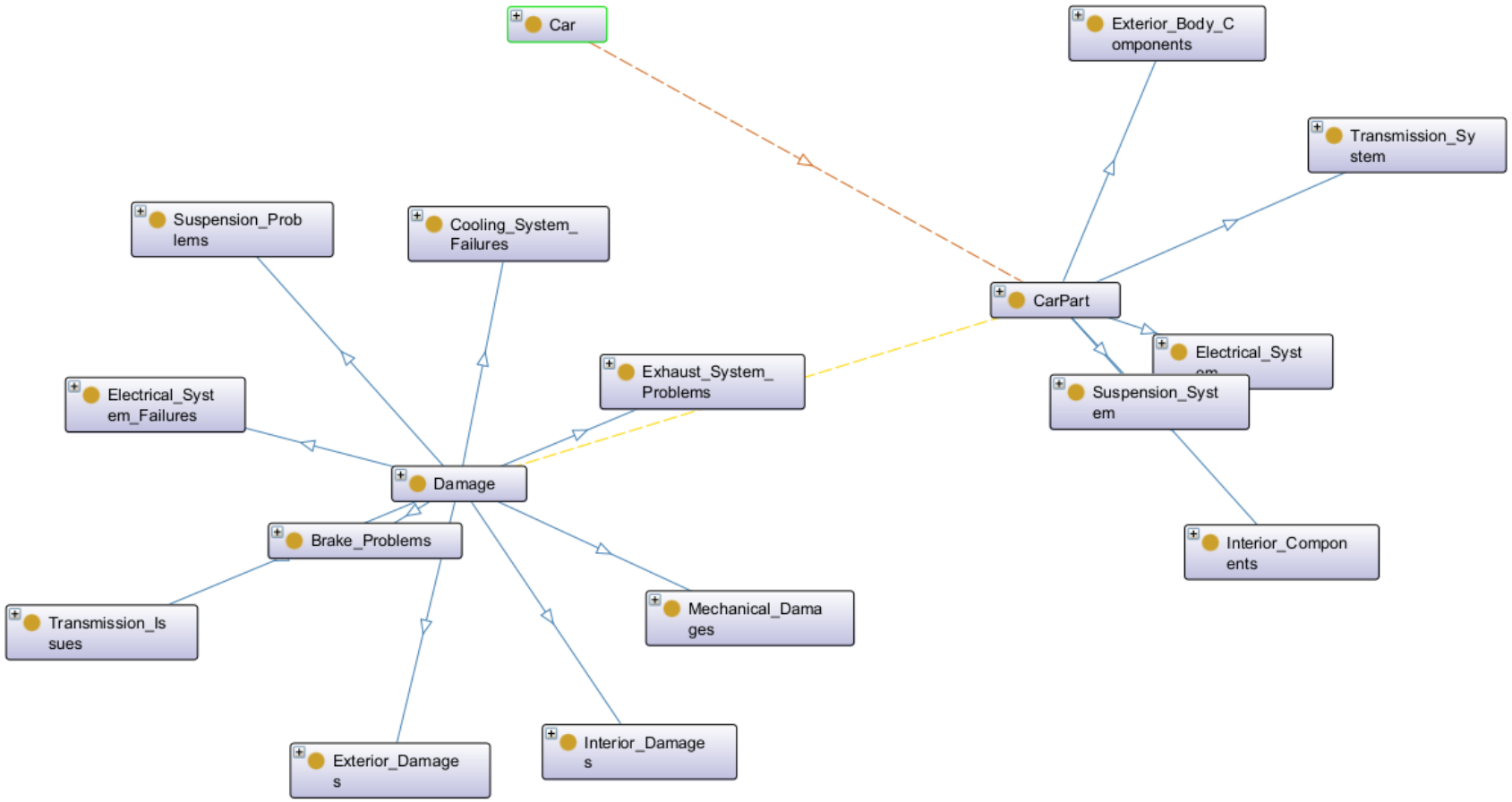
- **Q5:** Comment les informations extraites des rapports non structurés peuvent-elles être liées à des concepts spécifiques dans l'ontologie ?
- **Q6:** Comment l'ontologie peut-elle être utilisée pour améliorer la précision et la pertinence des relations extraites ?
- **Q7:** Quelle est la structure globale et la hiérarchie des composants de voiture et des types de dommages au sein de l'ontologie ?

Structure et Hiérarchie de l'ontologie

- Définition de la hiérarchie claire pour les composants de voiture et les types de dommages.
- Exemple: Organisation hiérarchique des pièces (p. ex. Carrosserie, Coffre, Couvercle du coffre).
- Catégorisation des types de dommages (p. ex. Dommage, Dommage de carrosserie, Éraflures).

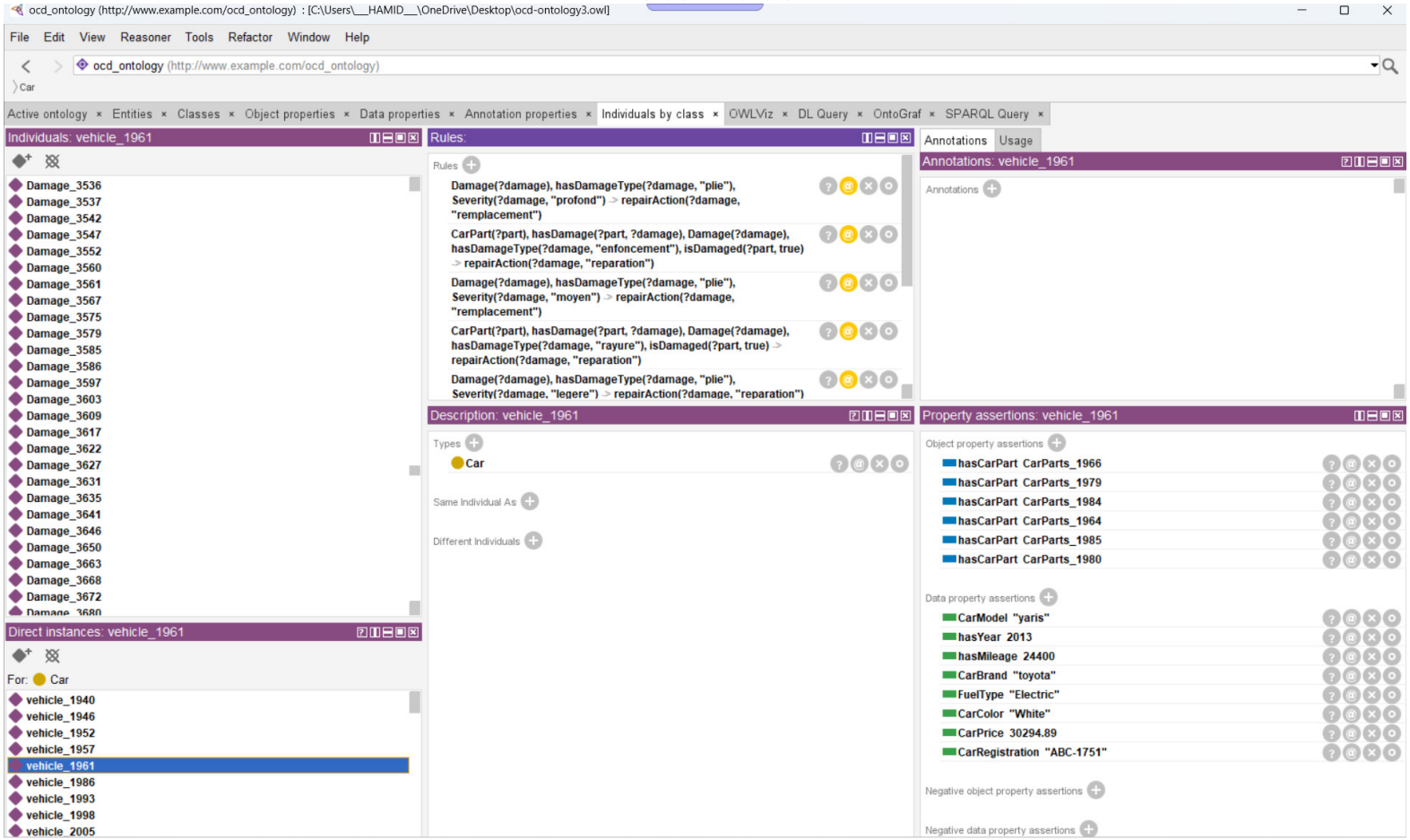
Construction de l'ontologie

Concepts



Construction de l'ontologie

Instances



Portal

Ontology for Car Damage

Last uploaded: October 30, 2023

Summary | Classes | Properties | Notes | Mappings | Instances | Souslesens

Details

Acronym	OCD
Visibility	Public
View Of Ontology	DEMO
Description	Ontology for Car Damage (OCD) is a domain-specific ontology designed for modeling car damage information. It provides a formal representation of entities and relationships involved in damage assessment, enabling automated data analysis for decision support in assessment, cost estimation, and repair planning. The ontology is structured using OWL (Web Ontology Language) standards to ensure accurate and consistent representation of car-related data.
Status	Alpha
Format	OWL
Contact	Hamid AHAGGACH, hahaggach@syartec.com
Categories	Other
Groups	OBO Foundry

Additional Metadata

URI	https://developer.iziflo.com/research/ocd

Links

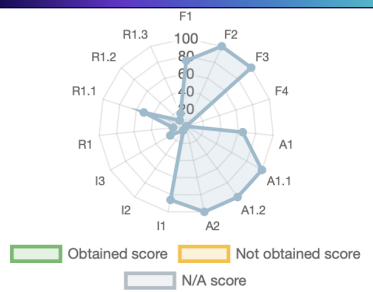
Go to the REST API JSON entry

Get my metadata back

N-Triple | Json-LD | RDF/XML

FAIR Scores beta

Total score : 258.0 (53.0%) details

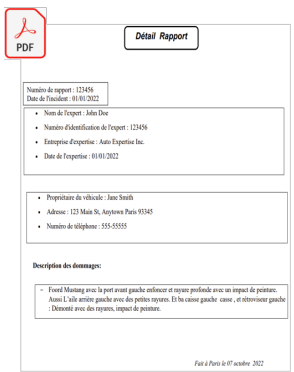


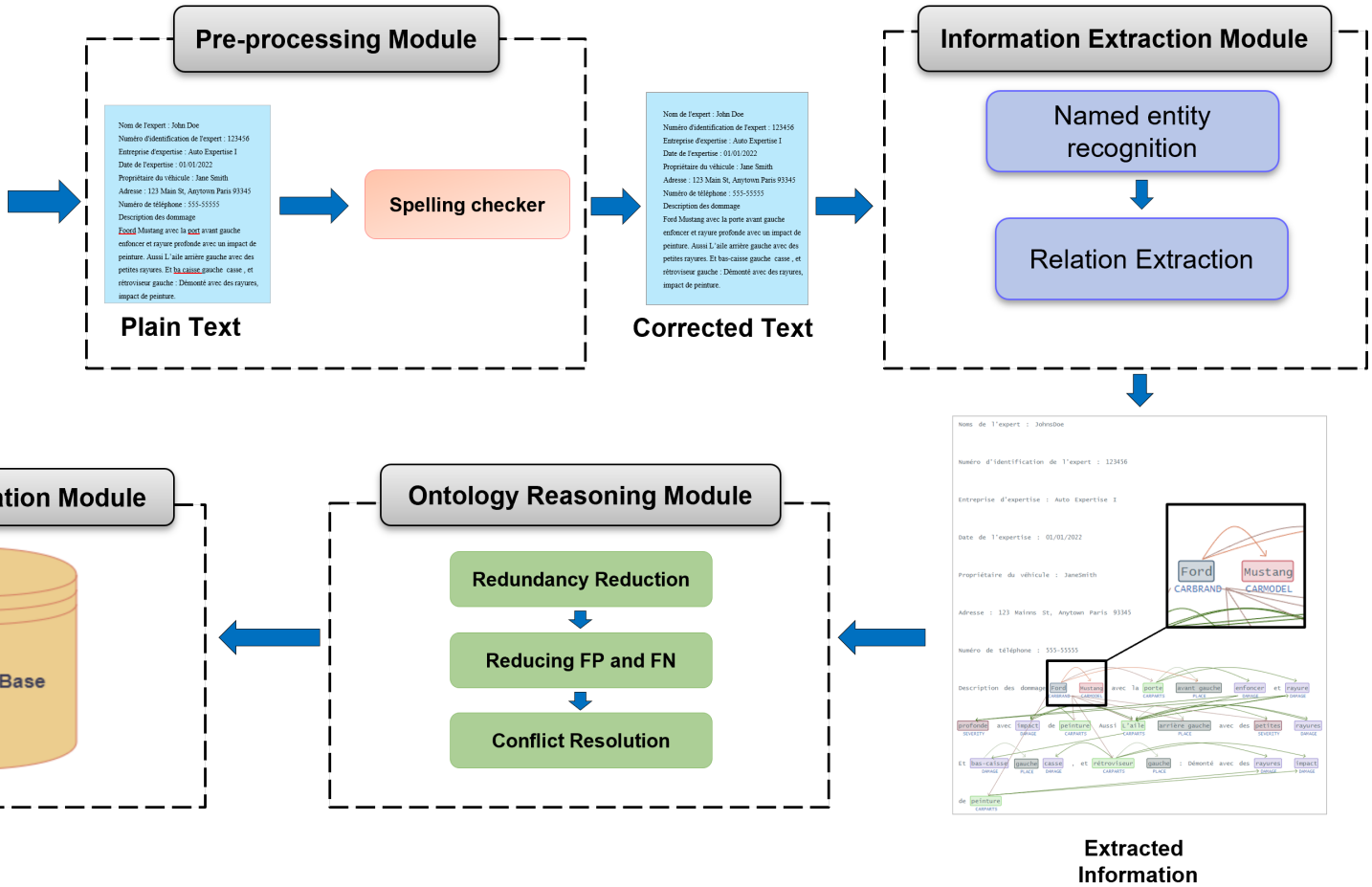
Approche : Évaluation de l'ontologie

- Validation et cohérence de l'ontologie
- Experts du domaine pour les concepts et relations
- Créée avec protégé 5.5.0
- Fact++ et HermiT : la consistance et la cohérence.

Approche : Peuplement de l'ontologie

méthodologie générale





Pre-processing module

- Texte extrait de rapports PDF et analysé avec un algorithme de vérification orthographique.
- Correction des erreurs d'orthographe pour augmenter l'efficacité et la précision de l'extraction d'informations.
- Plusieurs méthodes basées sur : es dictionnaires, les règles, les statistiques ou les réseaux de neurones.
- Approche : Basée sur les dictionnaires combiné à un algorithme basé sur distance de Levenshtein.

Information extraction module

- Extraction des informations sur les voitures :
  - Marque, modèle, couleur, etc.
  - Les composants endommagés.
  - Le type de dommage.
  - Les caractéristiques du dommage (sévérité, emplacement, etc.)
- Extraction des relations.
- Étiquetage des rapports.

Extraction d'entités nommées

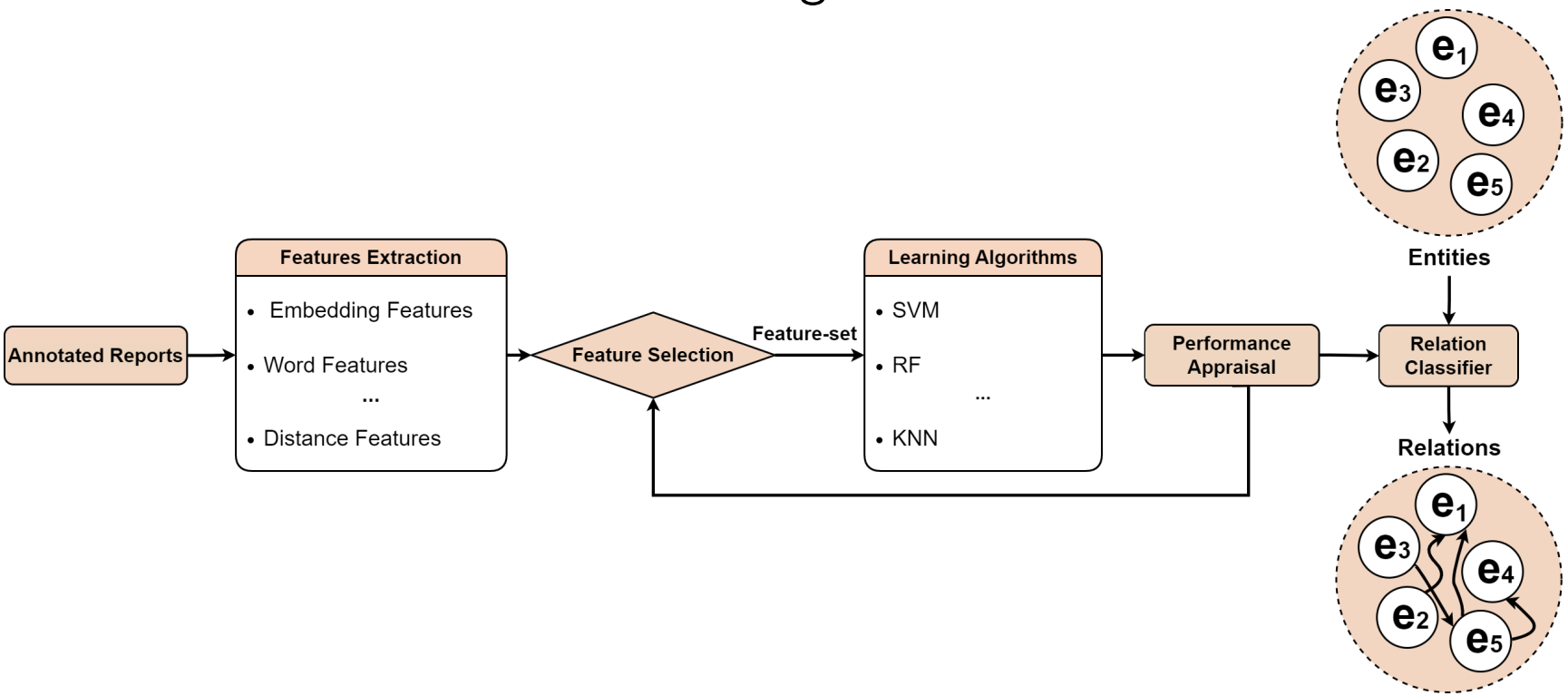
- Capturer le contexte des mots et des parties du discours de chaque entité dans le rapport.
- Comparaison de performances de plusieurs modèles :
  - Conditional Random Fields (CRF)
  - Long Short-Term Memory Bidirectional-CRF (BiLSTM-CRF)
  - FlauBERT
  - NER de SpaCy.

Extraction de relations

- Extraction de relations entre les entités extraites
- Extraction de features :
  - Caractéristiques de distance : Distance entre les mots, Distance entre les caractères, Distance entre les phrases, Orientation.
  - Caractéristiques de mot : fréquences des types d'annotation, type d'entité
  - Emebedding features : modeles d'embedding : Word2Vec andSpaCy.

Extraction de relations

Processus général



Expérimentations

- Rapports automobiles décrivant les dommages: pré-traités et étiquetés
- Entités

models	Entities																	
	Damage			CarParts			CarBrand			CarModel			Severity			Place		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CRF	0.89	0.94	0.91	0.89	0.89	0.89	**1.00**	0.97	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.95	0.91	0.93
FlauBERT	0.69	0.55	0.61	0.69	0.77	0.73	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.33	0.67	0.44	0.73	0.78	0.76
CRF	0.98	0.91	0.94	**0.97**	**0.95**	**0.96**	0.97	**1.00**	0.99	0.97	0.95	0.96	0.88	**1.00**	0.93	**1.00**	**1.00**	**1.00**
SpaCy Model	**1.00**	**0.95**	**0.97**	0.96	0.91	0.93	0.94	0.96	0.95	0.89	0.98	0.93	**1.00**	**1.00**	**1.00**	0.88	**1.00**	0.93

Expérimentation

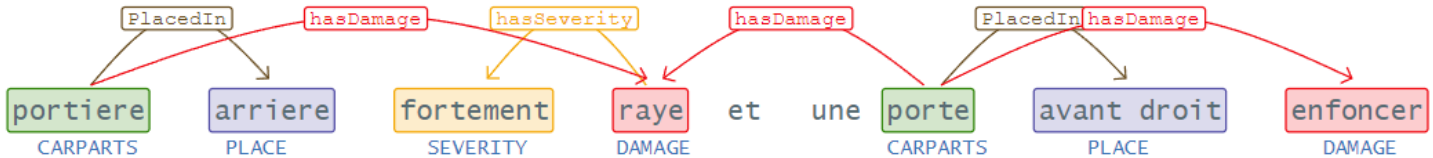
Models	Relation type											
	hasDamage			hasCarParts			PlacedIn			hasSeverity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM	0.95	0.55	0.70	**0.98**	0.54	0.70	**1.00**	0.82	0.90	0.50	0.40	0.44
KNN	0.75	0.75	0.75	0.75	0.75	0.75	0.98	0.95	0.96	0.60	0.60	0.60
DT	0.94	**0.92**	**0.93**	0.97	**0.97**	**0.97**	0.98	**0.98**	**0.98**	0.60	0.60	0.60
RF	**0.96**	0.90	**0.93**	0.97	**0.97**	**0.97**	0.98	**0.98**	**0.98**	**0.71**	**1.00**	**0.83**

Experimentation

Features	Relation type											
	hasDamage			hasCarParts			PlacedIn			hasSeverity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
baseline: distance & word feat-s	0.96	0.90	0.93	**0.97**	0.97	0.97	**0.98**	**0.98**	**0.98**	0.71	**1.00**	0.83
baseline - nb word dist	**0.99**	0.89	**0.94**	0.96	0.96	0.96	0.87	0.79	0.83	0.50	0.60	0.55
baseline-char dist	0.95	0.88	0.92	**0.97**	0.97	0.97	**0.98**	**0.98**	**0.98**	0.71	**1.00**	0.83
baseline sent dist	0.95	0.76	0.84	**0.97**	**0.99**	**0.98**	0.74	0.76	0.75	0.50	0.60	0.55
baseline + Embs: Spacy's vector	0.94	0.75	0.83	**0.97**	0.92	0.94	0.93	0.88	0.91	0.43	0.60	0.50
baseline + Embs: word2vec	0.98	0.76	0.85	0.96	0.94	0.95	0.95	0.92	0.93	0.43	0.60	0.50
baseline + Ontology-reasoning	0.97	**0.90**	0.93	**0.97**	**0.99**	**0.98**	**0.98**	**0.98**	**0.98**	**0.82**	**1.00**	**0.90**

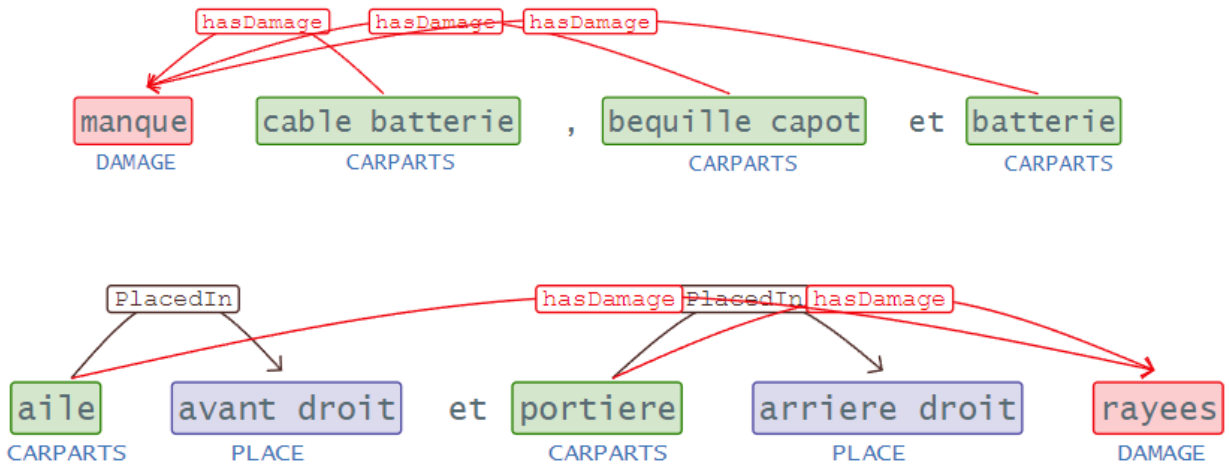
Expérimentations

- Intégration du raisonnement de l'ontologie : identification de relations correctes.



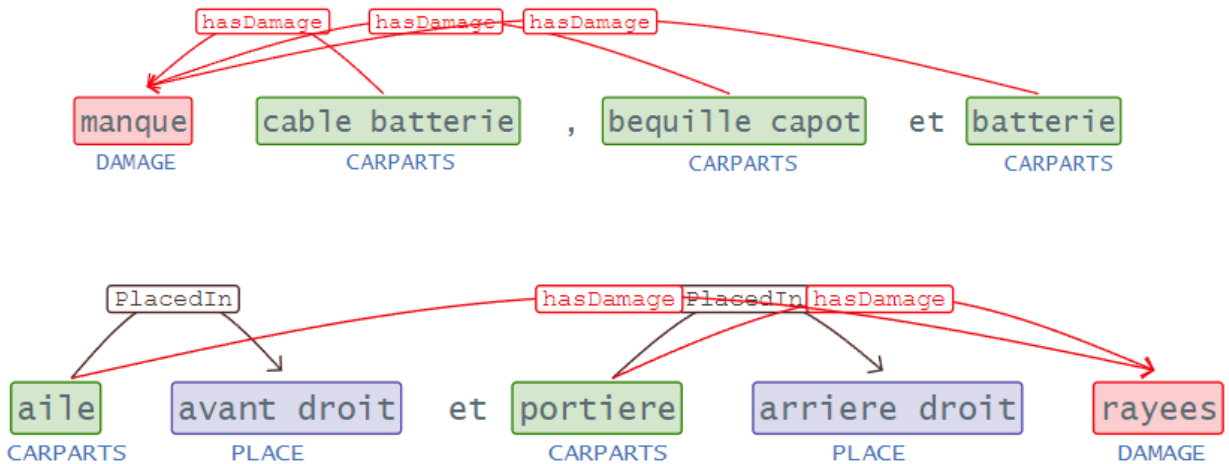
Expérimentations

- Extraction d'informations pertinentes
- Rapports automobiles complexes



Expérimentations

- Extraction d'informations pertinentes
- Rapports automobiles complexes



Conclusion

- Le développement de l'ontologie OCD.
- Une approche structurée et standardisée pour la modélisation des dommages dans l'industrie automobile.
- Approches d'extraction d'entités nommées et extraction de relations.
- Prédire les coûts de réparation en fonction de la gravité et de la nature des dommages.