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

Hamid Ahaggach[1,3], Lylia Abrouk,[1,2], Eric Lebon [3]

[1] Laboratoire LIB , Université de Bourgogne, Dijon, France
prenom.nom@u-bourgogne.fr
[2] MISTEA, Université de Montpellier, INRAE & Institut Agro, France
[3] Syartec, Aix-en-Provence, France

Résumé

*Dans cet article, nous présentons le travail publié dans la revue **Applied Ontology** intitulé **Information extraction from automotive reports for ontology population*** [1] *(Février 2024). Ce travail met en lumière nos recherches dédiées à l'utilisation des ontologies et à l'extraction d'informations (EI), dans le but d'analyser et de modéliser les dommages subis par les carrosseries de voitures. Notre approche consiste à analyser les rapports non structurés en appliquant des techniques de traitement automatique du langage, telles que la reconnaissance d'entités nommées (REN) et l'extraction de relations (ER), afin d'identifier et d'extraire les informations pertinentes des rapports.*

Mots-clés

Extraction d'informations, Ontologie, Reconnaissance d'entités nommées, Extraction de relations.

Abstract

*In this article, we present the paper published in the journal **Applied Ontology** titled **Information extraction from automotive reports for ontology population** [1] (February 2024). It highlights our research work dedicated to the use of ontologies and information extraction, aimed at analyzing and modeling the damages incurred by car bodies. Our approach analyzes unstructured reports using natural language processing techniques, such as REN and RE, to identify and extract relevant information from the reports.*

Keywords

Information extraction, Ontology, Named entity recognition, Relationship extraction.

1 Introduction

Dans le secteur automobile, la gestion du transport des voitures est une tâche complexe. À l'arrivée de chaque voiture, un contrôle de qualité est effectué pour identifier les dommages subis pendant le transport, incluant la prise de photographies et la rédaction de rapports. Cependant, ces rapports de dommages sont non structurés et ne suivent pas de standards uniformisés pour la description des dommages. Ceci entraîne un processus de saisie manuelle des données chronophage et susceptible d'erreurs. Pour cela, nous avons développé *OCD* (Ontology for Car Damage), une ontologie conçue pour la modélisation des dommages des voitures. Cette ontologie offre un cadre structuré permettant de décrire et de catégoriser les différents types de dommages de manière précise et uniforme. De plus, l'ontologie contribue à l'amélioration du système d'extraction d'informations que nous avons proposé. Ce système est conçu pour extraire les informations pertinentes des rapports automobiles non structurés afin de peupler notre ontologie. Notre approche facilite le processus de contrôle de qualité pour le transport des véhicules et offre une méthode standardisée pour documenter et catégoriser les dommages des voitures.

1. The accepted paper.

2 Travaux antérieurs

2.1 Ontologie

Dans le domaine de l'automobile, les ontologies ont été utilisées pour modéliser les accidents de la route en décrivant les circonstances, le lieu, les causes et les effets de l'accident. Cependant, l'accent mis sur la modélisation des dommages subis par le véhicule a été insuffisant. Le tableau 1 offre une comparaison complète des différentes ontologies automobiles selon plusieurs critères.

TABLE 1 – Comparaison des ontologies automobiles.

Critères/Travaux	[1]	[2]	[3]	[4]	[5]
Modélisation des dommages de voiture	✓	×	×	×	×
Modélisation des informations de voiture	✓	✓	✓	✓	✓
Modélisation des pièces	✓	×	✓	×	✓
Support multilingue	✓	×	×	×	×
Accès public	✓	×	×	✓	✓
Capacités d'inférence	✓	×	✓	×	×

2.2 Extraction d'information

Ces dernières années, plusieurs chercheurs se sont intéressés à l'extraction d'informations dans le domaine de l'automobile, en raison de son potentiel à améliorer divers pro-

cessus dans l'industrie. De nombreuses études se sont focalisées sur l'extraction d'informations telles que les marques et modèles de véhicules. Ces stratégies mises en œuvre demeurent conventionnelles, se basant principalement sur des systèmes de règles. Ces approches sont limitées, car elles ne parviennent pas à s'adapter aux variations du langage naturel et du contexte, ce qui affecte la précision et la fiabilité des données extraites. De plus, elles ne permettent pas d'extraire les relations entre les entités, élément essentiel pour une compréhension approfondie de la sémantique des informations.

3 Approche

Notre approche se compose de deux étapes principales : la construction de l'ontologie et l'extraction d'informations. Dans l'étape de construction de l'ontologie, nous définissons les concepts et les relations pour représenter le domaine de l'évaluation des dommages des voitures, L'ontologie est disponible sur *GitHub*[2] et *industryportal*[3]. Dans l'étape d'extraction d'informations (Figure 1), nous extrayons des informations (Reconnaissance d'Entités Nommées REN et Extraction de Relations ER) à partir du texte. Une fois les relations extraites, nous améliorons le résultat de cette extraction en utilisant le raisonnement de l'ontologie. Ceci permet de réduire les redondances en gérant les conflits et en minimisant les faux positifs et les faux négatifs dans les relations extraites. Nous peuplons l'ontologie en associant les entités et relations extraites aux concepts et propriétés de notre ontologie.

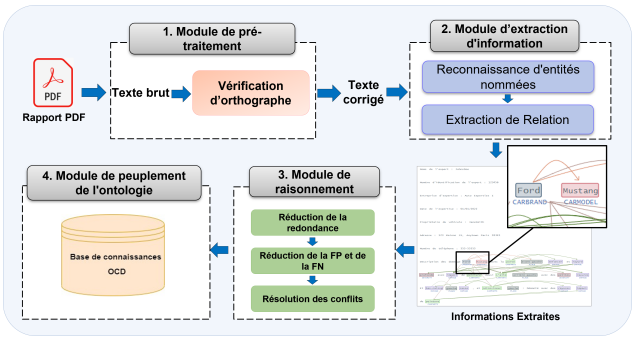

FIGURE 1 – Méthodologie générale

4 Résultats et Conclusion

Pour évaluer l'efficacité de notre approche, nous avons utilisé un ensemble de données comprenant des rapports automobiles décrivant les dommages aux voitures, fournis par l'entreprise *Syartec*[4]. L'ensemble de données a été prétraité et étiqueté à l'aide de l'outil *doccano* pour l'entraînement et le test de nos modèles REN et ER. Notre objectif est d'extraire des entités à partir de rapports de dommages automobiles. Plus précisément, nous nous sommes concentrés sur l'extraction de six types d'entités : $CarBrand$, $CarModel$, $Carparts$, $Damage$, $Severity$ et $Place$. Nous avons exclu les autres informations des rapports, car elles étaient déjà structurées et il n'était pas nécessaire de les extraire. Les résultats ont montré que le modèle $SpaCy$ est performant pour la plupart des entités, tandis que le modèle CRF est performant avec les entités spécifiques au domaine, et le modèle $BiLSTM - CRF$ est performant pour les entités composées. Une fois les entités extraites, nous avons utilisé des algorithmes d'apprentissage automatique pour identifier les relations entre elles. Nous avons cherché à extraire quatre types de relations : $hasDamage$, $hasCarParts$, $CarBrand$ et $Carparts$. Pour créer nos modèles d'extraction de relations, nous utilisons quatre algorithmes de classification utilisés pour l'extraction de relations : les machines à vecteurs de support, les k-plus proches voisins, les arbres de décision et les forêts aléatoires. En combinant le raisonnement de l'ontologie OCD, nous avons réussi à extraire des informations pertinentes de rapports automobiles complexes, particulièrement dans les scénarios où un seul événement de dommage est associé à plusieurs parties de la voiture. Nous avons créé l'ontologie OCD en utilisant *Protégé 5.5.0*[5] et l'avons peuplée avec les informations extraites en utilisant la bibliothèque $Owlready$.

Ces travaux ouvrent la voie à de nouvelles questions liées à la prédiction des coûts de réparation des véhicules. Nous avons initié une approche visant à répondre à cette problématique en proposant une approche hybride. En intégrant des règles $SWRL$ dans notre ontologie, nous pouvons identifier les composants réutilisables, réduisant ainsi la nécessité d'acheter de nouvelles pièces et, par conséquent, minimisant les coûts de réparation.

2. github.com/OntologyCarDamage/OCD
3. industryportal.enit.fr/ontologies/OCD
4. www.syartec.com
5. https://protege.stanford.edu

Références

[1] Ahaggach, H., Abrouk, L., & Lebon, E. (2024). Information extraction from automotive reports for ontology population. Applied Ontology, 1-30.

[2] Barrachina, J., Garrido, P., Fogue, M., Martinez, F. J., Cano, J. C., Calafate, C. T., & Manzoni, P. (2012, April). Caova : A car accident ontology for vanets. In 2012 IEEE wireless communications and networking conference (WCNC) (pp. 1864-1869). Ieee.

[3] Feld, M., & Müller, C. (2011, November). The automotive ontology : managing knowledge inside the vehicle and sharing it between cars. In Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 79-86).

[4] Hepp, M. (2010). Vehicle sales ontology. Retrieved from http ://www.heppnetz.de/ontologies/vso/ns

[5] Klotz, B., Troncy, R., Wilms, D., & Bonnet, C. (2018, October). VSSo : The Vehicle Signal and Attribute Ontology. In SSN@ ISWC (pp. 56-63).