Benchmarking Automated Review Response Generation for the Hospitality Domain

Tannon Kew, Michael Amsler, Sarah Ebling
Department of Computational Linguistics, University of Zurich
{kew, mamsler, ebling}@cl.uzh.ch

Abstract

Online customer reviews are of growing importance for many businesses in the hospitality industry, particularly restaurants and hotels. Managerial responses to such reviews provide businesses with the opportunity to influence the public discourse and to attain improved ratings over time. However, responding to each and every review is a time-consuming endeavour. Therefore, we investigate automatic generation of review responses in the hospitality domain for two languages, English and German.

We apply an existing system, originally proposed for review response generation for smartphone apps. This approach employs an extended neural network sequence-to-sequence architecture and performs well in the original domain. However, as shown through our experiments, when applied to a new domain, such as hospitality, performance drops considerably. Therefore, we analyse potential causes for the differences in performance and provide evidence to suggest that review response generation in the hospitality domain is a more challenging task and thus requires further study and additional domain adaptation techniques.

1 Introduction

Online customer reviews play a significant role in e-commerce and specifically in the tourism and hospitality industry. Websites such as TripAdvisor, Booking.com and Yelp offer customers the opportunity to share their experiences in the form of reviews relating to restaurants, cafés, hotels and other business types. To date, TripAdvisor boasts more than 860 million customer reviews and opinions worldwide[1].

While these reviews serve as an important point of reference for potential customers who value and rely on electronic word-of-mouth recommendations (Litvin et al., 2008), it has been shown that they also offer businesses an opportunity to influence the public discourse and foster positive customer relations by responding to these reviews appropriately (see Li et al. (2017); Li et al. (2018)). However, given the sheer amount of online reviews, composing these responses is a time-consuming and costly endeavour for any business.

In this paper, we report on automated response generation for hospitality reviews, specifically restaurants and hotels, in both English and German. While a similar task has previously been tackled for product reviews in online stores in Chinese (Zhao et al., 2019) and for smartphone app reviews in English (Gao et al., 2019), the application domain of hospitality reviews is, to the best of our knowledge, novel. In addition, our investigations are applied to two languages, which allows for a valuable idea of how well the approach generalises across languages.

The experiments conducted in this paper leverage the approach proposed by Gao et al. (2019), who extend the basic attentional sequence-to-sequence (seq2seq) neural network (Sutskever et al., 2014; Bahdanau et al., 2015) framework for conditioned text generation. We find that applying the proposed solu-

[1] http://ir.tripadvisor.com/static-files/f81e3d8c-2e18-409f-9ca3-8b3148238257 (last accessed: July 16, 2020)

Proceedings of the Workshop on Natural Language Processing in E-Commerce (EComNLP), pages 43–52
Barcelona, Spain (Online), Dec 12, 2020.

tion to hospitality review-response pairs yields a considerable decrease in overall performance compared to the results reported in the original paper in the context of app reviews. An empirical investigation into possible causes for the discrepancy provides evidence to suggest that the task is considerably more challenging in our target domain. Thus, we establish a preliminary baseline for future work on automatic response generation for hospitality reviews.

2 Seq2Seq Review Response Generation

Following previous work (e.g., Gao et al. (2019); Zhao et al. (2019)), we tackle the task of review response generation using the popular seq2seq encoder-decoder framework originally proposed by Sutskever et al. (2014) in the context of neural machine translation (NMT) and successfully applied in a broad range of NLP tasks, including abstractive text summarisation (Rush et al., 2015) and conversational dialogue systems (Vinyals and Le, 2015), among others. The central idea behind the seq2seq framework is as follows: Given an input (source) sequence of length n, $X = \{x_1, \ldots, x_n\}$, and a corresponding output (target) sequence of length m, $Y = \{y_1, \ldots, y_m\}$, we aim to learn a probabilistic mapping function $g(\cdot)$ that allows us to predict a probable target sequence $\hat{Y}$ for a novel source sequence.

2.1 Encoder-Decoder Network

The seq2seq model architecture proposed by Sutskever et al. (2014) comprises two recurrent neural networks (RNNs) which are jointly optimised during training. The first of these reads in the source sequence, *encoding* it into a fixed-length context vector $c \in \mathbb{R}$. The second neural network takes the resulting context vector c as input and effectively *decodes* it, outputting a sequence of target tokens over a number of timesteps $t \in T$ and stopping as soon as a special end-of-sequence token (<EOS>) is produced. Thus, the probability of a target token at timestep t is modelled as

$$\log p(y_t | y_{<t}, c) = g(h'_t, y_{t-1}, c).$$

A major limitation of the original encoder-decoder architecture is the fact that a given source sequence of arbitrary length is represented by the fixed-length context vector c. As a result, the longer the source sequence, the more difficult it becomes to capture and leverage important information from early on in the sequence when generating the target sequence. The attention mechanism, introduced by Bahdanau et al. (2015), helps to alleviate this.

2.2 Attention Mechanism

The attention mechanism provides the decoder access to the entire sequence of encoder hidden states $H = \{h_1, \ldots, h_n\}$ computed on the input sequence rather than only the final hidden state (Bahdanau et al., 2015; Luong et al., 2015). At each decoding timestep t, we dynamically calculate a new context vector c_t that considers the relative importance of elements in the source sequence for generating the current output y_t. This essentially allows the decoder to focus on particular parts of the input sequence during decoding and alleviates the information bottleneck associated with encoding the entire source sequence into a static fixed-length context vector.

2.3 Review Response Generation

While the attentional seq2seq approach described above has been proven to work well for tasks like NMT, it poses several challenges in the context of more free-form text generation. For example, in the context of hospitality reviews, context is typically very limited and responses can range from very generic, one-size-fits-all responses, e.g.,

> "thank you for your kind review . we are glad you enjoyed your visit and hope to see you again very soon !", (1)

to far more personalised responses that address specific topics raised in the review and also include aspects of external factual knowledge that may be relevant for a given response. The response below

provides such an example, where information relevant to the input review is set in bold and information that involves external 'world' knowledge is underlined.

> "thanks for your review . we are very happy to hear that you decided to enjoy the **famous cheese fondue and a classic sausage** at our restaurant and , of course , that **you liked it so much** . we 've been serving traditional swiss cuisine since <DIGIT> and are always happy to share our passion for it with our guest . **<NAME> and the whole team were very glad to receive such good feedback** and we are all looking forward to welcoming you back soon ." (2)

Therefore, instead of conditioning solely on the input text, as is done in standard NMT, we ideally want to be able to incorporate additional features that provide relevant contextual knowledge for the response text. To this end, Zhao et al. (2019) exploit a double-encoder-decoder network in which, in addition to the typical source text encoder, a second encoder reads in product-related information from a structured table. At inference time, the decoder has access to both contextual representation vectors by way of a gated multi-source attention mechanism (Arevalo et al., 2017). However, this approach requires appropriate factual tables to be generated for all relevant entities (in our case, restaurants and hotels listed on TripAdvisor), which we do not have access to.

The approach proposed by Gao et al. (2019), on the other hand, relies primarily on information that can be derived from the review text itself or additional metadata that is readily available. The authors identify four main attributes which are provided as additional input to the decoder (dubbed the A-component): (i) the length of the review text, (ii) the review rating, (iii) a sentiment score calculated for the entire review text, and (iv) the category of the app which is the focus of the review.

The motivation for including these attributes is given by Gao et al. (2019) as follows: review length is an important indicator for the length of an output response since it should be appropriate given the length and detail of the review text; the review rating provides valuable information that impacts the response style directly, e.g., expressing an apology given a negative review or expressing thanks given positive feedback (where the lower half of the numeric rating scale is interpreted as corresponding to negative feedback and the upper half to positive feedback, possibly with an ambiguous midpoint in case of odd-numbered scales); the sentiment score attempts to capture the attitude of the reviewer and accounts for instances where the review rating and review text are inconsistent with each other with respect to polarity; and finally, the category attribute provides general contextual information to the decoder since app reviews of different categories typically address different topics.

Additionally, in order to encourage the decoder to produce responses relevant for the input review, the authors exploit an external keyword dictionary developed by Di Sorbo et al. (2016) for the purpose of app review classification. This keyword dictionary (dubbed the K-component) identifies topical words and maps them to one of twelve aspects relevant for app reviews (e.g., 'version', 'GUI', 'pricing', etc.).

The approach proposed by Gao et al. (2019) is relatively straightforward to adapt to our task of review response generation for the hospitality domain in English and German. For this reason, we select it as a basis for our experiments. More details about the implementation of Gao et al. (2019) are given in Section 3.2.

3 Review Response Generation for the Hospitality Domain

3.1 Data

For the purpose of our experiments, we compiled two datasets of review-response pairs for hotels and restaurants published on TripAdvisor, roughly the same size as that of Gao et al.'s app dataset. We collected data in both English and German from nine different countries[2] and performed simple preprocessing. Specifically, we used `spaCy`[3] to mask personal names, toponyms, emails, urls and numbers. For English, texts were converted to lowercase, while for German, nouns and proper nouns, which are

[2] Australia, Canada, Ireland, New Zealand, United Kingdom, United States (for English); Austria, Germany, Switzerland (for German).

[3] `https://spacy.io/` (last accessed: July 16, 2020)

	Hosp. (en)	Hosp. (de)	Apps
train	320k	259k	280k
valid	40k	32k	14k
test	40k	32k	15k

Table 1: Overview of review-response pair datasets. For comparison, the size of the app review-response dataset by Gao et al. (2019) is also given on the right.

capitalised in standard German, were titlecased and all remaining words lowercased. We also removed duplicate review-response pairs from the dataset before randomly splitting it into training, validation and test sets at a ratio of 80:10:10. Table 1 provides an overview of the datasets.

3.2 Method

The proposed approach incorporates additional features into the encoder-decoder network, namely the A-component and the K-component, introduced in Section 2.3. The implementation by Gao et al. (2019) embeds each of the additional attributes in the A-component into 90-dimensional vectors which are then concatenated with the hidden representation for the source sequence before being passed to the decoder. The K-component is incorporated by concatenating a 20-dimensional vector, indicating whether or not a word represents a certain aspect, with the pre-trained word embeddings. For the most part, we follow the original implementation and use the same hyper-parameters[4]. To ensure comparability, we evaluate the performance of each model after training for three epochs. In this section we describe the few aspects in which our implementation differs from that of Gao et al.'s, namely the choice of pre-trained word embeddings, how we compute the external sentiment analysis attribute features, and how we derive keyword dictionaries for our data set of hospitality review-response pairs in German and English.

Pre-trained Word Embeddings In contrast to the original approach, which uses `GloVe` pre-trained word embeddings (Pennington et al., 2014), we opt for `fastText` pre-trained word embeddings (Grave et al., 2018). The motivation for this is two-fold. Firstly, `fastText` embeddings incorporate subword-level information in the form of character N-gram features. This allows us to derive reasonably well-estimated word embeddings even for out-of-vocabulary (OOV) words, which is crucial given that our corpus consists of user-generated web content, where typos and alternative spellings are frequent. For example, Table 2 shows some examples of commonly found misspellings in our dataset that are largely OOV and thus fail to receive an informative embedding using the pre-trained `GloVe` model. Secondly, pre-trained `fastText` embeddings are readily available for 157 languages[5], including English and German, whereas pre-trained `GloVe` embeddings are only available for English.

Sentiment Analysis As mentioned in Section 2.3, an overall sentiment score is provided to the decoder as part of the A-component. In the original implementation by Gao et al. (2019), the authors make use of SentiStrength (Thelwall et al., 2010) to automatically calculate the polarity of a given review. While SentiStrength is available for both English and German, and has been shown to successfully predict positive or negative sentiment for short social media web texts, where spelling mistakes are frequent, it is primarily only supported on machines running Microsoft Windows. Therefore, we resort to an alternative sentiment analysis tool. VADER (Valence Aware Dictionary for sEntiment Reasoning) (Hutto and Gilbert, 2014) employs a human-validated sentiment lexicon mapping lexical features (e.g., words) to a valence score, indicating its sentiment polarity and intensity. This lexicon is then combined with five general rules that consider grammatical and syntactic features which influence sentiment (e.g., use of punctuation, all-caps and degree modifiers such as 'very'). VADER has been shown to perform well

[4]Word embedding size = 100, number of hidden layers = 1, hidden layer size = 200, dropout rate = 0.1 and batch size = 32. Optimisation is performed with Adam (Kingma and Ba, 2017) with initial learning rate = 0.001 and weight decay = 0.00001. Since our texts are more varied (see Section 3.4), we limit the source and target vocabulary size to the 20,000 most frequent words, which is twice the size of the original implementation.

[5]`https://fasttext.cc/docs/en/crawl-vectors.html` (last accessed: July 16, 2020)

Canonical form	Common misspellings
restaurant	resturant (707), restuarant (362), restaraunt (163), restraunt (146), resteraunt (124), restaurent (50), resturaunt (42), restauraunt (28), restauarant (28), restarant (23), ...
accommodation	accomodation (1,130), accomadation (67), acommodation (12), accomdation (12), accommadation (9), accomidation (7), ...

Table 2: Common misspellings found in our corpus. Note, inspecting the two pre-trained embedding models shows that the most common misspellings 'restuarant' and 'accomodation' are indeed represented in the `GloVe` model, however, both words receive rather low similarity scores to their canonical forms (0.13 and 0.61, respectively). With the pre-trained `fastText` model, all variants receive similarity scores above 0.33.

across a broad range of domains, and particularly on social media texts (see Hutto and Gilbert (2014)).

While VADER was primarily developed for English, it is extensible and adaptable to other languages since it relies on a lexicon-based approach extended with simple heuristics. Tymann et al. (2019) propose GerVADER, which replaces the standard VADER lexicon with a slightly extended version of the German sentiment lexicon SentiWS (Remus et al., 2010).

It should be noted that GerVADER is shown to perform reasonably well on the classification of positive sentiment texts but performs poorly on negative sentiment texts due to the fact that negations in German often occur after the verb which they modify (e.g., *Ich gehe **niemals** zurück!* ('I go never back!')) as opposed to English, where they usually appear before (e.g., 'I never go back!'). The current version of GerVADER fails to account for such syntactic differences between the two languages and is thus suboptimal. Nevertheless, we make use of this off-the-shelf tool due to its availability and ease of application in the target domain.

Review Categories Gao et al. (2019) make use of the category attribute feature to distinguish reviews according to the genre of the relevant app. Since we do not have such fine-grained categories in our dataset, we simply distinguish between whether the review pertains to a hotel or a restaurant. We note, however, that a more fine-grained distinction regarding the nature of an establishment (e.g., 'take-away', 'fine-dining', 'coffee shop', etc. for restaurants) could be more beneficial for ensuring that the model learns relevant patterns from review-response examples.

Keyword Dictionary for Restaurant and Hotel Reviews In order to derive suitable keyword dictionaries for the target domain, we leverage a manually annotated corpus of English and German hospitality reviews in which 6,490 text spans have been categorised with a thematic aspect. For example, the span 'the waiter was very attentive' is categorised as *Service*. We aggregate all text spans for each category and filter for content words (i.e., nouns and adjectives) that are (a) common for the particular category and (b) uncommon for the other categories. Lastly, we manually filter out words for which there is no clear semantic relationship with the category *per se*.

Table 3 shows the different thematic aspects and the number of text spans associated with each. As can be seen, there is a large skew in the distribution over the categories such that the number of spans available ranges from 234 to 1,458 (German) and from 221 to 1,179 (English), which is also reflected in the resulting keyword dictionaries, whose sizes range from 28 to 120 and 18 to 173, respectively. This is, on the one hand, rooted in the distribution of the dataset, i.e., the number of text spans per category results from the overall frequencies in a small subset of reviews. On the other hand, the variety in the descriptions of the attributes per category is also different. We find a much richer vocabulary for descriptions of food, its preparation, and quality in comparison to content that refers to the pricing (cf. Table 4). This is in line with the distributions we found in the resources that were created by Di Sorbo et al. (2016) and adapted by Gao et al. (2019). Additionally, we keep the size of the dictionaries

Category	Text Spans de	Text Spans en	Keywords de	Keywords en
Ambiance	519	320	33	34
Facilities	234	221	28	81
Food	1,458	1,179	120	173
Service	1,085	763	98	36
Value	374	337	91	18
Total	3,670	2,820	370	324

Table 3: Overview of annotated text spans for each category in restaurant reviews, and the number of derived keywords for each dictionary.

Category	Keywords
Food	fondue, rosti, spaghetti, risotto, clams, pizzas ..., cappuccino, cocktail, wine, whisky, ..., coffees, pastries, sauce, ..., meat, fish, ..., taste, flavour, ..., veggie, vegan, gluten, ..., breakfast, brunch, lunch, dinner, ..., starters, appetiser, entrées, desserts, ... offers, option, combination, specialties, ... plate, courses, dishes, ..., allergies, ingredients, ...
Value	affordability, money, prices, price, charge, budget, value, fortune, cost, penny, pricing, price/quality, costs, fee, expense, price-performance, competitiveprices

Table 4: Example for keyword lists for the two categories *Food* and *Value*

approximately in the same range. Although it would be possible to expand these keyword lists further with corpus-based methods, we opt to compile them in such a way that the approach is comparable to previous work.

3.3 Evaluation and Results

We evaluate the performance of the proposed approach using automatic metrics which typically compare a system output hypothesis to its corresponding ground-truth reference. We report results using the following metrics:

BLEU (Papineni et al., 2001), a precision-based metric popular in machine translation evaluation, is commonly used for evaluating performance of response and dialogue generation systems (e.g., Zhao et al. (2019), Gao et al. (2019), Ghazvininejad et al. (2018), Xu et al. (2017), Li et al. (2016), Sordoni et al. (2015), among others) since it is claimed to correlate well with human judgements. BLEU-N calculates the number of N-gram matches between a hypothesis and one or more references and is commonly reported as the weighted average score for values of N ranging from $[1, 4]$.

ROUGE (Lin, 2004) is, in contrast to BLEU, a recall-oriented metric and is often applied in the evaluation of text summarisation systems. There are two main variants of ROUGE, namely ROUGE-N and ROUGE-L. The first of these considers the number of N-grams in a reference that are matched in the corresponding hypothesis. Typical values for N are 1 and 2. ROUGE-L, on the other hand, does not presuppose a defined sequence length, but simply computes the longest common sub-sequence between a set of references and a hypothesis text.[6]

Distinct-N was proposed by Li et al. (2016) in order to measure textual diversity in neural dialogue system output. Calculated over the whole test corpus, it measures the ratio of unique N-grams to

[6]We use the implementation provided at `https://pypi.org/project/rouge-score/` (last accessed: July 16, 2020).

Dataset	Model	BLEU	RGE-1	RGE-2	RGE-L	DIST-1	DIST-2	BERT-F1
Hosp. (en)	s2s (baseline)	**8.17**	**35.62**	**14.55**	**28.94**	0.00	0.01	**4.82**
	s2s +A +K	2.92	24.24	9.65	20.34	0.00	0.01	-13.07
Hosp. (de)	s2s (baseline)	**10.19**	**34.26**	**15.07**	27.14	0.07	0.13	**16.81**
	s2s +A +K	8.22	32.98	14.99	**27.39**	**0.15**	**0.27**	14.86
Apps (orig.)	s2s (baseline)	14.22	32.20	13.05	24.16	0.06	0.28	-4.76
	s2s +A +K	**29.36**	**44.33**	**27.29**	**37.69**	**0.09**	**0.62**	**14.18**
Apps (uniq.)	s2s (baseline)	15.22	33.82	13.33	25.74	0.05	0.27	-0.36
	s2s +A +K	**24.84**	**40.34**	**22.59**	**33.90**	**0.06**	**0.37**	**7.17**

Table 5: Results of automatic evaluations for hospitality (hosp.) review response generation. 's2s' stands for the basic attentional seq2seq model with no additional attributes, while 's2s +A +K' stands for the seq2seq model extended with additional attribute and keyword components, as proposed by Gao et al. (2019).

the total number of N-grams generated, thus providing an indication of how varied the generated responses are.

BERTScore was recently proposed by Zhang et al. (2020) for natural language generation evaluation. It utilises contextualised word embeddings from a pre-trained BERT language model (Devlin et al., 2019)[7] and calculates the pairwise cosine similarity score between tokens in a hypothesis and tokens in a set of references. Furthermore, BERTScore incorporates importance weighting by applying the inverse document frequency (idf) score to each reference word, which is calculated on the test corpus references. The major advantage of this metric is that it alleviates the constraint of surface-form and structural similarity imposed by N-gram matching metrics, and in addition, it is applicable in multiple settings since it relies solely on a task-agnostic pre-trained BERT model.

Table 5 shows the results of our experiments in which we apply the extended seq2seq model proposed by Gao et al. (2019) to the domain of hospitality reviews in both English and German. Since we aim to inspect potential performance improvements gained by incorporating the additional feature attributes described in Section 2.3, we use the basic attentional seq2seq model as our baseline. As can be seen, the extended architecture fails to outperform the baseline according to almost all metrics in the hospitality domain. The starkest difference occurs with English, where the baseline outperforms the extended approach by considerable margins across the board. The story is slightly different in the case of German, where the margins are much smaller, but still, the superiority of the baseline seq2seq model is apparent.

The low scores for both Distinct-N metrics, particularly for English hospitality responses suggest extensive N-gram repetition. Manual inspection of the system outputs confirms that the current seq2seq models tend to generate overly generic and uninformative responses such as in Example 1, as well as showing other signs of text *degeneration* (Holtzman et al., 2020). For this reason, we avoid conducting human evaluation on these outputs and aim to address this issue in future work.

As a point of reference, we also reproduce the experiment by Gao et al. (2019) on their original app review-response dataset and on a deduplicated version which removes overlap between train and test splits (shown as 'uniq.' in Table 5)[8]. Here, we see a convincing boost in performance over the vanilla seq2seq model (15.14 BLEU points), demonstrating that the additional features do indeed help for app review response generation.

[7]For English, BERTScore uses a 24-layer RoBERTa Large model, while for German, a 12-layer multilingual BERT model is used to compute the contextualised embedding representations.

[8]Note, the deduplicated app dataset ensures that there is no information leakage between training, test and validation splits. As expected, there is a slight drop in performance for all metrics. However, the benefits of the extended seq2seq archticure are still clearly visible over the baseline.

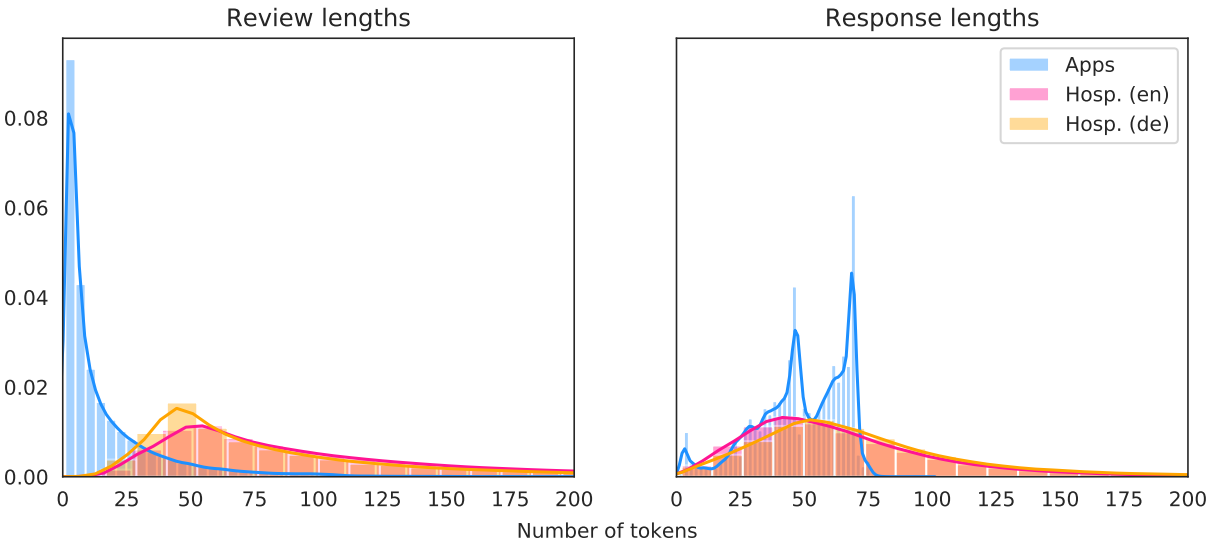

Figure 1: Distribution of review and response text lengths in the app dataset compared to the English and German hospitality datasets.

The large drop in performance of 20+ BLEU points when applying the proposed approach to a different domain and also another language is conspicuous. Therefore, in the following section, we discuss the nature of the different datasets and conduct an empirical investigation to determine potential causes.

3.4 Discussion

Given the results presented in Table 5, it is clear that the extended seq2seq architecture has difficulty generating suitable review responses when applied to the target domain. In order to investigate why this is the case, we take a closer look at the two different types of datasets and identify two potential causes, namely review length and textual variance in the responses.

Review Length The first noticeable difference between the review-response data for apps and that for the hospitality domain is the length of the source and target texts. Figure 1 shows the distribution of text lengths for all the three datasets. As can be seen, reviews for apps (blue on left) are generally much shorter in length than reviews for hotels and restaurants, with an average length of 16 tokens for the former and approximately 100 tokens for the latter. Furthermore, the distribution of app review response lengths (blue on right) is also far from a normal distribution, with multiple peaks corresponding to response text lengths of approximately 3, 45 and 70 tokens. This is largely due to a high number of frequently occurring template responses (e.g., 'thank you !').

In theory, an attentional encoder-decoder network should be able to comfortably handle source texts of 200 tokens in length. However, we conjecture that the brevity of app reviews contributes to the simplicity of generating these responses as the input signal is far less complex in many cases.

Textual Variance in Responses The non-normal distribution of app review response lengths raises considerable suspicion into the composition of the original app dataset. Inspecting the textual variance, i.e., the ratio of unique texts to the total number of texts, reveals that a large portion of reviews, responses and even review-response pairs occur multiple times in the app dataset. Table 6 shows textual variance among the three datasets.

As can be seen, only 40% of the responses in the app dataset are unique in contrast to approximately 94% for both English and German hospitality responses. This finding clearly indicates extensive repetition in the app dataset due to the frequent use of templated responses by app developers, which has also been reported by Hassan et al. (2018). For restaurant and hotel reviews, the story is closer to what we would expect, showing that the vast majority of responses are indeed unique. Naturally, this makes the

	Hosp. (en)	Hosp. (de)	Apps
Reviews	100.0%	99.99%	**70.04%**
Responses	94.2%	94.5%	**40.07%**
Review-response pairs	100.0%	100.0%	**88.09%**

Table 6: Percentage of unique review, response and review-response pair texts in each dataset.

task of learning to generate suitable responses in the hospitality domain much more challenging.

4 Conclusion and Future Work

We have applied a system designed for app review response generation to the domain of hospitality reviews in both English and German. According to a range of automatic evaluation metrics, the results of our experiments indicate that the benefits provided by the extended seq2seq model proposed by Gao et al. (2019) for their application domain do not transfer to our target domain. A subsequent empirical investigation into the different datasets has revealed that the nature of review-response pairs in the domain of smartphone app reviews differs significantly to that of the hospitality domain.

Given our results, we have established a preliminary baseline for automated review response generation in the domain of English and German hospitality reviews that uses a classic attentional seq2seq architecture. In future work, we intend to continue investigating potential improvements for this task, incorporating improved aspect-level sentiment detection and knowledge grounding techniques, as well as conducting more detailed evaluations with the help of human judges.

Acknowledgements

We are grateful to our industry partner re:spondelligent and the Swiss Innovation Agency InnoSuisse for their support of the ReAdvisor project (project number 38943.1 IP-ICT).

References

John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. 2017. Gated Multimodal Units for Information Fusion. *arXiv:1702.01992 [cs, stat]*, February.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May.

Andrea Di Sorbo, Sebastiano Panichella, Carol V. Alexandru, Junji Shimagaki, Corrado A. Visaggio, Gerardo Canfora, and Harald C. Gall. 2016. What would users change in my app? summarizing app reviews for recommending software changes. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering - FSE 2016*, pages 499–510, Seattle, WA, USA. ACM Press.

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

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. *arXiv:1802.06893 [cs]*, March.

Safwat Hassan, Chakkrit Tantithamthavorn, Cor-Paul Bezemer, and Ahmed E. Hassan. 2018. Studying the dialogue between users and developers of free apps in the Google Play Store. *Empirical Software Engineering*, 23(3):1275–1312, June.

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

C. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, Ann Arbor, Michigan, USA,.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January.

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]*, June.

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

Chunyu Li, Geng Cui, and Ling Peng. 2018. Tailoring management response to negative reviews: The effectiveness of accommodative versus defensive responses. *Computers in Human Behavior*, 84:272–284, July.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Stephen W. Litvin, Ronald E. Goldsmith, and Bing Pan. 2008. Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, 29(3):458–468, June.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *arXiv:1508.04025 [cs]*, September.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the 7th International Language Ressources and Evaluation (LREC)*, pages 1168–1171.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv:1509.00685 [cs]*, September.

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]*, June.

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

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.

Karsten Michael Tymann, Matthias Lutz, Patrick Palsbröker, and Carsten Gips. 2019. GerVADER - A German adaptation of the VADER sentiment analysis tool for social media texts. In *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen" (LWDA)*, pages 178–189, Berlin, Germany.

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

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3506–3510, Denver, USA, May. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*, February.

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.