Realizing a Denial of Expectation in Pipelined Neural Data-To-Text Generation

Maurice Langner[1], Ralf Klabunde[1]

[1]*Department of Linguistics, Ruhr-Universität Bochum, Germany*

Abstract

This paper aims at the generation of *denials of expectations* in the domain of vehicle reviews. A denial of expectation expresses an apparent contradiction between some probabilistically motivated rule and a current circumstance expressed by a contrastive sentence. For generating such an argumentative sentence, we present a new approach for content selection in a neural data-to-text generation framework. In addition to selecting relevant information from tabular data that should appear in the text, further methods are required for determining evaluations that are rooted in this data, but express individual appraisals of the respective vehicle. We show how a content selection module is able to decide when expressing a denial of expectation. We use multi-label and binary classification for content selection on automatically extracted training data and Random Forest Regression with varying knowledge limitation for predicting expectations about feature values. These predictions are compared to manually annotated corpus instances of contrast relations in order to show that the concept of denial of expectation is a reasonable approach to determining contrasts and evaluative content at the early stage of content selection.

Keywords

Natural Language Generation (NLG), denial of expectation, evaluative adverbs, content planning, document planning

1. Introduction

When arguing in favor of (or against) a new device in order to convince a potential buyer to purchase (or disregard) that device, it is often useful to compare features of that device with feature-related expected values, and to indicate possible consequences of discrepancies between real and expected value. One of the linguistic means for realizing this is the so-called *denial of expectation*, which is an apparent contradiction between some rule, be it grounded in domain knowledge, personal experience or norms of social behaviour, and a current circumstance expressed in the corresponding sentence. In sentences like:

1. *The sports car is not that big in terms of external dimensions, but it does weigh quite a bit.*
2. *Although the sports car is not that big in terms of external dimensions, it does weigh quite a bit.* (corpus example, engl. translation)
3. *(Un)fortunately, the sports car does weigh quite a bit.*

we have different linguistic realizations of the denial of expectation. Example (1) expresses a contradiction that is based on technical specifications and their consequences for a car's weight

6th Workshop on Advances In Argumentation In Artificial Intelligence (AI³ 2022)

Maurice.Langner@rub.de (M. Langner); Ralf.Klabunde@rub.de (R. Klabunde)

CEUR Workshop Proceedings (CEUR-WS.org)

and the weight of the corresponding sports car. According to the external dimensions, weight is expected to be lower than the real value. However, the conjunction *but* is ambiguous, it can also be used to express a contrast that is grounded in other means than rule violation (as in *John is short but Lea is tall*). Contrary to *but*, the concessive conjunction *although* in example (2) can only be used to express the denial of expectation. Adverbs like *(un)fortunately* are typically not considered as items expressing a denial of expectation, but in fact it is possible that they are grounded in the same semantic mechanism the conjunctions are based on. For example, *unfortunately* expresses that some part of the proposition that is in the scope of this adverb is typically evaluated neutrally, and in this special case it has been shifted to the negative end of this evaluation scale. Example (3) can be interpreted in the following way: The writer's use of this sentence is based on her estimation of the weight of sports cars, which is grounded in their typical dimensions and other technical factors, but the sports car that is relevant here violates this expectation. It should be noted that other uses of adverbs of this type are not grounded in a denial of expectation, but in other incompatibilities. For example, *unfortunately this car is red* might just express the difference between the speaker's colour preference and the actual colour of the car. Therefore, in this paper we consider only those adverb uses with a reference to a denial of expectation.

A denial of expectation as a special case of linguistically expressed contrast is inherently argumentatively motivated [1], since the addressee's inferred opposition between both conjuncts of such a sentence is quite decisive for the evaluation of the object or state of affairs at hand. Using Toulmin's argumentation schemes [2], examples (1) and (2) are claims based on expected and real values for the car as grounds, and the underlying expectation functions as the warrant.

In this paper, we are analyzing denials of expectations from a Natural Language Generation (NLG) perspective. Hence, we do not determine plausible oppositions for a given denial of expectation, but we are motivating decisions for realizing such a sentence. As application domain, we are using driving reports and an associated database with technical specifications for this. During content selection – the first stage in a pipelined NLG system – possible data correlations must be determined, together with a violation of this correlation in the current message to be verbalized.

2. Related Work

In traditional NLG pipelines for data-to-text generation, an interpretation module that encodes domain-expert knowledge decides which information should be contained in a message within the document plan. The avoidance of hand-coding a heuristic selection module is desirable for time and effort reasons and can be realized with the help of data-driven learning techniques. A premise to this is the availability of a reasonable amount of data. In general, data-to-text generation is the field of transforming tabular data into surface text [3, 4, 5]. Due to the rise of neural networks, traditional pipeline approaches were abandoned in favor of end-to-end trainable encoder-decoder networks [6, 7, 8], which do not separate content selection and document planning from surface realisation, often at the price of loosing controllability of generated content.

As an improvement in regard to content and informational correctness, copy mechanisms

[9, 10] are employed in order to directly copy content from input data to the output text in order to enhance correctness during the generation process, while maintaining end-to-end trainability. Mei et al. [7] use an intermediate aligner step between encoding and decoding in order to integrate more controllable content selection into the end-to-end network. As Wiseman et al. [8, p. 2259] point out, such a model performance is well below gold standard on the RotoWire dataset, regarding both, content selection and text coherence, although the copy mechanisms clearly improve the vanilla encoder-decoder networks in regard to BLEU. Their results also indicate the absence of correlation between the models' precision and recall for content selection and BLEU scores.

Another important point to mention is that most models are trained to generate short phrases only, e.g. short biographies from Wikipedia tables [6, 11]. Coherence of discourse and information structure decreases with increase of text length, which makes the encoder-decoder models a non-optimal choice for the generation of longer texts.

Ferreira et al. [4] propose a re-modularization of neural generation networks, chaining separately trainable and evaluable networks that are specialized for the different tasks of content selection, document planning and surface realisation. They show that these pipelined neural generation models outperform end-to-end networks, especially on unseen data, where the latter tend to produce topic-unrelated, incoherent texts and hallucinations. Turning back to a pipelined neural generation system necessitates to find a suitable content selection model that determines what the document plan shall contain.

Many papers have been published on research how to punish toxicity and inappropriate language use in neural NLG, resulting in more neutral word choice, but to our knowledge, the controlled generation of a denial of expectation as a model of contrast relations and evaluative content has not yet been dealt with in the context of neural NLG.

3. Methodology

We present a classification approach to content selection on automatically augmented data which predicts what information shall be present in the document plan. Furthermore, we use regression models in order to predict, given different knowledge limitations, whether the information to be produced in the surface text agrees with expectations about the information in order to determine in a data-driven manner whether expressive content is legitimate to use for putting the information into perspective or not.

3.1. Domain

A domain for data-to-text generation needs tabular data from which surface text shall be produced. The German Automotive Club ADAC supplied us with a proprietary data set of 1300 road test reports with 3000 to 6000 words including the respective database with 127 technical and economic properties for each vehicle. The road test reports are written by domain experts and contain a subset of the properties given in the database as well as surplus information, e.g. a guess of resale values of used cars, or their opinion on vehicle characteristics. Therefore, the texts also include evaluative information which is naturally grounded in subjective estimation. These evaluative decisions, in turn, will be reflected in the use of corresponding evaluative

- *The A3 completes the intermediate sprint from 60 to 100 km/h in a brisk 6.1 s.*
- *The petrol engine completes the intermediate sprint from 60 to 100 km/h in 5.6 s.*
- *The 108 kW/147 hp mean that the simulated overtaking maneuver (sprint from 60 to 100 km/h) ends after just 6.0 seconds.*

Table 1
Examples of verbalisations for *accelaration* (English translation of German corpus examples)

lexical items like adverbs or conjuncts. There is no information in the database which properties were named in the respective road test report. In order to produce a data set that is usable for a learning-based content selection module, we need a function that maps a road test review to a binary decision on properties in the database that either marks the absence of the respective piece of information from the text with 0 or its presence in the road test report with 1. The result is a triple (T, D, M) for each road test report T consisting of the text T, a database row D, which is a 127-tupel of mixed type information on the vehicle, and a map M which is a 127-tupel of binary values for the presence or absence of each property in D. We selected a subset of 15 properties from the database which are related to the technical details on engine, chassis and driving performance. In order to determine M, we extracted a sample of 200 database rows and respective test reports and build a heuristic information extraction module, which uses the tracking of keywords and pattern matching for determining the presence of a piece of information in the text. At the same time, we manually annotated a set of 50 texts from the corpus in regard to the presence of the 15 target features we selected beforehand. This set of annotated texts serves as the gold corpus for evaluating the heuristic IE module. The set of 200 reports for building the IE heuristic and the set of 50 texts for manual annotations are disjoint, and only the remaining 1050 road test reports and database rows are used for the machine learning models of content selection and denial of expectation.

3.2. Information Extraction

For information extraction, we searched the subset of 200 reports for keywords and patterns in which the 15 target properties are verbalized in order to extract rules in form of regular expressions. Fortunately, across test reports, the verbalisation of each piece of information is relatively uniform, despite the self-evident differences between properties (see Table 1).

As the examples in Table 1 show, the data point *acceleration* is stereotypically verbalized as a quantity of seconds the car needs to accelerate from 60 to 100 km/h. Additionally, key words like *sprint* and *overtaking maneuver* often indicate the target information *acceleration*. Not all of the features are as straightforward in extraction as *acceleration*. The feature *displacement*, for example, occurs in highly aggregated compounds like *the 2.5 litre turbo combustor motor* and is less accurately extractable.

This heuristic information extraction approach is applied to the remaining 1050 pairs of road reports and database rows in order to produce the necessary training instances for neural content selection. The output is a table containing 15 columns such that for each road test report there is a 15-tuple of binary values that indicate whether the respective property was found in the text or not.

3.3. Content Selection

Content selection has only little bearing in neural end-to-end networks and encoder-decoder models, especially in regard to the fact that the generated texts often only comprise a few sentences or a single paragraph [6, 12, 13, 14]. In encoder-decoder systems even with copy mechanism, the generated texts loose coherence and informational correctness when unfolding the texts. Even the most recent GPT-3 models, despite perfect grammatical correctness and language style, show rather weak correctness of information. Puduppully et al. [15] succeed in generating longer texts in a domain with an average of 330 words. The authors built an encoder-decoder model that learns relations between NBA basketball game records (numerous repetitive events with 4 features each) and the respective verbalisations in the corresponding game summary text. The difference between their dataset and the car review corpus is the structural non-repetitive nature of the latter – there are no repetitive events that can be mapped onto a summarization; each property in the ADAC database is a unique informational unit that is dealt with mostly separately or in paragraphs containing several related properties. Ferreira et al. [4], who competitively evaluate end-to-end models against their pipelined GRU and Transformer, use the webNLG corpus, which maps RDF triples onto a short paragraph containing the information. The authors do not deal with content selection in the actual sense, but rather content ordering of the preselected triples in the corpus.

Our content selector takes the very first step in an NLG pipeline and decides which pieces of information shall be integrated into the document plan, which in turn, after adding the database values for the respective properties, can also be represented as RDF triples, based on the ADAC database and the human written car reviews. We assume that the authors of the car reviews have domain-specific reasons for choosing certain properties from the database and leaving others aside. For example, *horse power* is used in nearly all texts, whereas *valves* are never even mentioned. Hence we assume that given the texts, the database entries, and our generated training data, we are able to train a neural classifier that is capable of predicting which features should be produced. This would mean that the classifier could encode domain expert knowledge on what to say by finding the same patterns in the technical data as the experts would do.

Before using classification, we tried to determine feature importance patterns in the data. Assuming that some of the technical details are interdependent, not all pieces of information are relevant for predicting the presence of each property. This is also reasonable from an engineering perspective, since properties of the motor might condition each other, while these have no influence on the design of the interior.

The heatmap in Figure 1 reveals some interesting relations. The slightly recognizable line, which resembles a linear function with gradient -1, is self-evident, since each piece of information is mapped onto itself. More important are the really strong outliers beside this line, showing that each data point only has a few, but important points that condition the target point's presence.

According to the feature importance and usability of feature type, we selected 40 categorical and numerical values for predicting the 15 properties. The 15 data points were first label-encoded and then one-hot encoded. By applying the heuristic information extraction module to the texts, we gained about 931 usable text array pairs. The remaining car reviews only provided partial or fragmental test reports, which we therefore excluded. These arrays contain 15 binary

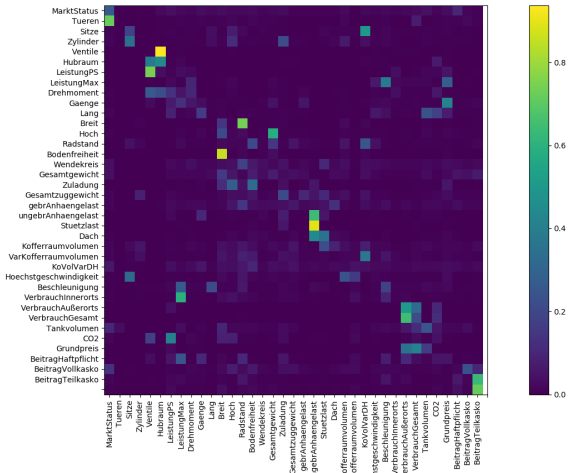

Figure 1: Heatmap of feature importance for content selection

values [1,0] indicating presence or absence of the target properties.

An important point to mention here is that the input to the classifier is not a tensor with 15 elements of the binary decision on presence, but the real values of the 40 relevant properties, for e.g. *HP*, *torque*, *weight* and so on taken from the database. The network is therefore trained on determining the presence or absence of a piece of information on the basis of the vehicle's data, and not on the presence or absence of the other data points.

A multi-label classifier [16, 17] with a dense network of 5 layers (Keras, Tensorflow), RELU activation function and sigmoid activation as final layer did not lead to satisfying results. We received only a few percent of accuracy per feature when predicting all 15 target features at once. Reasonable accuracy is only reachable when performing binary classification for each target feature separately.

3.4. Surface Realisation

For surface realization we used the multi-lingual T5 (MT5, [18]); the model is fine-tuned on the German webNLG corpus ([19]) with varying numbers of aggregated triples. In order to fine-tune the MT5 for car reviews, we extracted triples from our ADAC corpus of car reviews by matching the corresponding surface text with the features and values given in the corresponding database. The generated sample below shows what the neural pipeline generated on the basis of the technical data of the *Renault Clio TCe 130 GPF*. The content selector predicted to generate the features *torque*, *supercharging* and *motor power*, which were aggregated with feature values from the database, then linearized by the document planner and fed into the custom MT5 in order to produce the output surface text. The reference text from the original car review is also

listed for comparison.

Input:	type: combustor; motor power: 130 HP; cubic: 1.3L; torque: 240 Nm; ...
Content:	torque, supercharging, motor power
Planning:	Renault Clio TCe 130 GPF | torque | 240 Nm Renault Clio TCe 130 GPF | supercharging | Turbo Renault Clio TCe 130 GPF | motor power | 130 PS (torque, supercharging, motor power (HP))
Output:	*Der Turbobenziner leistet 130 PS und entwickelt ein maximales Drehmoment von 240.* 'The turbocharged petrol engine makes 130 PS and produces a torque of 240.'
Reference:	*Der 1,3 Liter Vierzylinderbenziner leistet dank Turboaufladung 130 PS und entwickelt ein maximales Drehmoment von kräftigen 240 Nm, das bereits bei 1.600 Umdrehungen pro Minute bereitsteht.* 'The 1.3 liter four-cylinder petrol engine has, thanks to supercharging, 130 PS and produces a maximal torque of powerful 240 nm, which is available already at 1.600 turns per minute.'

3.5. Denial of Expectation as Contrast Relation

Vehicle testers have certain expectations about features of a vehicle when writing the reviews. These are based on domain-specific experience, such that experts have an intuition what weight a sports car should have, given a set of exterior dimensions, motor block size and so on. Many evaluations and contrasting details are formulated in regard to technical details which contradict each other or do not agree to the expected value, both in positive and negative polarity.

A denial of expectation when the expected value does not agree to the real value may be a trigger for generating a concessive or evaluative marker that signals this mismatch to the reader. Independent of the question how such a contrast or evaluation shall be lexicalised, the underlying semantic mechanism for determining such a mismatch is data-driven and is, therefore, installed at the interface of content selection and document planning.

A straightforward approach to predicting values, given a set of features, is regression. Using a Random Forest regression implementation (Scikit, 100 estimators), which often faces limitations for linguistic data [20], but is a reasonable choice for our task at hand, we predict numerical values for each of the 15 target features on the basis of the set of residual features in the database (excluding the target feature). By comparing these predicted values to the real values, we can determine whether there is a significant deviation that may trigger a contrast relation or the usage of an evaluative adverb.

From the corpus of 1300 car reports we automatically extracted instances of concessive markers, namely *obwohl* ('although'), and the evaluative adverbs *erstaunlicherweise* ('surprisingly'), *bedauerlicherweise* ('regrettably') and *leider* ('unfortunately'), and filtered them by the assessibility of the contrasting information in our database. Contrast relations, to which the denial of expectation applies, but which cannot be modelled in a data-driven way, dealed with subjective driving experience, e.g. the adjustability of the arm rest or the noise level of the motor, or would necessitate additional reasoning or information we did not have access to.

19 instances of concessives and evaluative expressions remain for analysis, listed in Table 4. Note that we confined the evaluative adverbs to those cases with a denial of expectation as underlying contrasting motivation. This is a small number of instances, but it relates to the fact that we only searched for the specific markers above. The polarity defines either positive or negative direction of the denial of expectation, whereas the arrows beside the source and target properties names indicate whether their values need to be high(-er) (↑) or low(-er) (↓) in order to match the polarity.

Furthermore, Table 4 lists four numerical values: the real values for the target data point which were retrieved from the database and three possible thresholds for modeling denial of expectation. These are the predicted value on the basis of all 40 relevant data points, the 'naive' prediction where the input to the regressor is limited to the source information the authors name in their contrasting relation, and finally the average value of the target data point across all vehicles in the database. The numerical values that differ with correct sign from the real value such that the denial of expectation can be captured by the threshold are printed in boldface.

A difference to the real value can be interpreted as a denial of expectation. The expected value is higher or lower than the real value - the sign of the difference and semantic relation between the source and target features, e.g. higher *HP* may entail higher *maximal velocity*, determine the polarity of the whole expression. The polarity (+/-) determines whether the denial is a surprise (+) or a disappointment (-). For example, instance (8) in Table 4 can be paraphrased as *although the car has many horse powers, the mileage is comparably low.* The expected values (both regression values, 5.94 and 6.1) are higher than the real value (5.9), meaning that the real *mileage* falls below what is expected for a car of the respective *HP*. Since less *mileage* is positive, the polarity of the whole expression is positive.

Before going into the analytic details, a few technical relations need clarification in order to understand the expression for the target property *range(↑)@mileage(↓)*. Five of the contrasting relations are established between the size of the fuel tank and the possible range of the car, indicating that despite a given comparably small tank, a high range is possible. The range as a numeric value is listed in the database only for electric cars, but for combustors, the range can be inferred by the tank size and the minimal consumption. Range and mileage are anti-proportional, meaning that the smaller the mileage, the higher the possible range at constant tank size. Our target feature is therefore the fuel consumption, where less is generally considered better. Furthermore, *acceleration* should be explained. Acceleration is quantified in seconds needed to reach a certain velocity, e.g. 100 km/h. Higher acceleration leads to fewer seconds needed.

4. Evaluation

4.1. Information Extraction

In Table 2 the results of the heuristic information extraction approach tested on the gold corpus of 50 reports are listed. We received perfect or near perfect values for *horse power*, *torque*, *fuel consumption*, *acceleration* and *brakes*. While *velocity* and *transmission* still show good recall and precision values around 0.85, *displacement* and *supercharging* reveal that the highly aggregated expressions for the motor description are too diverse to be captured as accurately as the other

feature	precision	recall	counts
horse powers	1.0	1.0	927
torque	0.97	1.0	805
fuel consumption	0.97	1.0	926
acceleration	1.0	0.97	419
motor type	0.97	0.91	800
price	1.0	0.89	926
brakes	0.97	1.0	831
max. velocity	0.87	0.91	399
transmission	0.85	0.85	766
cylinders	0.84	0.88	119
displacement	0.65	0.77	288
supercharging	0.6	0.46	461
weight	0.25	0.27	889
valves	-	-	0

Table 2
Information Extraction: Precision and recall for extracted properties, plus occurence counts in the resulting training data

feature	acc	err	precis.	recall
horse powers	0.997	0.005	0.996	0.997
valves	0.997	0.004	0	0
fuel consumption	0.992	0.013	0.99	0.99
price	0.991	0.016	0.99	0.99
weight	0.914	0.112	0.96	0.97
motor type	0.815	0.098	0.88	0.85
brakes	0.797	1.0	0.90	0.86
cylinders	0.79	0.12	0.11	0.18
torque	0.765	0.173	0.88	0.87
transmission	0.72	0.184	0.79	0.75
displacement	0.686	0.11	0.29	0.24
supercharging	0.61	0.075	0.68	0.53
max. velocity	0.56	0.054	0.52	0.38
acceleration	0.53	0.071	0.45	0.48

Table 3
Content Selection: Accuracy, precision and recall values for feature-wise binary classification

attributes. For *valves* we cannot offer data since the feature did not occur in the reviews at all, which shows its irrelevance to the reader from the domain experts' perspective. The low recall and precision values of *weight* is due to the rather indefinite nature of the car's weight. Different weight-related features exist, e.g. trailing load, maximal allowed weight of cargo on the roof or in the trunk, and sums of varying subsets of those are used in the text. The weight of the car is therefore complicated to distinguish from other weight-related expressions.

4.2. Content Selection

For evaluating the content selection module, we used 8-fold cross validation and calculated the average scores. In Table (3), the evaluation metrics for each feature are listed separately. The features *acceleration* and *max. velocity*, which were very well recognized by IE, are the properties with worst prediction accuracy and only marginally better than random. The data points *supercharging* and *displacement* agree in accuracy with the low precision and recall values the IE already suggested. *Horse power* and *valves* are predicted with best accuracy, which is explainable by the simple categorical use of the former and the utter absence of the latter, making their selection deterministic.

4.3. Surface Realisation

The output of the system is evaluated against the original human-produced car reviews using BLEU. Surprisingly, the model reaches 0.57 BLEU despite the small data set. Depending on the order of triples, the BLEU score may drop to 0.19 (random order), showing that the linear order, based on topic extraction, is of essential importance for the successful transformation into surface text. This is a clear indication that the discourse planner plays a central role in the acceptance of the generated texts.

4.4. Denial of Expectation as Contrast Relation

Figure 2 and 3 depict the sampling variance [21] of the Random Forest regressor.

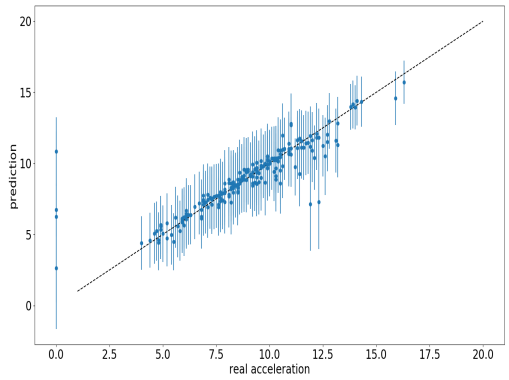

Figure 2: Confidence intervals of *acceleration*

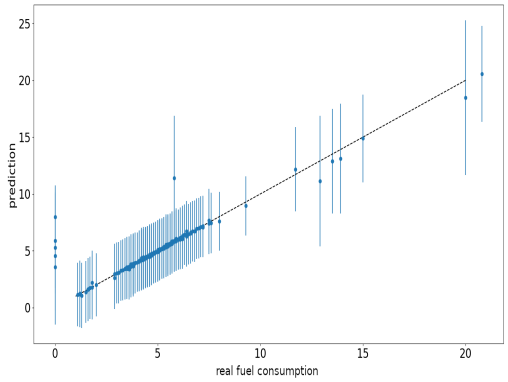

Figure 3: Confidence intervals of *fuel consumption*

The graphs show the error of the Random Forest Regression model for the two features *fuel consumption* and *acceleration*, which are from the upper and the lower end, respectively, of the accuracy scale for content selection. Figure 3 shows that the predicted values agree perfectly with the real values of fuel consumption, and the confidence intervals are uniform, at least in the area lower than 9 liters of fuel per 100 km. Above this threshold, only few data points are available in the data, for which the prediction is much less accurate and the confidence much

id	token	pol.	source property	target property	real	threshold for denial of expect. pred.	naive	avg.
1	obwohl	+	fuel tank(↓)	range(↑)@mileage(↓)	4.1	4.1	**4.4**	**5.4**
2	obwohl	+	fuel tank(↓)	range(↑)@mileage(↓)	6	5.997	5.7	5.4
3	obwohl	+	fuel tank(↓)	range(↑)@mileage(↓)	3.9	**3.98**	**4.3**	**5.4**
4	obwohl	+	weight(↑)	vehicle payload(↑)	520	549.18	545.5	**497**
5	obwohl	+	weight(↑)	acceleration(↑)	7.8	**7.93**	**8.6**	**9.17**
6	obwohl	-	dimensions(↓)	weight(↑)	2173	**2134**	**1985**	**2028**
7	obwohl	+	fuel tank(↓)	range(↑)@mileage(↓)	3.7	3.53	**4.93**	**5.4**
8	obwohl	+	hp(↑)	mileage(↓)	5.9	**5.94**	**6.1**	5.4
9	obwohl	+	weight(↑)	acceleration(↑)	7.6	**7.7**	**8.8**	**9.17**
10	obwohl	+	fuel tank(↓)	range(↑)mileage(↓)	5.1	5.08	4.7	**5.4**
11	obwohl	+	weight(↑)	mileage(↓)	4.8	4.78	**8.3**	**5.4**
12	obwohl	+	weight(↑)	acceleration(↑)	5.2	4.95	**6.75**	**9.14**
13	leider	-	–	price(↑)	31900	35120	–	42250
14	leider	-	–	price(↑)	26295	**26161**	–	42250
15	leider	-	–	mileage(↑)	5.2	5.149	–	**5.4**
16	obwohl	+	Supercharging(↓)	torque(↑)	213	**195.86**	**150.0**	301.8
17	obwohl	+	displacement(↓)	max. velocity(↑)	182	**179.95**	**175**	200
18	obwohl	+	supercharging(↓)	acceleration(↑)	6.1	5.9	**6.2**	**9.17**
19	erstaunl.	+	hp(↓)	trailing payload(↑)	4900	**3551.15**	**1100**	**1719**

Table 4
Contrasting relations and evaluative adverbs from the ADAC car review corpus

weaker. A less accurate picture is drawn by the predictions for *acceleration*, which are still close to the real values, but fewer predictions are perfectly on point. There is also more variance in the confidence intervals, but in contrast to *fuel consumption*, there are fewer outliers and the values are spread across the interval of 2.5 seconds to 15 seconds in a more balanced way.

On average, regression seems to yield good results when predicting the 15 target features on the basis of the 40 most relevant features in the database. But does regression model the domain experts' decisions in a sufficiently accurate way? A closer look at instances of denial of expectation in the corpus sheds light on the relation of author expertise and the motivation for generating this kind of contrast.

As the column of predicted values in Table 4 displays, many of the expected values related to fuel consumption are modeled nearly perfect on point, leaving no or only marginal differences. Reducing the input features of the regression model to the source information causes more deviation, which allows us to model contrast with 'naive' prediction with limited knowledge. The average mileage across all database entries is also a good threshold estimator. The outlier in regard to mileage is instance 2. The original text gives a reasonable explanation for this - the scale by which the positiveness of reduced fuel consumption is explained is a direct comparison to the previous version of the same car model. The newer version has a smaller tank, but longer range. Therefore, the contrast is triggered by the direct comparison of tank and mileage of two versions of the same car model.

All contrasts concerning *acceleration* can either be modeled with the predicted value or the naive prediction that is limited to the knowledge given by the respective source information. The average of *acceleration* also captures all instances correctly. Example 3, stated in section 1, is the only instance of *obwohl* with negative polarity, denying the expectation of a lower weight given

the comparably small dimensions. This contrast is also correctly modeled with all of the three possible thresholds. The evaluative adverb *leider*, which semantically expresses a negative point, is not correctly predictable with the regressed values. Only for the mileage-related instance (15) the average seems to be a reasonable threshold. The other instances of *leider* deal with price, which is highly dependent on build quality, brand and prestige and therefore the possibly most problematic feature for evaluative content. The average price is not a good estimator in this case. Instances (16) and (18), contrasting the lack in *supercharging* with surprisingly good *torque* and *acceleration* values, are captured by naively predicted values, the former even by the value predicted on the whole relevant dataset. Example (17) contrasts the minor motor *displacement* with a surprisingly high *maximum velocity*, which is captured by both predictions, while the average value is far away from proximity. The contrast relation in (19), marked by *erstaunlicherweise* ('surprisingly'), deals with an extraordinarily high payload given a rather low *HP*. All thresholds capture this contrast correctly, while the regressed value with full input features is still the closest to the original value. The usage of *erstaunlicherweise* instead of using *obwohl* may indicate that the authors have a proper classification of contrast markers and evaluative adverbs that express a certain degree of deviation from the expectation – the distance to the expected value in either positive or negative direction may trigger the usage of an expression that semantically quantifies this distance.

5. Conclusion

First, the heuristic approach to information extraction and data augmentation introduces noise into the training data, possibly to the same extend to which errors occur in the manually annotated data. The consequence to be drawn from this is that any network trained on this data may also incorrectly predict on the basis of error-prone input. The only limitation this imposes on the neural models is that it cannot outperform the correctness of neither heuristic nor the human annotations, which is an inherent issue of the challenge to apply machine learning models to noisy data.

Regarding the content selection module in general we can state that the performance of the content classifier is acceptable, given the small data set and the complexity of the task of both IE and annotation, which produced the model input. The content selector can, for unseen data, generalize from encoded input properties, and predict the presence of the pieces of information in the text to be generated. Although the accuracy values for some properties are still imperfect, this is a huge leap towards modeling domain expert knowledge for content selection with minimal resources of annotations. As already mentioned, the error rate of the heuristic IE approach is passed on to the content selector, which also entails that its accuracy will presumably improve with increasing recall and precision of its input.

In regard to surface realisation, it is important to mention that using Transformer models for producing surface text has a downside; neural models often hallucinate facts [22] and different methods have already been applied to prevent it [23]. A special case is expressive content such as evaluative adverbs and contrast relations, which go beyond the purely propositional content. Intentional production of such expressive, non-propositional, constructions means that any sort of non-at-issue content should be suppressed for decoding where the input does not demand

for such verbalisation. Measures against hallucination intend to do that, but for expressives it would work on word level only. Measures against hallucination that base on fact-checking and ranking cannot be applied here, because non-at-issue content only puts facts into perspective; it reflects the author's opinion. In training instances, where the input data motivate expressive content, the expressive content in the text output is either preserved, or it has been enriched with it in case no expressives are present. This training includes the systematic annotation of non-at-issue content [24]. Further research will show how such an approach dovetails with methods to minimize fact hallucination.

Summing up the empirical findings on modeling the denial of expectation, we can state that 42% of the evaluative expressions and contrasts we dealt with are explainable through regression of the target feature with full domain knowledge. Limiting the knowledge to the source features of the contrasting relations improves the coverage to 86%, excluding the instances of *leider*, where no source features are mentioned explicitly (72% including 13, 14 and 15). The average value as a threshold scores differently for different features. Although no reliable conclusion can be drawn from the statistics due to data sparseness, features like $fuelconsumption$ and *acceleration* seem more comparable to a global average than details like *torque* and *price*.

A very interesting point is the difference between predictions on the full set of input features and the knowledge limitation of 'naive' prediction the authors of the report anticipate. The empirical data shows that limiting the input knowledge of the regression to the features on the basis of which the car reviewers make their assumptions doubles the percentage of correct predictions. Due to data sparseness, we only have a very small number of instances we can use for evaluation. Nevertheless, the score of 42% and 86% respectively exceed our expectations and are a good indicator that the acquisition of more instances for evaluation will validate our findings and the adequateness of our modeling approach.

6. Future Works

A topic we did not yet address in regard to the generation of denials of expectation are false positives. Regression may determine a trigger for evaluative expressions where no such expression shows up in the corpus of texts. There might be non-linguistic motivations behind (not) producing evaluative content, e.g. an economic bias or personal preferences. These motivations cannot be modeled, but may explain a mismatch between observed and data-driven usage of evaluatives. This amounts to the necessity, once we have built a data set of verbalisations of properties containing both, neutral ones and those enriched with evaluative content, to determine the minimal difference between expected value and real value that triggers a contrast or evaluative adverb, such that the distribution of evaluatives in the empirical data best matches their distribution in the document plans.

With more data, we are sure to be able to determine these property-wise minimal deviations and deploy them for determining evaluative content at content selection level.

References

[1] G. Winterstein, What *but*-sentences argue for: An argumentative analysis of *but*, Lingua 122 (2012) 1864–1885.

[2] S. Toulmin, The Uses of Argument, Cambridge University Press, Cambridge, 1958.

[3] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, 2000.

[4] T. C. Ferreira, C. van der Lee, E. van Miltenburg, E. Krahmer, Neural data-to-text generation: A comparison between pipeline and end-to-end architectures, EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference (2020) 552–562. doi:10.18653/v1/d19-1052. arXiv:1908.09022.

[5] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 1–64. doi:10.1613/jair.5714. arXiv:1703.09902.

[6] R. Lebret, D. Grangier, M. Auli, Neural text generation from structured data with application to the biography domain, EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings (2016) 1203–1213. doi:10.18653/v1/d16-1128. arXiv:1603.07771.

[7] H. Mei, M. Bansal, M. R. Walter, What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment, 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference (2016) 720–730. doi:10.18653/v1/n16-1086. arXiv:1509.00838.

[8] S. Wiseman, S. M. Shieber, A. M. Rush, Challenges in Data-to-Document Generation (2017) 2253–2263.

[9] S. Gehrmann, F. Z. Dai, H. Elder, A. M. Rush, End-to-End Content and Plan Selection for Natural Language Generation, E2E NLG Challenge System Descriptions (2018) 46–56. URL: http://www.macs.hw.ac.uk/InteractionLab/E2E/final_papers/E2E-HarvardNLP.pdf. arXiv:1810.04700v1.

[10] A. Moryossef, Y. Goldberg, I. Dagan, Step-by-step: Separating planning from realization in neural data-to-text generation, in: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, volume 1, Association for Computational Linguistics (ACL), 2019, pp. 2267–2277. arXiv:1904.03396.

[11] A. Chisholm, W. Radford, B. Hachey, Learning to generate one-sentence biographies from Wikidata, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 633–642. URL: https://aclanthology.org/E17-1060.

[12] T. Liu, K. Wang, L. Sha, B. Chang, Z. Sui, Table-to-text generation by structure-aware seq2seq learning, CoRR abs/1711.09724 (2017). URL: http://arxiv.org/abs/1711.09724. arXiv:1711.09724.

[13] L. Sha, L. Mou, T. Liu, P. Poupart, S. Li, B. Chang, Z. Sui, Order-planning neural text

generation from structured data, in: AAAI, 2018.
[14] L. Perez-Beltrachini, M. Lapata, Bootstrapping generators from noisy data, in: NAACL, 2018.
[15] R. Puduppully, L. Dong, M. Lapata, Data-to-text generation with content selection and planning, 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019 (2019) 6908–6915. doi:10.1609/aaai.v33i01.33016908. arXiv:1809.00582.
[16] D. Gkatzia, H. F. Hastie, An ensemble method for content selection for data-to-text systems, CoRR abs/1506.02922 (2015). URL: http://arxiv.org/abs/1506.02922. arXiv:1506.02922.
[17] C. Kelly, A. Copestake, N. Karamanis, Investigating content selection for language generation using machine learning, in: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), Association for Computational Linguistics, Athens, Greece, 2009, pp. 130–137. URL: https://aclanthology.org/W09-0623.
[18] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer (2020). URL: http://arxiv.org/abs/2010.11934. arXiv:2010.11934.
[19] T. C. Ferreira, D. Moussallem, S. Wubben, E. Krahmer, Enriching the WebNLG corpus, INLG 2018 - 11th International Natural Language Generation Conference, Proceedings of the Conference (2018) 171–176. doi:10.18653/v1/w18-6521.
[20] S. Gries, On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement, Corpus Linguistics and Lingustic Theory 16 (2019). doi:10.1515/cllt-2018-0078.
[21] S. Wager, T. Hastie, B. Efron, Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, Journal of machine learning research : JMLR 15 (2014) 1625–1651.
[22] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, CoRR abs/2202.03629 (2022). URL: https://arxiv.org/abs/2202.03629. arXiv:2202.03629.
[23] C. Rebuffel, M. Roberti, L. Soulier, G. Scoutheeten, R. Cancelliere, P. Gallinari, Controlling hallucinations at word level in data-to-text generation, CoRR abs/2102.02810 (2021). URL: https://arxiv.org/abs/2102.02810. arXiv:2102.02810.
[24] C. Hesse, M. Langner, A. Benz, R. Klabunde, Discrepancies Between Database- and Pragmatically Driven NLG: Insights from QUD-Based Annotations, in: D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, B. Heinisch (Eds.), 3rd Conference on Language, Data and Knowledge (LDK 2021), volume 93 of *Open Access Series in Informatics (OASIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021, pp. 32:1–32:9. URL: https://drops.dagstuhl.de/opus/volltexte/2021/14568. doi:10.4230/OASIcs.LDK.2021.32.