CarDRP - An advanced tool for price prediction from unstructured damage reports

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

[1] *LIB Laboratory, University of Burgundy, Dijon, France*
[2] *Syartec, Aix-en-Provence, France*

Abstract

In the automotive industry, accurately estimating the cost of repairing car damages is crucial for both customers and service providers. The process of manually analyzing unstructured reports describing car damage and predicting repair prices is time-consuming and prone to errors. To address this challenge, this article introduces *CarDRP* (Car Damage Repair Price), an advanced tool designed to automate the prediction of repair prices from unstructured damage descriptions.

CarDRP leverages the power of machine learning and natural language processing techniques to analyze and extract relevant information from textual reports. A regression model is then employed to predict the approximate cost of repairing the reported damages. This automated approach significantly simplifies the pricing process, saving time and effort for both car owners and repair service providers.

Keywords

Cost estimation, machine learning, natural language processing, entity extraction, relationship extraction.

1. Introduction

The automotive industry faces numerous challenges [1, 2], and one of them revolves around the efficient and accurate estimation of car repair damages. During vehicle transportation, damages may occur. To ensure quality control, every vehicle undergoes inspection, and any identified damages are documented in a car damage report. However, the estimation process is complicated by the unstructured nature of these reports. The manual estimation process for repair costs has been time-consuming and error-prone, relying heavily on manual analysis and interpretation. In response to this problem, we created a software tool called *CarDRP*. This advanced tool aims to automate the estimation of repair costs. To the best of our knowledge, there are no other applications currently available in this domain, making *CarDRP* a groundbreaking and unique solution in this field. The primary objective of *CarDRP* is to automatically analyze and extract relevant information from the unstructured car damage reports. By using named entities recognition and relation extraction techniques, the tool is capable of extracting entities and relationships from textual descriptions of car damages. This structured data is then utilized to train a regression model that predicts the approximate cost of repairing the reported damages.

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6-10, 2023, Athens, Greece

*Corresponding author.

†These authors contributed equally.

Hamid.ahaggach@u-bourgogne.fr (H. Ahaggach); lylia.abrouk@u-bourgogne.fr (L. Abrouk); elebon@syartec.com (E. Lebon)

CEUR Workshop Proceedings (CEUR-WS.org)

In the following sections, we will delve deeper into the functionality and capabilities of *CarDRP*. We will explore how the tool leverages machine learning and natural language processing techniques to analyze and extract crucial information from the unstructured insurance reports. Furthermore, we will present the results of estimation repair price.

2. CarDRP Tool

In this section, we will provide a detailed overview of the two main phases of our *CarDRP* tool.

2.1. Information Structuring

The first phase of *CarDRP* involves the crucial task of structuring the information extracted from the unstructured reports. In this phase, the tool utilizes natural language processing techniques to analyze and understand the textual descriptions of car damages. It extracts entities such as the type of damage (e.g., dented bumper, cracked windshield), its severity, and relevant attributes (e.g., location, size). Furthermore, the tool establishes relationships between these entities to provide a comprehensive understanding of the reported damages.

2.1.1. Named Entity Recognition

Named Entity Recognition (NER) is a task in natural language processing that involves identifying and recognize named entities in text. There are several approaches to NER, including: rule-based approach, dictionaries approach, machine learning and, deep learning [3]. The aim of NER is to label specific entities, such as names of people, organizations, and other predefined categories. In the context of *CarDRP*, NER is utilized to identify and extract named entities related to car damages, car parts, damage location etc. In the task of NER, we compare different machine learning algorithms such as CRF [4], BILSTM [5], FlauBERT [6], etc. and we select the best model, in our case, we have chosen SpaCy NER model. SpaCy[1] is a powerful natural language processing library that offers advanced features for entity recognition and text processing. SpaCy utilizes a combination of rule-based matching, statistical models, and deep learning techniques to identify and classify entities in text. It uses pre-trained models and custom training data to achieve accurate and efficient entity extraction. We fine-tuned *SpaCy* NER model with 1000 iterations, a *batch_size* of 32, a *dropout* rate of 0.35, and the *sgd* optimizer.

To evaluate the performance of the models, we measure various metrics such as F1 score, precision, and recall. Table 1 shows how different models performed when they were tested on various entity types such as *Damage*, *CarParts*, *CarBrand*, *CarModel*, *Severity*, and *Place*.

2.1.2. Relation Extraction

Relation Extraction (RE) is a task in natural language processing that focuses on identifying and extracting relationships between entities mentioned in text. In the context of the information structuring phase of *CarDRP*, once named entities related to car damages have been recognized,

[1] https://spacy.io/models/fr

Table 1
Comparative results of NER models. The highest precision (P), recall (R) and F1-scores (F1) are in bold.

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	0.88	**1.00**	0.93
SpaCy Model	**1.00**	**0.95**	**0.97**	0.96	0.91	0.93	**1.00**	**1.00**	**1.00**	0.89	0.98	0.93	0.94	0.96	0.95	**1.00**	**1.00**	**1.00**

the tool aims to establish relationships between these entities to gain a comprehensive understanding of the reported damages. For instance, it aims to extract the relationship *hasDamage* between the car part and the damage. The RE task can be viewed as a classification task. The goal is to classify and categorize the relationships between entities into different predefined classes or categories.

In this task, we evaluate the performance of multiple machine learning models for recognizing and classifying relationships between entities. The models under consideration include Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forests (RF). The objective is to determine the most effective model for relation extraction. Our evaluation reveals that the RF model outperforms the others, providing better results in this context. This achievement was realized through the optimization of particular hyperparameters, specifically by setting the *num_estimators* to 8 and the *maximum_depth* to 10. Table 2 displays a comparative analysis of various models for relation extraction covering relationships such as *hasDamage*, *hasCarParts*, *PlacedIn*, and *hasSeverity*.

Table 2
Comparative results of models for relation extraction. The highest precision (P), recall (R) and F1-scores (F1) are in bold.

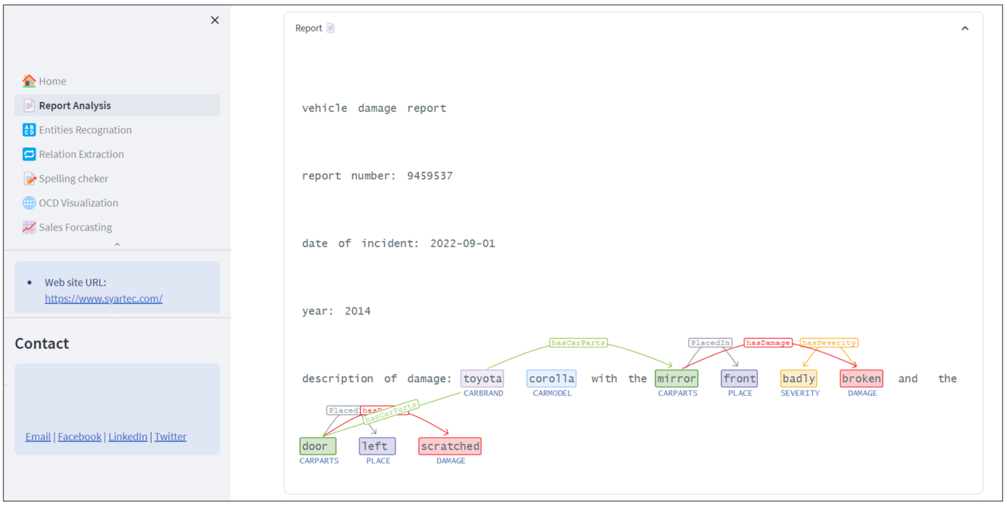
Models	Relation type											
	hasDamage			hasCarParts			PlacedIn			hasSeverity		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM	0.95	0.55	0.70	**0.98**	0.54	0.70	**1.00**	0.82	0.90	0.50	0.40	0.44
KNN	0.75	0.75	0.75	0.75	0.75	0.75	0.98	0.95	0.96	0.60	0.60	0.60
DT	0.94	**0.92**	**0.93**	0.97	**0.97**	**0.97**	0.98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

2.1.3. Discussion

In structuring the information phase, *CarDRP* transforms the unstructured textual data into a structured format (Figure 1) that can be readily utilized for further analysis. This phase enables the accurate predictions and ensuring the reliability of the subsequent price estimation.

2.2. Price Prediction

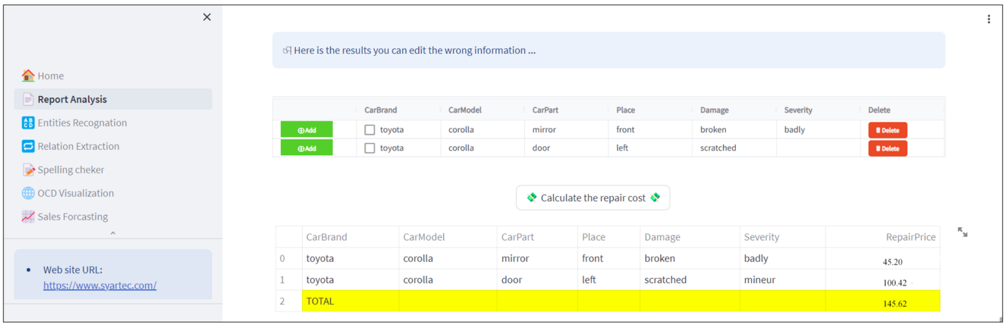
In the price prediction phase, machine learning algorithms and statistical techniques are employed to estimate the repair cost. These models utilize the structured data obtained from the



- Web site URL: https://www.syartec.com/

```
vehicle damage report

report number: 9459537

date of incident: 2022-09-01

year: 2014

description of damage: toyota corolla with the mirror front badly broken and the
door left scratched
```

Figure 1: Extraction of Entities and Relations.



- Web site URL: https://www.syartec.com/

Here is the results you can edit the wrong information ...

	CarBrand	CarModel	CarPart	Place	Damage	Severity	Delete
⊕Add	□ toyota	corolla	mirror	front	broken	badly	Delete
⊕Add	□ toyota	corolla	door	left	scratched		Delete

	CarBrand	CarModel	CarPart	Place	Damage	Severity	RepairPrice
0	toyota	corolla	mirror	front	broken	badly	45.20
1	toyota	corolla	door	left	scratched	mineur	100.42
2	TOTAL						145.62

Figure 2: Estimating Repair Costs with Regression Models.

Information structuring to predict the price of repair cost. The price prediction models are trained using historical data, which encompasses structured data with actual price values. By analyzing the relationships between the input features and the prices in the training data, the models learn to make predictions on new, unseen data. In our study, we conducted a comparison of several regression models to predict repair costs. The performance of these models was evaluated using different evaluation metrics, including mean absolute error, mean squared error, root mean squared error, and R-squared score. Among the evaluated models, *XGBRegressor* model demonstrated the best performance, achieving an mean absolute error of 0.92 with the following hyperparameters: `n_estimators=100`, `learning_rate=0.1`, `max_depth=5`. This indicates that, on average, the predicted repair costs deviate from the actual costs by only 0.92 euros, Figure 2 present an example of estimating repair costs with *XGBRegressor* model.

In Figure 1 and Figure 2, we present the interface of our *CarDRP* system, illustrating the flow from structured information extraction to price prediction. These figures provide a visual representation of the different stages involved in our approach.

2.3. Experimental Setup:

We conducted all our experiments using a real dataset of damage reports, and we ran them on a machine equipped with 16 GB RAM and an Intel Core $i7 - 12700H$ processor, ensuring ample computational resources for both training and testing our models. The *CarDRP* tool was developed using Python programming language and the *Streamlit*[2] framework. The demo of `CarDRP` is available online[3].

3. Conclusion and Perspectives

In this study, we propose a tool for predicting repair costs using unstructured data, which operates in two phases. The first phase, *information structuring* transform unstructured textual data into a structured format to facilitates accurate predictions and enhances the reliability of subsequent price estimations. Furthermore, the *price prediction* phase employs machine learning specially regression models to forecast the repair costs. Looking forward, there are several perspectives to consider for future research. Firstly, it would be beneficial to explore the incorporation of additional data to train deep learning models and further improve prediction accuracy. Additionally, investigating the integration of semantic ontologies can provide a deeper understanding of the underlying concepts and relationships within the data. This integration can enhance the overall predictive capabilities of the system and enable more accurate estimations.

References

[1] P. M. Kyu, K. Woraratpanya, Car damage detection and classification, in: Proceedings of the 11th international conference on advances in information technology, 2020, pp. 1–6.

[2] H. Ahaggach, L. Abrouk, S. Foufou, E. Lebon, Predicting car sale time with data analytics and machine learning, in: IFIP International Conference on Product Lifecycle Management, Springer, 2022, pp. 399–409.

[3] H. Ahaggach, L. Abrouk, E. Lebon, Information extraction and ontology population using car insurance reports, in: International Conference on Information Technology-New Generations, Springer, 2023, pp. 405–411.

[4] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[5] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

[6] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, arXiv preprint arXiv:1912.05372 (2019).

[2] https://streamlit.io/

[3] https://drive.google.com/drive/folders/1o0NOBqxj3rxFURFBmuH-C6kjj_d85TOU