Towards a Framework for Risk-Neutral Autonomous Vehicle by Semantic Similarity Analysis

Ingyu Lee†*,Byung-Won On‡
†Yeungnam Univ. Institute of Information and Communication,
‡Kunsan National Univ. Department of Software Science & Engineering

Abstract : Many companies have worked on autonomous vehicle and some have already driven on the street to test for almost ten years. However, incidents involving autonomous vehicle are frequently reported on media, and people are still worrying about the safety of autonomous vehicle. In this paper, we are proposing a framework to understand the risk factors of autonomous vehicles by analyzing the semantics of autonomous vehicle accidents and to present some preliminary insights for risk-neutral autonomous vehicle technologies.

Keywords : Autonomous Vehicle, Accident Reports, Semantic Similarity, Risk-Neutral

I. Introduction

According to NHTSA, 36,096 people were died by car accidents in 2019[1], and human errors were the cause of 94% of the crashes. The Thales Group[2] has also reported that 90% of traffic deaths can be avoided by autonomous vehicle. To achieve the latter, many companies have worked on the technologies but still a majority of people worries about the safety of autonomous vehicles.

In this paper, we are proposing a framework to understand the risk factors of autonomous vehicle by analyzing the autonomous vehicle accident reports from California[3]. California DMV shared the autonomous vehicle accident from 2014 to present. The full accident reports from 2019 are available on their website. We used the accident reports and applied semantic similarity analysis to understand the causes of the accidents.

On the other hands, the accident description inside the accident reports are various according to the company and personnel reporting the accidents. Therefore, naive term frequency counting cannot reveal the causality from the accident reports and semantic analysis on the accident description is needed to understand the accidents.

The Word2Vec[4,5] is a statistical method for efficiently training word embedding from a large text corpus based on deep neural-networks. The Word2Vec was proposed by Tomas Mikology et. al. at 2013 and has been popularly used as the standard word embedding technique in natural language processing community. Additionally, the Word2Vec allows the word vector algebraic operations. For example, subtracting the “man” from “King” corresponds to “Queen” after subtracting “woman” as shown in Equation (1).

$$\text{King} - \text{man} = \text{Queen} - \text{woman} \quad (1)$$

The Global Vectors for Word Representation (GloVe)[6] is an extension of the Word2Vec for efficiently learning word vectors proposed

* Corresponding Author
Ingyu Lee : Yeungnam Univ. Institute of Information System and Communication.

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03039493).

by Pennington, et. al. at Stanford. The GloVe i s an approach to combine both the global statistics of matrix factorization techniques like in Latent Semantic Analysis (LSA)[7] with the local context-based learning as in the Word2Vec. In this paper, we are using the GloVe to analyze the semantic meanings of the accident reports to get benefits from the global and local context information.

This paper consists of the followings. The research methodology and algorithm is described in Section 2 and the preliminary experimental results will be followed in Section 3. Section 4 will present the insights from the analysis and concluding remarks.

II. Research Methodology

The Figure 1 shows the overall framework of the process and the Figure 2 shows the corresponding algorithm. Firstly, we collected the accident reports from the California DMV and convert the accident description to vectorized repositories using the GloVe which is a well-known pre-trained semantic representation of word vectors as described in the previous section [6]. At the same time, we convert the questionnaires to a vectorized version using the GloVe. Finally, we compare the semantic similarity between the accident report repositories and the questionnaires. The questionnaires we used for our preliminary analysis are the followings.

- Which part of the autonomous technology is a major problem, is it from Hardware or Software?
- Which sensors are more involved with the accidents? Is it lidar, radar, camera, GPS/GNSS, or MAP errors?
- In which phases of autonomous driving technology is a major issue? Is it from a perception, tracking, localization, planning, trajectory, or control?
- Is the accident involved with pedestrian, bicyclist, vehicle or truck?

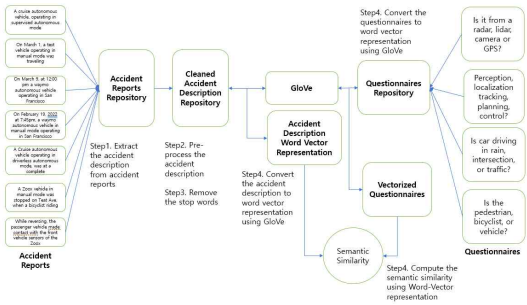

Fig. 1. Process Framework.

Algorithm 1: Semantic Analysis based on GloVe

Input: Rs, Qs, GloVe
Output: SemSim(Qs,Rs)
Let Rs is the repository of description of accident
Pre-process Rs and drop the common words
Let Qs is the repository of questionnaires
Let $GloVe$ is the vector represented trained model
for $r \in Rs$ **do**
  Convert the r into word vector representation using GloVe
end
for $q \in Qs$ **do**
  Convert the q into word vector representation using GloVe
end
while $Qs \neq$ **do**
  Remove q from Qs
  Compute the similarity on word vectors $sim(q, Rs)$
  Add the results to $SemSim(q, Rs)$
end
return $SemSim(Qs, Rs)$

Fig. 2. Semantic Analysis based on GloVe.

- What were the driving environments when the accident happened?

III. Experimental Results

We collected the autonomous vehicle accident reports from California DMV from 2019 to 2022. The total accident reports are 286 cases from 10 different companies. During the accidents, the 137 vehicles were driving in conventional mode and the other 149 were in autonomous driving mode. The detail descriptions of the data are shown in Table 1.

Then, we pre-processed the accident repository including removing symbols, tokenization, stemming, and removing common words. After that, we convert the pre-processed repository to word vector representation forms using the GloVe. We applied the same process for the questionnaires and compute the semantic

distance between accident repository with the questionnaires.

Table 1. Autonomous Vehicle Accidents Records from California in 2019-2022

	Autonomous Mode	Conventional Mode
Aimotive	-	1
Apple	1	8
ArgoAi	2	1
Aurora	-	2
Cruise	81	32
Lyft	1	7
PonyAi	5	1
Waymo	46	61
WeRide	2	1
Zoox	11	23

Table 2 shows the preliminary results of our analysis. The table shows the semantic distances between the accident repository with the questionnaires when we used 300 dimensional Glove vectors. We also used the threshold value 0.3 to get the major semantic similarity and the results are shown in the table.

At first, hardware and software shows similar tendency in terms of causing errors or accidents based on the autonomous vehicle accident repository. Secondly, camera is the major concerns following with radar. Considering the camera is the most popular sensor in almost every autonomous vehicles equipped with, the results are not surprising. Radar is also one of the most popularly equipped autonomous vehicle sensor and also causes more accidents than other sensors. Third, control phase is the biggest issue with planning phase. The errors might start with perception phase but the accident usually happen during the control phase. Sometimes, it might be a little bit difficult to figure out when the error starts with but finding the error in early phase is very important to avoid the accidents. Fourth, vehicle is the main entity involved in accidents with a pedestrian. Pedestrian causes accidents in many cases considering the semantic similarity bigger than the threshold is much higher than others. In this experiment, we did not consider the weight or severity of the accidents. Fifth, the accidents happens more in urban traffic situation with constructions around the area with unusual light condition. The latter is expected since the highway driving does not require sudden changes or movements of vehicles. However, urban driving needs all sensors are working properly and decision should be made within a short time period. Finally, Intersection is more dangerous as expected and weather are equally worse than other driving environments.

Table 2. Semantic similarity analysis on California accident reports

Questions	Semantic Similarity	Ratio (Similarity > 0.3)
H/W and S/W	H/W (0.26) S/W (0.27)	H/W (0.12) S/W (0.22)
Sensors	radar (0.29) lidar (-0.1) camera (0.38) GPS (0.10) MAP (0.25)	radar (0.39) lidar (0.0) camera (0.99) GPS (0.0) MAP (0.02)
Phases	perception (0.15) tracking (0.29) localization (-0.04) planning (0.35) trajectory (0.17) control (0.48)	perception (0.0) tracking (0.38) localization (0.0) planning (0.88) trajectory (0.0) control (0.99)
Entities	pedestrian (0.33) bicyclist (-0.02) vehicle (0.63)	pedestrian (0.83) bicyclist (0.0) vehicle (0.99)
Environments	light (0.44) construction (0.38) traffic (0.54) intersection (0.37) rain (0.28) weather (0.33) debris (0.31)	light (0.99) construction (0.99) traffic (0.99) intersection (0.90) rain (0.25) weather (0.87) debris (0.62)

Table 3 shows the differences between two companies, C and W, based on the accident

reports. According to the analysis, the W has more issues in terms of Hardware and Software, but the radar sensors in C has more issues than that of W. The C has more issues in tracking and planning than those of W. The W has more accidents involving pedestrian than that of C. The C shows better performance in driving with rain and debris condition than those of W.

Table 3. Comparison between two companies.

Questions	C (Similarity > 0.3)	W (Similarity > 0.3)
H/W and S/W	H/W (0.01) S/W (0.03)	H/W (0.17) S/W (0.36)
Sensors	radar (0.7) lidar (0.0) camera (0.99) GPS (0.0) MAP (0.0)	radar (0.15) lidar (0.0) camera (0.99) GPS (0.01) MAP (0.0)
Phases	perception (0.0) tracking (0.43) localization (0.0) planning (0.97) trajectory (0.0) control (0.99)	perception (0.0) tracking (0.25) localization (0.0) planning (0.78) trajectory (0.0) control (0.99)
Entities	pedestrian (0.87) bicyclist (0.0) vehicle (0.99)	pedestrian (0.93) bicyclist (0.0) vehicle (0.99)
Environments	light (0.99) construction (0.99) traffic (0.99) intersection (0.93) rain (0.09) weather (0.90) debris (0.39)	light (0.99) construction (0.99) traffic (0.99) intersection (0.97) rain (0.38) weather (0.85) debris (0.74)

IV. Conclusion

In this paper, we proposed a framework to analyze the autonomous vehicle accident reports using the semantic analysis to understand the accident risk factors of autonomous vehicle. We also showed some preliminary analysis results using the proposed framework. The latter will be beneficial to autonomous vehicle makers to prepare ways to avoid the accidents.

This research has a lot of flaws which should be improved with more data set. Currently, we used only four years accident reports from California which does not include all the different cases in the real driving situations. In the future, we will generate more synthetic data based on the current repository which will help to understand the risk factors of autonomous driving.

References

[1] NHTSA, “Automated Vehicle Safety”, https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety, Last accessed at April, 2022.

[2] Thales Group “7 Benefits of Autonomous Cars” https://www.thalesgroup.com/en/markets/digital-identity-and-security/iot/magazine/7-benefits-autonomous-cars, Last updated at January, 2021.

[3] California DMV “Autonomous Vehicle Collision Reports” https://www.dmv.ca.gov/portal/vehicle-industry-services/autohttps://www.overleaf.com/project/61ceb62d3f346cc334e79867nomous-vehicles/autonomous-vehicle-collision-reports/, Last accessed at April, 2022.

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean “Efficient Estimation of Word Representations in Vector Space” International Conference on Learning Representation (ICLR), 2013

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean “Distributed Representations of Words and Phrases and their Compositionality” International Conference on Neural Information Processing Systems (NIPS), pp. 3111-3119, 2013.

[6] J. Pennington, R. Socher and C. Manning “GloVe: Global Vectors for Word Representation” Empirical Methods in Natural Language Processing (EMNLP), pp.1532-1543, 2014.

[7] T. Landauer, P. Foltz, D. Laham “An Introduction to Latent Semantic Analysis” Discourse Processes, 25, 259-284, 1998.