Application of machine learning techniques - signal mapping & description generation

Degree programme: BSc in Computer Science
Specialisation: Data Engineering
Thesis advisor: Violeta Vogel
Industrial partner: HaslerRail AG, Bern

This work advances the development of a future where signals recorded from trains are automatically assigned well-fitted descriptions. Combined with a mapping between these signals, serving as a foundation for tackling the new challenges of cloud-based signal analytics.

Introduction

The world is continuing to transform; being a more cloud oriented one. So does the software used for analyzing recorded signals from trains. Event recorders (EVR) or juridical recording units (JRU) are standardly installed on every train. As their name suggests, they are responsible for recording signals such as a door being opened or closed. HaslerRail which I have partnered up with, delivers such EVR/JRU devices and distributes them under the name TELOC®. It used to be that those recorded signals were only analyzed in case of an accident, emergency brake or similar. It used to be that the data is manually extracted and analyzed on a train-by-train level. Nowadays, with the power of collaboration on the cloud, we are shifting to automated data extraction, implemented by default on every new device. Ready to be deployed when needed. The cloud-based signal analysis software provided by HaslerRail is named EVA+. It widens the perspective of analysis to a fleet level, allowing for visualizations such as heatmaps; showing areas where the wheels of the trains slip, due to losing traction and therefor wear and tear both, the train and the tracks unnecessarily.

Focus

The cryptic nature of signal names with poor standardization, currently makes it a time-consuming guessing game, understanding their underlying meaning. A vast amount of background knowledge within the railway industry is most often a requirement. Creating a mapping of signals between trains representing how similar one signal from another one is, is the follow up step in understanding signals.

Methodology

This work is driven by Machine Learning approaches as the title suggests. The image below represents the approach chosen to do so.

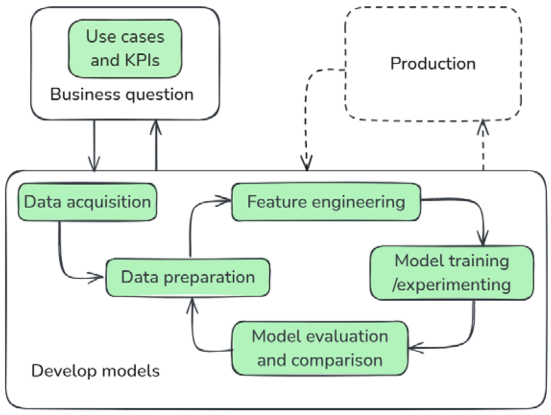
Results

The tokenizer developed to split the signal names into its meaningful parts has an accuracy of 86%. This is more than good enough to allow a locally run large language model (LLM) generating detailed description of a given signal name. Should an LLM assisted description generation not yield the expected results, a #tag like system, using all the extracted attributes will serve as a middle way. Those attributes further serve to map signals to one another. A big enough golden standard (control data set labeled by a human allowing to run the model against) representing equal signals under different names for the evaluation of the mapping has not been archived so far.

Applicaitons

The developed algorithms deployed into the existing infrastructure combined with the adjustment of the frontend would benefit EVA+. Not only have such adaptations the potential to massively flatten the steep learning curve by providing signal descriptions, but also reducing the initial setup-time needed for new users by beeing able to providing them with suggestions of commonly used visualization from others, by mapping those to the given trains.

Mirco Fabio Gall



ML life cycle inside an average organisation