Improving Specificity in Review Response Generation with Data-Driven Data Filtering

Tannon Kew **Martin Volk**
Department of Computational Linguistics,
University of Zurich
{kew,volk}@cl.uzh.ch

Abstract

Responding to online customer reviews has become an essential part of successfully managing and growing a business both in e-commerce and the hospitality and tourism sectors. Recently, neural text generation methods intended to assist authors in composing responses have been shown to deliver highly fluent and natural looking texts. However, they also tend to learn a strong, undesirable bias towards generating overly generic, one-size-fits-all outputs to a wide range of inputs. While this often results in 'safe', high-probability responses, there are many practical settings in which greater specificity is preferable. In this work we examine the task of generating more specific responses for online reviews in the hospitality domain by identifying generic responses in the training data, filtering them and fine-tuning the generation model. We experiment with a range of data-driven filtering methods and show through automatic and human evaluation that, despite a 60% reduction in the amount of training data, filtering helps to derive models that are capable of generating more specific, useful responses.

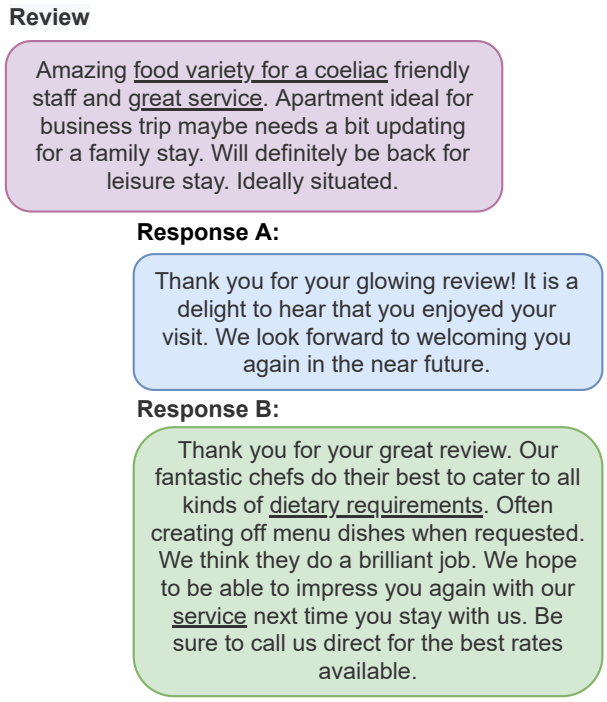

Figure 1: A user-written hotel review with two potentially valid responses. Response A (in blue) is a generic, one-size-fits-all style response, while Response B (in green) addresses and reiterates some, but not all, of the positive points raised in the review.

1 Introduction

Sequence-to-sequence (Seq2Seq) modelling with neural networks has proven to be an extremely popular and effective paradigm for a wide range of conditional text generation tasks (Sutskever et al., 2014; Vinyals and Le, 2015; Nallapati et al., 2016; Lebret et al., 2016, etc.). More recently, the development of large, pre-trained Seq2Seq models (e.g. Lewis et al., 2019) has lowered the bar on the amount of labelled in-domain data required to train models on a particular task and still achieve highly grammatical and fluent text. However, generative models often tend to produce bland and generic text, which significantly inhibits their potential utility (Holtzman et al., 2020). This problem is especially prevalent in tasks with valid many-to-one mappings, where generic outputs occur frequently in the training data; in dialogue modelling it has been referred to as the "I don't know" problem (Khayrallah and Sedoc, 2021).

In this work we consider the task of automatically generating responses to online hospitality reviews. Figure 1 provides an example of the task and presents a user-written hotel review along with two potentially valid responses. While Response B is highly specific, addressing the opening comment of the review and the positive mention of the service, Response A is generic. Such a response would be applicable to a broad range of positive reviews, highlighting the many-to-one problem.

Defining exactly what constitutes a *good* review response is not straightforward. Formal require-

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 121 - 133

ments such as structure, style, intent and grammaticality are all important to consider, however, in this work we focus on content. Popular web-based review platforms (e.g. Google, Tripadvisor, etc.) recommend that responses should address topics raised in the review specifically. Such an approach is also supported by the Gricean maxims of quantity (be informative) and relation (be relevant) (Grice, 1975). Thus, we aim to avoid generating generic responses such as Response A in Figure 1. Yet, given a lack of constraints in response authorship, a significant portion of data that is available from online platforms consists of generic responses which are potentially of little benefit or even detrimental. In order to derive models that are capable of producing more specific, contentful responses, it is essential to mitigate the negative impact of these generic responses in the training data.

A simple yet effective method for improving a model's performance toward a specific goal is to increase the amount of training examples that exhibit the associated target quality and decrease the amount that do not. However, depending on the objective, this can be difficult. For example, classifying an arbitrary piece of text for specificity is challenging since there is limited consensus on what exactly constitutes specificity (Li et al., 2017b). Nevertheless, we investigate this idea and apply unsupervised scoring techniques to hotel review responses that aim to indicate a text's genericness. Given these scores, we infer suitable thresholds and filter out highly generic training data examples. We find that refining the training data and using just 40% of the original training examples allows us to derive models that are capable of producing fewer generic review responses according to both automatic metrics and human evaluation. Our code to reproduce the data used and relevant experiments is available on GitHub.[1]

2 Background and Related Work

Review Response Generation Thanks to an increasing awareness of the benefits associated with addressing online customer feedback (Proserpio and Zervas, 2018; Li et al., 2017a), there is a growing body of literature on automated review response generation. Previous work in this area has considered various domain applications and extended the basic encoder-decoder architecture to incorporate additional contextual information alongside a review text. Zhao et al. (2019) generate responses for product reviews on an e-commerce platform using tabular product information as additional context, while Gao et al. (2019a) focus on generating responses for smartphone app reviews and incorporate discrete external attribute features, such as the review rating and app category. Kew et al. (2020) later applied the same model to restaurant and hotel reviews in English and German and showed that extensive variability in hospitality responses (compared to app review responses) leads to considerably worse performance according to automatic metrics.

Combating Generic Outputs Considerable work has been dedicated to mitigating generic outputs in dialogue models. One popular approach is to feed the model additional contextual information in order to encourage more 'contentful' responses. Depending on the availability of relevant data, this might include the dialogue history (Sordoni et al., 2015), free text from an external knowledge source (Ghazvininejad et al., 2018; Bruyn et al., 2020), or embedded topic signals derived from the input query (Xing et al., 2017). Meanwhile, a number of works have focused on improving the model architecture (Serban et al., 2016a,b; Zhao et al., 2017; Bao et al., 2019; Gao et al., 2019b) or modifying the decoding strategy (Baheti et al., 2018; Li et al., 2016).

Since generic responses occur with high frequency in the dialogue training data they induce a strong, undesirable bias. Thus, it also makes sense to tackle this problem at its source. Previous work in this direction has aimed to remove uninformative training examples in a *conditional* framework by performing comparisons between source and target pairs (Xu et al., 2018; Csáky et al., 2019). In contrast to dialogue data, review-response pairs typically consist of multiple sentences, resembling paragraphs rather than single sentences. This leads to extensive variance in the surface form on both the source and target side, rendering conditional approaches less suitable. For instance, initial investigations revealed that the best-performing approach presented in Csáky et al. (2019) identifies only 5% of hospitality review-responses as generic. Therefore, in contrast to previous works, we set out to identify generic responses independently of their corresponding source texts to improve our training data.

[1] https://github.com/ZurichNLP/specific_hospo_respo

3 Methods

In order to derive models that are capable of producing fewer generic responses, we consider removing them entirely from the training data. Our hypothesis is that generic responses seen often during training encourage the model to learn 'safe' but uninformative responses and are thus detrimental to the model's ability to generate more specific responses. To investigate this, we define three potential methods for scoring a text's genericness within a corpus, operationalising these at the word, sentence and document level. We then derive suitable thresholds for each scoring method and filter training data examples according to their genericness. Formally, given a response text in the training corpus $\mathcal{R} \in \mathcal{T}$, we aim to assign a numerical score $\mathcal{S}$, indicating how unique $\mathcal{R}$ is in relation to all other responses in the corpus.

Lexical Frequency To operationalise our scoring techniques at the word level, we define a response text as a bag of words $\mathcal{R} = \{w_1, w_2, ..., w_m\}$. The frequency distribution of words in natural language corpora tends to follow a long-tailed power law (Zipf, 1935). We exploit this property to easily identify words that occur with such high-frequency that they can be considered to contribute little to no specific information.

Following Wu et al. (2018), a response may then be considered universal if it consists predominantly of words whose rank in the frequency table $\geq n$. Based on this intuition, this scoring method calculates the ratio of high-frequency words to less frequent ones. Specifically for each response text, we compute,

$$\mathcal{S}_{\text{lex\_freq}} = \frac{\sum_{i=1}^{m} \mathcal{I}(w_i)}{m}, \tag{1}$$

where $I(w_i)$ is defined as

$$\mathcal{I}(w_i) = \begin{cases} 1 & \text{if } count(w_i, \mathcal{T}) \geq t \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

In our experiments, we set a frequency threshold $t = 500$ in order to capture a reasonable amount of generic content words (e.g. 'hotel', 'review', etc.) as well as typical stop words.

Sentence Average Considering only the occurrence of unigrams within a response fails to take into account the effect of larger semantic units that may be considered as generic phrases (e.g. personalised greetings, salutations, expressions of gratitude, etc.). Therefore, we also consider a scoring method aimed at quantifying a response's genericness at the sentence level. To operationalise this method, we define a response text as a bag of sentences, $\mathcal{R} = \{s_1, s_2, ..., s_n\}$. Similar to the lexical frequency-based score described above, given a means of reliably identifying generic sentences, we could simply calculate the ratio of generic to non-generic sentences comprising a response text. However, this is less straightforward since sentences do not share the same distributional property.

Works such as Reimers and Gurevych (2019) and Artetxe and Schwenk (2019) demonstrate that deep contextualised sentence representations work well for a wide range of sentence-level semantic textual similarity (STS) tasks. Inspired by these works, we consider scoring a response sentence for genericness by computing its semantic similarity against a pool of generic example sentences $\mathcal{G} = \{g_1, g_2, ..., g_n\}$.

Initial experiments showed that LSTM baseline models exhibit a strong bias towards generating universal responses with little specificity to the themes raised in reviews. We gathered all response sentences generated more than once by an earlier model, considering them as our pool of generic examples $\mathcal{G}$, and compute the maximum similarity for each $s \in \mathcal{R}$ as follows:

$$\xi(s) = \max_{g \in \mathcal{G}}(\cos(s, g)). \tag{3}$$

Then, we compute average sentence-level genericness as

$$\mathcal{S}_{\text{sent\_avg}} = \frac{1}{n} \sum_{i=1}^{n} \xi(s_i). \tag{4}$$

This method constitutes a two-step approach to improve the training data based on outputs from a less performant model and may be seen as similar to the idea behind the iterative 'data distillation' approach presented by Li et al. (2017b). However, unlike that work, which dealt with sentence-level outputs and compared them directly, we compute the final score for a response text by averaging the genericness scores of its constituent sentences.

LM Perplexity In order to score a response text for genericness at the document level, we rely on a causal language model (LM) and compute the perplexity (PPL) of each response. Intuitively, generic responses that occur frequently and with relatively little variation are less surprising and should thus receive a lower LM PPL in contrast to a highly

specific response. Since we do not provide the review text as context, the LM is forced to score the response in isolation, thus maximising surprisal for less generic responses that contain more unexpected events.

To this end, we use a distilled GPT-2 model that is fine-tuned to our domain (Radford et al., 2019; Wolf et al., 2020) and for each training response compute

$$\mathcal{S}_{\text{LM\_PPL}} = \exp(\text{CE}_{\text{LM}}, \mathcal{R}). \quad (5)$$

4 Experiments

4.1 Data Set

Our primary data set comprises a total of 500k unique hotel review-response pairs published on Tripadvisor[2]. We collected data from seven different countries with reviews for more than 7.5k establishments, ranging from luxury hotels to backpacker's hostels and small bed-n-breakfasts. Of the 500k review-response pairs, we take approximately 90% for model training, setting aside 5% for validation purposes and the final 5% for evaluation.

In addition to investigating the proposed techniques on hospitality review responses, we also conduct a small generalisability study on a related data set from a different domain. Specifically, we use the mobile app review responses, originally introduced by (Gao et al., 2019a). These review responses were collected from the Google Play Store and differ considerably in terms of both style and length to those found in the hospitality domain (Kew et al., 2020). Table 1 provides a brief overview of both data sets.

Domain	Split	Rev-resp pairs	Sents
Hospitality	Training	450,367	5.4M
	Validation	24,897	299k
	Test	24,736	297k
Mobile Apps	Training	278,374	1.7M
	Validation	14,602	90k
	Test	15,404	95k

Table 1: Overview of the review response data sets used in our experiments. The hospitality domain refers to pairs collected from TripAdvisor, while the mobile app domain refers to the data set introduced by (Gao et al., 2019a). Numbers indicate the size of the training data before performing targeted filtering.

4.2 Training Data Filtration

After having scored each response text in the corpus with the methods described in Section 3, we inspected the distributions of the resulting scores on the training set. Figure 2 shows the distribution of values for each scoring method. As can be seen, the majority of the distributions follow relatively smooth normal distributions, with various degrees of skew, indicating that different scorers appear to detect different qualities. In order to make all experiment runs comparable, we aim to extract the 'best' 40% of training data examples according to each individual scoring method. To derive appropriate thresholds for data filtering, we inspect samples along the range of x-axis values and align these with the following intuitions:

(i) **Lexical frequency** – a higher ratio of high-frequency words indicates more genericness, thus lower is better;

(ii) **Sentence Average** – a higher score indicates a higher degree of generic sentences within a response, thus lower is better;

(iii) **LM PPL** - a lower PPL indicates less surprisal, while a high PPL potentially indicates a large degree of noise and possibly ungrammatical text, thus a mid-range score is better.

Since we filter the training data according to each scoring method independently, it is reasonable to expect that there may be considerable overlap between the resulting training subsets. Figure 3 shows that most overlap occurs between the word and sentence-level scored subsets with 65%, while the LM PPL filtered subset contains only 57% shared examples.

4.3 Model Training and Inference

Our response generation models are built on top of BART (Lewis et al., 2019), a large pre-trained model for Seq2Seq tasks. All models are initialised with the same BART-base model from Hugging Face (Wolf et al., 2020), which comprises six encoder and six decoder layers. We fine-tuned our models with default hyperparameters and an effective batch size of 40 for a maximum of 8 epochs.[3] The best model from each training run was selected according to ROUGE-2 performance on a 25% sample of the validation set.

[2] https://www.tripadvisor.com/

[3] Depending on the amount of data used, fine-tuning typically runs for two to 5 days on a single 12GB GPU.

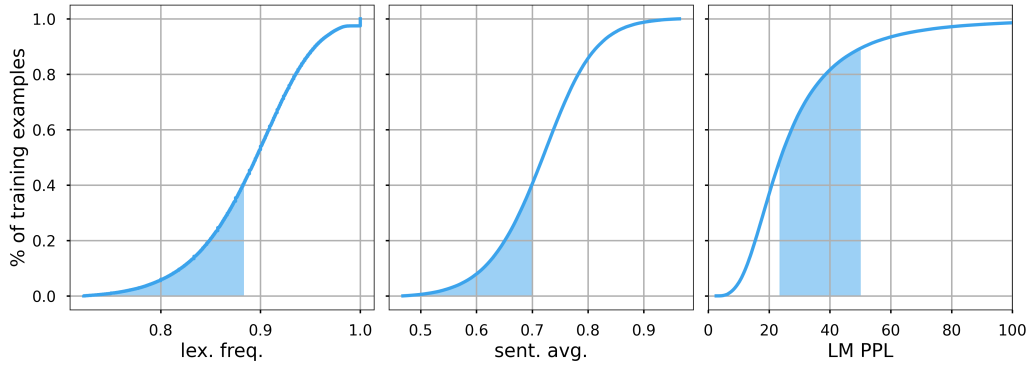

Figure 2: Cumulative distribution plots for each scoring method on the training data. The shaded areas show the 'optimal' 40% of the training data (review-response pairs) identified by the method.

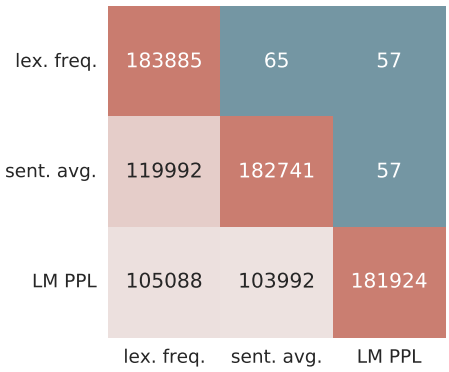

Figure 3: The amount of overlapping examples between each of the three filtered training sets. Numbers in the bottom-left show the raw counts of overlapping target texts, while the numbers in the top-right show the amount of overlap as a rounded percentage.

For all models, inference was performed using standard beam search with k=5 on the full test set, i.e. no filtering is applied to the test set. As a baseline, we use a model fine-tuned on all available training data and compare this to the three experimental systems, each fine-tuned on one of the filtered training sets.

4.4 Evaluation

Evaluating short text-based conversation is inherently difficult since responses are, to a large degree, open-ended. For any given input sequence, the space of potentially valid outputs is extremely large. As a consequence, it is necessary to analyse various characteristics of the generated texts. We employ a selection of automatic metrics that act as approximate but useful indicators of textual quality along multiple axes. In addition, we conduct a human evaluation and compare model outputs in order to measure the effect of different data filtering methods on model performance.

4.4.1 Automatic Metrics

Reference-based Metrics Ground truth responses are unlikely to serve as reliable references for comparison with surface-level or embedding-based automatic metrics due to the open-ended nature of the task. Despite this, and other criticisms (Reiter and Belz, 2009), popular N-gram overlap metrics, such as BLEU have been reported in the relevant literature (Gao et al., 2019a; Zhao et al., 2019). An alternative, easy-to-compute metric is **chrF** (Popović, 2015), which operates on character N-grams rather than full tokens and balances both precision and recall. This makes it considerably more flexible than BLEU, especially for noisy web-based text where spelling errors are common. In addition to reporting chrF against the ground truth responses, we also separately compute chrF using the corresponding input reviews as 'stand-in' references. This provides an approximate measure for specificity in model outputs.

Lexical Diversity and Range Automatically generated review responses should exhibit a decent amount of both inter- and intra-textual diversity. Low inter-textual diversity implies that models repeatedly generate the same or highly similar texts, while low intra-textual diversity indicates that model outputs contain lexical repetitions, possibly as a result of getting stuck in repetitive *de*generate loops (Welleck et al., 2019; Holtzman et al., 2020).

To measure inter-textual diversity, we employ **Self-BLEU** (Zhu et al., 2018). This metric computes for each system-generated output the BLEU score, regarding all other system-generated outputs as makeshift references. Thereby, it effectively measures the amount of textual similarity in terms

	chrF-tgt ↑	chrF-src ↑	DIST-1 ↑	Self-BLEU ↓	Uniq. ↑	Len ↑
Ground truth	-	20.7	75.66	1.18	37649	80.92
Rule-based	19.7	10.1	86.44	55.19	153	35.91
Baseline	30.47	15.87	76.84	24.6	7174	59.39
Lex. freq.	33.6	20.63	74.33	15.37	11859	82.37
Sent. avg.	32.53	20.2	73.46	13.11	11858	75.69
LM PPL	32.63	21.0	74.51	4.24	13366	73.82

Table 2: Model performance under all automatic evaluation metrics considered. Values reported for all BART-based models are averaged over three individual inference runs from models trained with different random seeds to account for potential variation between training runs. Note, metrics reported here are multiplied by 100 where applicable for improved readability.

of N-gram overlap between generated responses. A higher Self-BLEU score indicates less diversity.

For intra-textual diversity, we follow Choi et al. (2020) and use **Distinct-N** (Li et al., 2016). This metric calculates the ratio of unique N-grams to the total number of N-grams generated *within* a text, taking the macro average as the final score. Following Welleck et al. (2019), we also report the total number of **unique** words generated by a model over the entire test set. This provides a simple indicator of a model's lexical range.

Finally, we report the average length of generated texts to provide a rough idea of a model's ability to generate adequate responses under the assumption that shorter responses indicate a greater degree of genericness. Where possible, we also compute these metrics for the human-written ground truth responses in the test set to provide a valuable idea of expected or appropriate values.

4.4.2 Human Evaluation

In order to assess a model's ability to generate fewer generic, one-size-fits-all responses on the basis of training data filtering, we conduct a human evaluation. We sampled 200 reviews from the test set and generated responses with all four models. We then recruited four evaluators, all of whom are familiar with the field of NLP, and asked two evaluators to rate examples 1-100 and the other two evaluators to rate examples 101-200. The evaluation schema was designed to make pairwise comparisons between randomly selected model outputs and asked judges to indicate which response is more specific to the input review. Note that while framing the evaluation question in this way inverts our main aim of reducing generic responses, it simplifies the task for the judges by encouraging them to focus more on the content that *is* generated by the model, rather than what is not. To facilitate decision making and allow for more nuanced judgements, we use a continuous scale that allows evaluators to indicate the degree to which they believe one response is better than the other (Belz and Kow, 2011). Judges were also able to accept or reject both responses if they were equally specific or generic, respectively.

5 Results

Automatic Metrics Table 2 compares model performance under the automatic metrics considered. In addition to the baseline and experimental models discussed above, we also compute automatic metrics for a naïve rule-based baseline, which simply returns a single, hand-crafted response from a small set of candidates based on the rating associated with the input review. These are intended to be highly generic responses that fit the context and thus provide a useful comparison and motivation for more complex approaches.

According to both versions of chrF, all models trained on filtered data sets show considerable gains over the baseline model, trained on the entirety of the data. Specifically, chrF computed against the true target shows smaller improvements over the baseline, while chrF computed against the corresponding source shows a relatively large improvement for all experimental models, bringing the degree of overlap in model outputs much more in line with human-written responses.

DIST-1 shows that intra-textual lexical diversity is consistent against the human-written responses for most models, indicating the lexical repetitions occur within a reasonable range for this relatively restricted domain. Thus, there is no indication that our models are getting stuck in repetitive *de*generate loops (Holtzman et al., 2020).

Self-BLEU reveals considerable variation among all models in terms of the diversity between generated responses. According to this metric, the LM PPL filtering method ensures the most diverse response texts, while both the rule-based and BART baseline generate the least diverse texts by far. Noticeably, the best scoring models are not quite on par with the diversity of human-written responses. However, this is to be expected given that neural models generally stick to generating higher-frequency words (Holtzman et al., 2020). This phenomenon is further indicated by a large discrepancy between the counts of unique lexical items observed.

In terms of average response length, most models under-generate when compared to the human-written ground truth. Here, the baseline models fall the shortest, which may be a useful proxy indicating higher genericness. Meanwhile, the models trained on the lexical frequency-filtered subset show a tendency to generate longer responses. This may be due to the score being directly related to the word count of a text, despite normalisation.

Human Evaluation In analysing the results of the human evaluation, we considered only those examples on which two judges agreed in terms of preference towards a particular response candidate and acceptability. This resulted in 129 valid pairwise comparisons. Following recommendations by Novikova et al. (2018), we derive the overall model rankings using the Bayesian ranking algorithm TrueSkill™ (Herbrich et al., 2007).

According to the results of our human evaluation, all filtering methods help to improve the specificity of model outputs, thereby reducing genericness. Figure 4 depicts the final model rankings derived through applying the TrueSkill™ algorithm to accepted pairwise comparisons from our evaluation. Here, it can clearly be seen that all experimental models outperform the baseline, with a clear tendency towards filtering for genericness based on larger semantic units, i.e. sentence or document-level.

Taken together, the results of our automatic evaluation and the human evaluation strongly suggest that fine-tuning on a filtered subset of data is beneficial, reducing the model's tendency to produce generic responses. In particular, chrF-src and Self-BLEU are useful indicators for gauging relative genericness and diversity of generated texts. Table 5 in Appendix A provides some examples of the responses generated by all models and how they improve in terms of specificity to the input review.

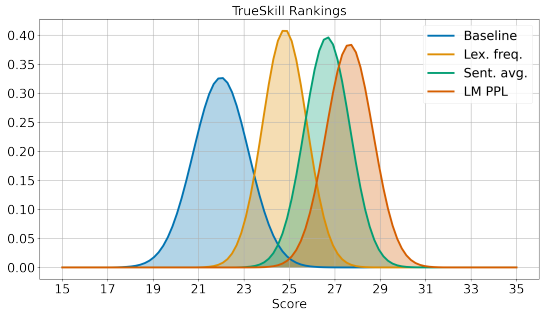

Figure 4: Final model ranking posteriors as computed with TrueSkill™ on 129 human evaluated pairwise comparisons.

5.1 Ablations

How much filtering is too much filtering? In the above experiments we select thresholds for each filtering method to retain approximately only 40% of the original training data. Results of automatic and human evaluations reveal that the LM PPL method performs best on balance. To investigate the ideal amount of filtering, we train and evaluate additional models by incrementing the lower bound of the target text LM PPL filter to train on different quantities of data. Furthermore, we also consider combining all filtering methods together to train a model on only the least generic responses according to the thresholds set in Section 4.2. The results of these ablations are presented in Table 3.[4]

Comparing the results for the LM PPL filter in isolation, we see that the largest performance gains are achieved when training with between 40 and 80% of the total amount of data. According to metrics used as proxies for specificity, chrF-src, Self-BLEU and Uniq., more aggressive filtering (e.g. 40%) works best with very little cost in terms of chrF-tgt. Meanwhile, extremely aggressive filtering (20%), leads to a large performance drop across the board. Interestingly though, combining all filtering methods to filter aggressively has a more positive impact, suggesting that the overall quality of the training data used can indeed be further improved by considering multiple filtering methods. That said, the relatively high Self-BLEU score indicates that this model tends to generate the same response to different input reviews to a greater extent than those trained on more data.

[4]To reduce the computational cost of ablations, we perform a single training run for these models with a fixed random seed.

	chrF-tgt ↑	chrF-src ↑	DIST-1 ↑	Self-BLEU ↓	Uniq. ↑	Len ↑
20%	27.0	15.3	74.39	31.21	7449	50.46
40%	32.7	21.1	74.54	4.49	13548	73.31
60%	32.8	19.6	74.55	8.08	11405	71.73
80%	32.9	18.9	74.68	12.67	8615	72.96
100%	30.5	15.8	76.95	27.1	7273	59.09
ALL 15%	33.3	22.7	72.29	18.19	14374	83.48

Table 3: Results of ablation runs investigating a) performance as a function of the percentage of data filtered using LM PPL and b) performance as a result of combining *all* filtering methods with the thresholds shown in Figure 2 and training on only the 'best' 15% of the data.

	chrF-tgt ↑	chrF-src ↑	DIST-1 ↑	Self-BLEU ↓	Uniq. ↑	Len ↑
20%	26.0	18.0	83.76	0.31	2362	40.12
40%	29.8	17.5	81.98	1.56	2111	45.04
60%	32.3	16.7	80.91	1.47	1978	49.53
80%	33.5	15.8	80.34	2.72	1545	50.65
100%	35.5	15.5	79.04	2.41	1459	53.67

Table 4: LM PPL filtering at varying thresholds for mobile app review response generation (Gao et al., 2019a).

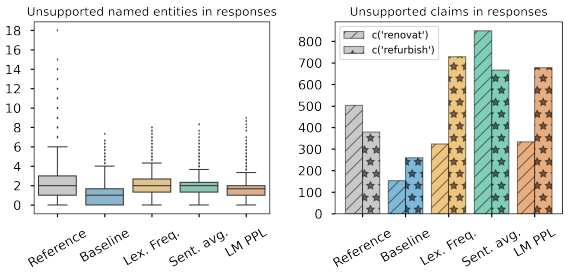

Figure 5: Left: averaged occurrences of named entities in responses that do not appear in the corresponding input review. Right: total occurrences of the stems 'renovat' and 'refurbish' in responses texts.

Generalisability Targeted data filtering is effective for reducing genericness in hospitality review response generation. To investigate whether such an approach generalises to other domains we also consider applying our best performing filtering method to the related task of mobile app review response generation using the data set presented in Gao et al. (2019a). Following the LM PPL approach described in Section 3, we again experiment with a range of thresholds for filtering both low-PPL and overly high-PPL responses from training data. Table 4 shows that targeted filtering in this domain also leads to increased specificity according to automatic metrics used as proxies for measuring genericness.

Are there any side effects? Encouraging a model to consistently produce fewer generic outputs may also have potential side effects. For example, it is possible that this could lead to an increase in hallucinated content that is unsupported by the input and thus may be factually incorrect or misleading. To investigate whether or not the proposed approach compromises the generated outputs in this way, we search for candidate hallucinations in the generated responses and compare their occurrence frequencies to the reference texts and the outputs from the baseline model.

As candidates, we consider named entities, which generally constitute common hallucination errors (Dziri et al., 2021) and mentions of renovations or refurbishments. The latter is observed frequently in hotel review responses as a suitable reply to a criticism about outdated infrastructure or decor. Naturally, the models themselves have no knowledge of whether renovations are planned, so a conservative approach would be to consider all generated responses that mention renovations as hallucinations.

First, we searched review-response pairs for named entities and computed the amount of named entities in the response that do not appear in the corresponding review and are thus 'unsupported'.[5] On the left of Figure 5, we can see that unsupported named entities occur more frequently in the experimental models in contrast to the baseline. Second, we counted occurrences of the stems 'renovat' and 'refurbish' in all response texts. On the right of Figure 5, we can see that all experimental models are guilty of over-producing claims involving renovations or refurbishments and thus could be at risk of generating more factually incorrect claims.

[5]For identifying named entities, we used spaCy (https://spacy.io/).

5.2 Discussion and Future Work

Based on the observations from the previous sections, it is clear filtering uninformative instances from the training data is an effective approach to reduce genericness in model outputs for response generation. However, it does not come without risk. Our analysis revealed that generated responses tend to contain more hallucinated content. Thus, further work is required to mitigate this and better ensure the factual accuracy of generated outputs.

While removing generic training examples is effective at reducing unwanted predictive biases, it does not provide any means to steer the amount of genericness. In certain application scenarios, it may be more desirable to be able to control the degree of genericness at inference time in order to handle difficult or ambiguous cases (Li et al., 2017b). To this end, our methods for quantifying textual genericness, might also be used to derive categorical labels for training examples that are provided to the model in order to be able to steer the generation appropriately at inference time (Filippova, 2020; Martin et al., 2020). We leave a detailed investigation in this direction for future work.

We also acknowledge that there is a considerable risk in deploying fully automatic review response generation in online settings. The societal impacts of computer-generated language are still relatively unknown and thus it is unclear what effects such an application may have on customer satisfaction and business-customer relations in e-commerce and online settings. This work is intended to support response authors in improving their efficiency and extending the capabilities.

6 Conclusion

State-of-the-art approaches to conditional text generation involving transfer learning can be adapted to perform a wide range of domain-specific tasks with strong and convincing results. However, the content and quality of a model's outputs largely reflect that of the in-domain data used for fine-tuning. Thus, care should be taken when deciding which data to use for training. In this paper we presented three unconditional scoring techniques for identifying and filtering generic responses in a parallel corpus of review-response pairs. Results of both automatic and human evaluation revealed that this is an effective approach for helping to reduce the production of generic, one-size-fits-all outputs for review response generation in the hospitality domain, as well as for mobile app reviews. We have also shown that such an approach has potential side effects that must be handled appropriately before being utilised in a real-world scenario.

Acknowledgements

This work was partially supported by the Swiss Innovation Agency InnoSuisse as part of the ReAdvisor project (project number 38943.1 IP-ICT) under the direction of Sarah Ebling. We would like to thank Jannis Vamvas, Janis Goldzycher, Nicolas Spring and Noëmi Aepli for their assistance with conducting evaluations and Chantal Amrhein and Rico Sennrich for their valuable feedback. We also thank the anonymous reviewers for their helpful and constructive comments.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980, Brussels, Belgium. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. PLATO: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Anja Belz and Eric Kow. 2011. Discrete vs. Continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, Oregon, USA.

M. D. Bruyn, E. Lotfi, Jeska Buhmann, and W. Daelemans. 2020. BART for knowledge grounded conversations. In *Converse@KDD*.

Byung-Ju Choi, Jimin Hong, David Keetae Park, and Sang Wan Lee. 2020. Fˆ2-Softmax: Diversifying Neural Text Generation via Frequency Factorized Softmax. *arXiv:2009.09417 [cs]*.

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. *arXiv:2104.08455 [cs]*.

Katja Filippova. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. *arXiv:2010.05873 [cs]*.

Cuiyun Gao, Jichuan Zeng, Xin Xia, David Lo, Michael R. Lyu, and Irwin King. 2019a. Automating App Review Response Generation. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 163–175, San Diego, USA. IEEE.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019b. Jointly Optimizing Diversity and Relevance in Neural Response Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1229–1238, Minneapolis, USA.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5110–5117, New Orleans, USA.

Herbert P Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill(TM): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]*.

Tannon Kew, Michael Amsler, and Sarah Ebling. 2020. Benchmarking Automated Review Response Generation for the Hospitality Domain. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 43–52, Barcelona, Spain. Association for Computational Linguistics.

Huda Khayrallah and João Sedoc. 2021. Measuring the 'I don't know' Problem through the Lens of Gricean Quantity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5659–5670, Online.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*.

Chunyu Li, Geng Cui, and Ling Peng. 2017a. The signaling effect of management response in engaging customers: A study of the hotel industry. *Tourism Management*, 62:42–53.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. *arXiv:1510.03055 [cs]*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017b. Data Distillation for Controlling Specificity in Dialogue Generation. *arXiv:1702.06703 [cs]*.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gùlçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Davide Proserpio and Giorgos Zervas. 2018. Study: Replying to Customer Reviews Results in Better Ratings. *Harvard Business Review*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783, Phoenix, Arizona. AAAI Press.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016b. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *arXiv:1605.06069 [cs]*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. *arXiv:1506.06714 [cs]*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs]*.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv:1506.05869 [cs]*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural Text Generation with Unlikelihood Training. *arXiv:1908.04319 [cs, stat]*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*.

Bowen Wu, Nan Jiang, Zhifeng Gao, Suke Li, Wenge Rong, and Baoxun Wang. 2018. Why do neural response generation models prefer universal replies? *CoRR*, abs/1808.09187.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, USA.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Lujun Zhao, Kaisong Song, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2019. Review Response Generation in E-Commerce Platforms with External Product Information. In *The World Wide Web Conference on - WWW '19*, pages 2425–2435, San Francisco, CA, USA. ACM Press.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100, Ann Arbor MI USA. ACM.

George Kingsley Zipf. 1935. The psycho-biology of language. In *Indogermanische Forschungen*.

A Example Outputs

	Review rating: 4 stars
Review	**Basic clean & lovely** We stayed here whilst doing a road trip between the snowy mountains and Canberra. After doing the caves and coming through Talbingo (beautiful countryside) we were deciding whether to stay in Tumut or Yass or Gundagai. Gundagai had more choices avail. We chose **** due to $ location. There are 3 tiers of accom. Budget $89 basic. Upstairs w a bed shower and a v v tiny tv. Middle +$30 for ground level bigger room and Spa room for $more again. We chose basic as it was 1 night. It was clean. Shower pressure good and comfortable bed. All u need. Wifi didn't happen. Reception was fine. If u don't need large tv and wifi go basic and stay. If u want/need more this wouldn't be the place for you.
Ground truth	Thank you **** for your comments. Our upstairs rooms are small compact rooms and are well priced for the size of room and location (1/2 way between Sydney & Melbourne) downstairs rooms are much more roomy with large flatscreen TV's, parking at the door, spacious bathrooms and of a deluxe standard - worth the extra price for more comfort.
Baseline	Thank you for taking the time to write a review. We are glad you enjoyed your stay with us.
Lex. freq.	Thank you for taking the time to write a review. We are glad that you enjoyed your stay with us. We do offer a range of accommodation options to suit all budgets. Our budget rooms are priced at $89 per night for a standard room and $89 for a superior room. We also offer a variety of room types to suit different needs. We hope to see you again in the future.
Sent. avg.	Thank you for taking the time to write a review of your stay with us at ****. We are sorry to hear that you did not enjoy your stay and that you found the TV to be small. We do have a number of larger TV's in the rooms and we are currently in the process of upgrading them. We hope to see you again in Gundagai.
LM PPL	Thank you for taking the time to write a review. We are glad that you enjoyed your stay with us and found the rooms clean, comfortable and clean. We do offer a range of accommodation at **** and we are sorry that you were not able to experience this during your stay. We hope you enjoyed the rest of your road trip.

	Review rating: 3 stars
Review	**I frequent this hotel for business** Rooms are nice. Great location. Staff should be drug tested. They have some employees that wreak of marijuana... probably will take my business elsewhere. Bye bye. Just say no! This was not an isolated incident.
Ground truth	Thank you for taking the time to comment and I will look into your comments. I am sorry you may take your business elsewhere.
Baseline	Dear Valued Guest, Thank you for taking the time out of your schedule to complete our Guest Survey regarding your recent stay at our hotel. Please know that your feedback is extremely valuable to us and we will utilize your response by continuing to focus on making each and every guest experience the best it can be. We hope to be of service to you in the future. ...
Lex. freq.	Thank you for taking the time to complete our guest survey about your recent experience at ****. Your feedback is critical to our continued improvement and success! We hope you will come back and stay with us again soon. Please contact me personally to arrange your next visit so I can ensure your next stay is nothing less than exceptional. ...
Sent. avg.	Dear Valued Guest, Thank you for taking the time to write a review about your recent stay at ****. I am sorry to hear that your stay was not a pleasant one. We take the cleanliness and safety of our guests very seriously. I have shared your comments with our staff to ensure that this does not happen again in the future. Again, thank you for your review and we hope to see you in the near future. ...
LM PPL	Dear Valued Guest, Thank you for taking the time to provide us with your feedback from your recent stay at ****, a very good place to be. We are happy to hear that you found our rooms to be nice and our location to be great. However, we are very sorry to hear about the incident that occurred during your stay with us. We take the safety and security of all of our guests very seriously and we will be following up with the appropriate departments to ensure this does not happen again. We hope to have the opportunity to welcome you back in the future. ...

Table 5: Example model outputs for reviews of various user ratings (continued on next page). Coloured passages in the responses indicate content that addresses aspects of the input review specifically. We also italicise potentially hallucinated content, which is unsupported by the source text. Note, bold text prefixing a review indicates the review's title. Personalised greetings and salutations are truncated from the response texts, as indicated with ellipsis. Similarly, identifiable information such as personal names and establishment names are anonymised with ****.

	Review rating: 2 stars
Review	**Good for location only. Disappointing overall** Location is great but room faced trash bins and was loud. Also, room was very tiny and looked dated. Windows were tough to close and no air circulation other than the portable fan. Bathroom was hard to even walk in and out of.
Ground truth	We appreciate you taking the time to offer us feedback on your stay. Please allow me to apologise that you did not enjoy your experience on this occasion. ...
Baseline	Thank you for taking the time to write a review. We are sorry to hear that you did not enjoy your stay with us. We hope that you will consider staying with us again in the future.
Lex. freq.	Thank you for taking the time to provide us with a review based on your last experience with us. I would like to apologize if you felt as though we had fallen short of your expectations - please allow me to express my sincerest apologies. We are in the midst of a multi-year renovation project that will touch every corner of our hotel - we hope to warmly welcome you back soon for a stay that is uniquely memorable for all the right reasons.
Sent. avg.	Thank you for taking the time to write a review about your recent stay with us. I am sorry to hear that you did not enjoy your stay. We are in the process of renovating our rooms and hope to have this completed by the end of the year. I hope that you will give us another opportunity to show you that your unsatisfactory stay was an exception. If I can be of any further assistance, please do not hesitate to contact me directly.
LM PPL	Thank you for taking the time to provide a review based on your last experience with us. We would like to apologize for falling short of not only meeting your expectations, but ours as well. It is only through feedback such as yours that we are able to maintain and indeed, where necessary, improve upon the service we provide to our valued guests. We hope to warmly welcome you back soon for a stay that you are most contented with.

	Review rating: 1 star
Review	**False advertising?** Due to the favorable reviews on the Internet we have booked family room (2 rooms with door in between). When we walked in after a full day of driving we were tempted to leave the rooms and continue the trip, as everything but location was disappointing. Rooms were dusty (particularly carpet) and not refurbished for more than 30 years. Beds were short (not king size), uncomfortable and with cheap pillow. Almost every piece of furniture was chipped and broken, bathroom doors couldn't be closed properly, Shower head was falling down. TV reception was terrible. Acoustic insulation was inappropriate. Photo on the Internet is not the proper representation of the motel; I almost took photos to put it on the Internet to show how our rooms looked like. Do not be tempted to come to this nominally 3 star motel. I personally wouldn't give it a single star.
Ground truth	We appreciate all feed back from customers good and bad, and thankyou for taking the time to review. This is the 1st time we have had such a review and feel just a few points should be made. We do not cater for Familys , groups or people under 40 years old, thus the Golden Chain logo, Senior logo and the fact there is nowhere to add children to any internet booking. When presented with a Family as occainionally does happen we open 2 large double rooms with 2 bathrooms 2 lounge rooms and 2 balconies with 100% waterview over ****. and full breakfast for 5 delived to your room for $129 per night. All but these two rooms are Theme rooms completed early 2010 with artworks and decoration valued at $20,000 per room. The risk of damage is too high to have children at the motel. Not all guests like our contingency plan however I beleive it is fair. That said we will look at a standard renovation for these rooms.
Baseline	Thank you for taking the time to write a review. We are sorry to hear that you did not enjoy your stay with us. We have taken all your comments on board and are *currently in the process of refurbishing our rooms*. We hope that you will consider staying with us again in the future.
Lex. freq.	Thank you for taking the time to write a review. We are sorry to hear that you did not enjoy your stay with us. *We do not advertise that we are a 3 star motel*, but we do have a very high standard of cleanliness and customer service. Our rooms are cleaned on a regular basis, and we have never had any issues with any of our carpets. All of our *rooms have been refurbished over the past 3 years*, and our rooms are regularly inspected by our housekeeping team. We have also *recently replaced all of our beds, and are currently in the process of replacing all of the furniture in our rooms*...
Sent. avg.	Thank you for taking the time to write a review. We are sorry to hear that you did not enjoy your stay with us. We do have a wide range of room types to suit all budgets and requirements. We have a variety of different room types available to suit different needs and budgets. *All of our rooms have been refurbished over the last 2 years* and we are sorry that you found the room to be dusty. We would have been more than happy to move you to a different room had we been made aware of your concerns...
LM PPL	Thank you for taking the time to write your review. We are sorry to hear that you did not enjoy your stay with us. Our rooms are cleaned to a very high standard and we are sorry that you found the carpet dusty and uncomfortable. *The carpets in the rooms have been replaced over the last couple of years* and *we have not received any complaints about the carpet in the room you stayed in*. We *have not had any complaints regarding the TV reception or the shower head*. We would have been more than happy to move you to another room if you had let us know of your dissatisfaction.