Validating Predictive Models Of Evaluative Language For Controllable Data2Text Generation

Maurice Langner (Maurice.Langner@rub.de), Ralf Klabunde (Ralf.Klabunde@rub.de)

Linguistic Data Science Lab, Ruhr-Universität Bochum

Introduction

In the current upsurge of Large Language Models and their application to data2text generation, controllability is an essential topic in the light of hallucinations and toxicity of language [3]. Controllable generation of evaluative markers is a challenge in text generation in general, mainly because the motivation for an evaluative tone must be grounded in the data. We conducted an empirical study on how evaluative adverbs influence the reader's expectations on certain car features in road test reports. Additionally, we show how to use regression models for approximating the features of cars and delineate Boolean borders where data qualifies the generation of evaluative language. Finally we show how well the reader's expectations in comprehension agree with the input data and therefore validate the use of regression to approximate decision boundaries.

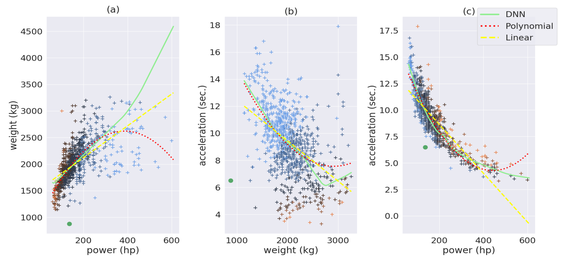
Corpus and input data

We use a corpus provided by the German Automotive Club **ADAC**, containing database entries of technical details on cars and the respective car review. We use the numerical values of car features such as *mileage* for training regression models that predict an expected value for the target feature.
For the empirical study, we extracted sentences from the car reviews that contain evaluative adverbs expressing a denial of an expected value. We devised seven different categories of sentences. Categories 2 and -2 contain evaluative adverbs and respective adjectives in the noun phrases, categories 1 and -1 contain only an adverb, categories 0.5 and -0.5 contain only the adjective(s) and category 0 is neutral. The signed categories are understood as negative, the unsigned express positive polarity. For each corpus finding, the remaining categories are constructed.

polarity	item
-2	Disappointingly, the car goes slowly from 0 to 60 mph in [...] seconds with a power output of 200 hp.
0	With a power output of 200 hp, the car goes from 0 to 60 mph in [...] seconds.
2	Amazingly, the car goes from 0 to 60 mph in only [...] seconds with a power output of 200 hp.

Regression models

In order to be able to approximate expected values of a target feature such as *acceleration*, given one or more input features, such as *maximum speed*, we implemented linear regression, polynomial regression and a DNN model. Applied to the input features, the model predicts what would be expected given our ADAC corpus data, such that a deviation from the real value in the database indicates a denial of expectation and therefore a data-driven motivation for producing evaluative language [1]. Below are given a few examples of 1 onto 1 predictions in the group of *acceleration*, *power* and *weight*.



Overall, linear and polynomial regression often underfit, while the DNN fits best (as expected). The green dot represents the Lotus Elise, which is an outlier in any regard with an exceptionally low kerb weight (1931 pounds) that allows for 'surprisingly' high acceleration (6.5 seconds) at a comparably low power of 136 hp. The graph shows that the predictions license the generation of evaluative language in any regard. For reference, chatGPT generated 'Honestly, the car's performance was underwhelming with a 6.5 second acceleration despite its 136 hp power output.' Such tasks prove that LLMs struggle to produce adequate evaluative expressions.

Comprehension study

We conducted an online comprehension study with two groups of each 50 participants from Prolific (German as first language). The web application presented randomized sentences belonging to one of the seven categories of evaluative language. Each sentence contrasts two car features, e.g. *acceleration* given *power* as in the table and the graph below. One of the features was masked and the participants task was to adjust the lower and upper threshold of the numerical interval that they judged as realistic for the masked feature given the evaluative stance of the sentence. Several groups with different feature pairs and evaluative categories were tested. The graph below shows the results for 2, 0 and -2 categorized phrases that agree with sentences in the table.

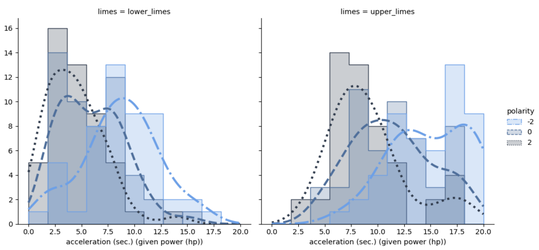


Figure 1:expectations for 2, 0 and -2

The results show clearly and consistently that evaluative adverbs of positive and negative polarity shift the reader's expectation of the value away from the neutral expression in agreement with the polarity of the expression [2]. All differences between distributions in both thresholds are highly significant (> 0.006) in both studies. The influence of adjectives shows a tendency towards the same effect but are less significant.

Matching regression with expectation

When superposing expectations collected in the comprehension study (lower graph in Figure 2) with the distribution of real data points (middle block in 2) in the corpus as well as the regression values (upmost block in 2), we see a surprisingly good match (swarm intelligence).
The maxima of the neutral curve matches the mean of the real data points well. The regression value (light blue asterisk) also agrees nearly perfectly. The maxima of the distributions for the positive and negative expressions align well with the points marking the regression value +/- the standard deviation (blue triangles pointing upwards and downwards). We hence consider these as binary decision boundaries for generating a positive or negative stance in the output text. Consequently, this proposes to use regression methods for determining a mismatch between real value and expected value at the level of document planning. Therefore, the respective information on evaluative language use can be included in AMRs as inputs to Text2text transformers that generate the surface text.

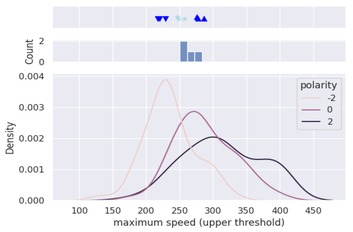


Figure 2:data match

Conclusion

We have shown that regression with DNNs is a suitable means to approximate expected values in a D2T corpus with numerical features. These expectations serve as an anchor point for comparing the real values and thus determine whether there is data-driven motivation for producing evaluative language. We furthermore empirically showed that evaluative markers shift reader's expectations. Expectations, real data and predictions fit surprisingly well, validating our approach to determining binary decision boundaries for generation of positively or negatively polar evaluative stance in d2t systems. Including these findings in AMRs also allows for using transformers for surface text generation in neural NLG pipelines.

References

[1] M. Langner and R. Klabunde.
Realizing a denial of expectation in pipelined neural data-to-text generation.
In R. Confalonieri and D. Porello, editors, *Proceedings of the 6th Workshop on Advances in Argumentation in Artificial Intelligence 2022 (AIxIA 2022)*, 2022.

[2] S. Mahamood, E. Reiter, and C. Mellish.
A comparison of hedged and non-hedged nlg texts.
In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, page 155–158, USA, 2007. Association for Computational Linguistics.

[3] C. Rebuffel, M. Roberti, L. Soulier, G. Scoutheeten, R. Cancelliere, and P. Gallinari.
Controlling hallucinations at word level in data-to-text generation.
CoRR, abs/2102.02810, 2021.

RUHR UNIVERSITÄT BOCHUM RUB LDSL Linguistic Data Science Lab