

Sentence Alignment for Automatic Review-Response Generation in the Hospitality Domain

Benchmarking Sentence Alignment Techniques for Automatic Review-Response Generation in the Hospitality Domain

Renate Hauser^{1,*}, Tannon Kew^{1,*}

¹Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Abstract

Recently, online customer reviews have surged in popularity, placing additional demands on businesses to respond to these reviews. Conditional text generation models, trained to generate a response given an input review have been proposed to facilitate human authors in composing high quality responses. However, this approach has been shown to yield rather unsatisfying, generic responses while, in practice, responses are required to address reviews specifically and individually. We hypothesise that this issue could be tackled by changing the alignment paradigm and using sentence-aligned training data instead of document-aligned. Yet, finding correct sentence alignments in the review-response document pairs is not trivial. In this paper we investigate methods to align sentences based on computing the surface and semantic similarity between source and target pairs and benchmark performance for this rather challenging alignment problem.

1. Introduction

Online reviews have become an extremely popular and useful tool for both businesses and consumers. Today, there are numerous online platforms such as TripAdvisor, Yelp, or Booking.com, where customers can rate restaurants and hotels and write reviews about their visit. These reviews are an increasingly important source of information for potential or future customers [1]. This has led to a growing emphasis on effective online customer feedback management, which gives businesses the opportunity to influence the public discourse. However, many businesses lack the resources required to efficiently respond to such a high influx of reviews. This has given rise to research into how artificial intelligence can support the process of writing a review-response. Katsiuba et al. [2] investigate a neural sequence-to-sequence model trained to automatically generate a full response for a given review text. However, their proposed system tends to produce generic responses rather than addressing specific issues raised in the input review, which limits its applicability in practice. Since review responses vary greatly in both style and specificity, we hypothesise that the high degree of generic automatically generated responses is due to a fundamental alignment problem: at

the document-level, alignments between semantic units in the review text and the response text are often scarce. One potential solution would be to go below the document level and investigate review response generation at the sentence level. However, such an approach involves first extracting aligned sentence pairs from document-level review responses. In this paper, we investigate sentence alignment methods for review-response texts in the hospitality domain. Specifically, we consider two different approaches; one working at the surface level by making use of character n-grams, the other leveraging sentence embeddings to assess the semantic similarity of a given sentence pair.

2. Related Work


Automatic Review-Response Generation The success of sequence-to-sequence (seq2seq) encoder-decoder models [3] in the task of conversational modelling [4] has led to a significant interest in chatbots and conversational agents in industry as well as in academic research [5]. The task of automatic review response generation is similar to that of conversational agents, where the goal is to generate an adequate response for a given input. Also, both tasks face the challenge of modelling social skills in order to make the behaviour of the system more human-like [2, 6]. Consequently, seq2seq modelling techniques have been a popular choice for the task of automatic review response generation in various domains at the document level [7, 8, 9]. In an attempt to encourage indi-

SwissText 2022: Swiss Text Analytics Conference, June 08–10, 2022, Lugano, Switzerland

*Corresponding author.

✉ renate.hauser@uzh.ch (R. Hauser); kew@uzh.ch (T. Kew)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

vidual or specific responses these proposed approaches have typically focused on extending the basic seq2seq architecture to incorporate additional contextual information. Yet this fails to alleviate the problem entirely.

Sentence Alignment Sentence-aligned parallel corpora are a crucial prerequisite for language transduction tasks such as machine translation (MT) or conversational modelling. Yet the quality of systems trained on parallel data is largely dependent on the quality of the training data and poor alignments can severely harm the performance of the downstream application [10]. Consequently, there has been a great deal of research towards improving alignment algorithms. Algorithms operating on surface-level overlap as well as more complex neural approaches that consider deep semantic representations, have been proposed for MT [11, 12], automatic text simplification [13, 14] and paraphrasing [15]. To the best of our knowledge, we are the first to study the alignment of sentence pairs from online review-response documents and thus set out to benchmark this task.

3. Review-Response Sentence Alignment

Review response pairs typically resemble paragraphs, often containing multiple sentences or ideas. Yet responses differ greatly in terms of how specifically and individually they reply to a review. Additionally, responses frequently contain additional comments or information that does not refer to a point mentioned in the review. This leads to an inherent alignment problem at the document level, where semantically aligned units between the review and response are scarce and sometimes nonexistent. In contrast to other more common alignment tasks [e.g. 12, 14], review-response pairs exhibit a number of qualities that add complexity to this task. Firstly, alignments do not follow monotonicity constraints. Secondly, there is no guarantee that for any given review sentence a corresponding response sentence exists. This leads to a considerable number of documents that do not contain any alignment units. At the same time, N:M-alignments are also common due to a large degree of writing styles and the largely informal, free-form expression.

3.1. Data

With the overall aim of learning better semantic mappings between reviews and appropriate responses, we would like to identify and extract sentence pairs from a large collection of review-response document pairs. As a first step, we compile a dataset of approximately 500,000 documents consisting of review-response pairs for hotels

published on TripAdvisor. Scripts to reproduce our data will be made publicly available¹.

3.2. Method

In order to quantify an alignment, we rely on the intuition that aligned sentences should be semantically similar. For example, a review sentence that praises the quality of a hotel bed should be aligned with a response sentence that mentions sleep. As a first step, we need to segment the review-response pair documents into their constituent sentences. For this we use spaCy². We keep preprocessing minimal and simply apply lowercasing since casing is of little importance for the alignment task [11]. Secondly, following our underlying assumption, we compute a similarity score for each combination of review and response sentences in a document. To this end, we investigate two different approaches, namely, surface-level similarity based on character n-gram overlap and semantic similarity based on dense sentence embeddings. While the former offers a computationally cheap approach, it fails to account for sentences that are semantically similar but expressed differently, such as in the example above. Thus, we expect the latter approach to be most suitable. In a final step, we determine suitable thresholds for classifying an alignment unit and derive alignments based on these scores (Section 3.2).

Surface-Level Similarity To compute surface-level similarity between source and target sentences, we use the chrF metric presented by Popović [16]. The formula of chrF is as follows:

$$\text{chrF}\beta = (1 + \beta) * \frac{(\text{chrP} * \text{chrR})}{(\beta^2 * \text{chrP} + \text{chrR})}$$

Where chrP is the percentage of character n-grams in the hypothesis that are also present in the reference (i.e. precision) and chrR is the percentage of character n-grams in the reference that is also present in the hypothesis (i.e. recall). We investigate several settings. As we want to focus on content words rather than stopwords, we only consider n-gram orders starting from $n=4$. On the other hand, too high n-gram lengths might be too restricting to derive any useful alignments. We therefore set an upper limit of $n=6$.

Semantic Similarity To compute semantic similarity, we make use of BERT-based sentence embeddings (SBERT) [17]³ and compute the cosine similarity between sentence pairs. We consider two alternate framings for our task and compare SBERT models accordingly. The

¹<https://github.com/renatehauseruzh/rev-resp-sentalign>

²<https://spacy.io/>

³https://www.sbert.net/docs/pretrained_models.html

first of these frames a response sentence as a paraphrase of a review sentence for which we use the paraphrase-MiniLM-L3-v2 model. The second considers the task as a type of natural language inference (NLI), in which a response sentence may be logically inferred by a review sentence. Thus we also test the `nli-mpnet-base-v2` model.

To Align or not to Align Since we cannot assume the alignments to be monotonic, every pair of review and response sentences in a document is a potential candidate. For each such pair, a similarity score needs to be computed, resulting in a similarity matrix. An example is provided in Appendix A. As the time complexity of this comparison is $O(|S| * |T|)$, where S is the review text and T the response text, this is an expensive step. However, since the vast majority of the review-response documents contain less than ten sentences, this is still feasible. Given this matrix of similarity scores, the challenge is to determine an appropriate threshold for classifying an aligned sentence pair. In the following section we investigate suitable thresholds by inspecting the trade-off between precision and recall on a small manually-annotated gold-standard.

4. Experiments

Gold Standard To be able to automatically validate our candidate aligners, we compiled a manually annotated gold standard containing 115 review-response pair documents. These pairs were randomly sampled from the test split of our dataset. We then tasked two annotators, who were familiar with the alignment task, to annotate each review sentence with zero, one or multiple corresponding response sentences. This is a non-trivial task, as there is often no obvious distinction between a vague, generic response and no correspondence. The manual annotation yielded approximately 130 aligned sentence pairs. To measure the inter-annotator agreement (IAA) we used the Kappa statistic [18, 19]. The IAA for the gold standard reached a Kappa value of 0.64. This rather low agreement reflects the difficulty of the task.

Metrics We validate the output of the aligners with **precision**, **recall** and **F1 score**. The total number of alignments in the gold standard serve as the expected number of alignments that an aligner should extract. Because of the range of possible correct alignments, only considering complete matches would be too restricting. Therefore, we follow Jiang et al. [14] and report metrics for completely matching alignments (vs. partially matching alignments + non-alignments) as well as for partially + completely matching alignments (vs. non-alignments). We considered an alignment to be partially correct, if

at least one review sentence *and* one response sentence assigned by the aligner appears in an alignment in the gold standard.

Similarity Thresholds Low thresholds lead to large, unmotivated N:M-alignments, while high thresholds constrain the space of possible aligned segments too harshly. Therefore, we considered thresholds ranging from 0.02 to 0.16 and 0.1 to 0.6 for the chrF based approach and the cosine similarity approach, respectively. Manual investigation showed that 0.16 and 0.6, respectively, were reasonable thresholds, above which alignments were not found.

Performance We consider the results for complete matches to be a measure for how well the alignments reflect the human judgement in the gold standard. As can be seen in Figure 1, the higher n-gram orders (n), as well as higher thresholds (t), yield better measures for complete matches. However, looking at Figure 3, we can see a clear trade off between the total number of alignments and a high F1 score using the expected number of alignments of 130 from the gold standard. In fact, the three aligners with $n=6/t=0.06$, $n=5/t=0.08$, and $n=4/t=0.12$ yield comparable results in all three metrics while yielding a reasonable number of alignments.

As the focus lies on a good precision, we consider thresholds that are above the critical point where precision exceeds recall for the semantic similarity based approach, namely between $t=0.4$ and $t=0.5$. This choice is also confirmed by the number of extracted alignments that starts to drop off steeply at $t=0.4$. As can be seen in Figure 2, while partial matches are consistently higher for the NLI-BERT model, performance in terms of complete alignments is relatively equal for both SBERT models.

4.1. Results

To assess, how well the five most promising candidate aligners reflect the judgement of a human annotator, we conducted a small manual evaluation. We randomly sampled 50 alignments produced by each aligner, including 1:1 and N:M-alignments. We manually labeled these with either "valid" or "not valid". For both surface-level and semantic approaches, misalignments are typically caused by occurrence of named entities such as hotel or personal names. This influence is particularly observable for the $n=4$ chrF aligner. Meanwhile, NLI SBERT aligner is somewhat greedy and suffers from large, unmotivated N:M-alignments. Despite this, we found no evidence of one of the candidate aligners substantially outperforming the others with all yielding between 27 and 34 valid alignments out of 50.

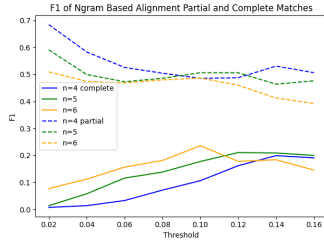


Figure 1: F1 score for partial and complete matches of the n-gram based Aligner for n-gram orders 4-6

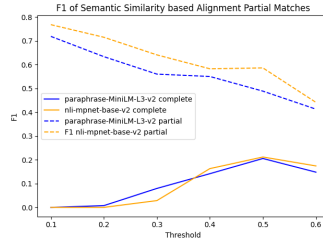


Figure 2: F1 score for partial and complete matches of the semantic similarity based Aligner for both models

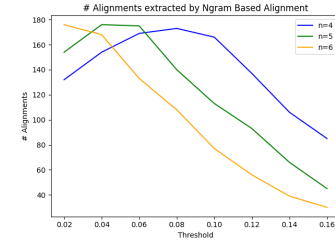


Figure 3: Total number of extracted alignments by the n-gram based Aligners for n-gram orders 4-6

4.2. Discussion & Future Work

Based on the results, we are able to derive five candidate approaches that yield approximately equally good results when evaluated manually. Our qualitative evaluation shows that our methods are capable of extracting relatively high-precision alignments, but suffer in terms of recall, leading to a low overall F1 score.

Given a large dataset of review-response documents, future work will benefit from the methods presented here to derive aligned sentence pairs for training review-response generation models on the sentence level. We hope that this will encourage the model to learn more semantically related mappings between the source and target texts.

We acknowledge that the gold standard used to validate a range of thresholds for our methods is relatively small and a larger gold standard would be beneficial for enhancing the reliability of these results. Furthermore, evaluation of sentence-level review response generation systems is also dependent on sentence-level test data. Thus, additional human annotation is required to construct a suitable evaluation set.

5. Conclusion

In this paper we investigated possible methods for deriving aligned sentences from hospitality review-response pairs. We believe that such alignments will be useful for improving the performance of downstream review response generation models by better mapping semantically related segments between the source and target texts. Automatic validation results and a small qualitative evaluation reveal that a relatively cheap character n-gram overlap metric allows us to align sentence pairs based purely on surface-level similarity with comparable results to a more expensive approach based on semantic similarity.

References

- [1] V. Browning, K. K. F. So, B. Sparks, The Influence of Online Reviews on Consumers' Attributions of Service Quality and Control for Service Standards in Hotels, *Journal of Travel & Tourism Marketing* 30 (2013) 23–40. URL: <https://doi.org/10.1080/10548408.2013.750971>. doi:10.1080/10548408.2013.750971.
- [2] D. Katsiuba, T. Kew, M. Dolata, G. Schwabe, Supporting online customer feedback management with automatic review response generation, in: *The 55th Hawaii International Conference on System Sciences, HICSS, 2022*, pp. 226–236. URL: <https://doi.org/10.5167/uzh-212773>.
- [3] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to Sequence Learning with Neural Networks, in: *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, Montreal, 2014. URL: <https://arxiv.org/abs/1409.3215v3>.
- [4] O. Vinyals, Q. Le, A Neural Conversational Model, *arXiv:1506.05869 [cs]* (2015). URL: <http://arxiv.org/abs/1506.05869>.
- [5] P. Gentsch, *Künstliche Intelligenz für Sales, Marketing und Service*, Springer Fachmedien Wiesbaden, Wiesbaden, 2018. URL: <http://link.springer.com/10.1007/978-3-658-19147-4>. doi:10.1007/978-3-658-19147-4.
- [6] S. Diederich, M. Janßen-Müller, A. Brendel, S. Morana, Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent, in: *Proceedings of International Conference on Information Systems (ICIS)*, Munich, 2019. URL: https://aisel.aisnet.org/icis2019/smart_service_science/smart_service_science/2/.
- [7] C. Gao, J. Zeng, X. Xia, D. Lo, M. R. Lyu, I. King, Automating App Review Response Generation, in: *Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering*

- (ASE), San Diego, USA, 2019, pp. 163–175. doi:10.1109/ASE.2019.00025.
- [8] L. Zhao, K. Song, C. Sun, Q. Zhang, X. Huang, X. Liu, Review Response Generation in E-Commerce Platforms with External Product Information, in: The World Wide Web Conference on - WWW '19, San Francisco, 2019, pp. 2425–2435. URL: <http://dl.acm.org/citation.cfm?doid=3308558.3313581>. doi:10.1145/3308558.3313581.
- [9] T. Kew, M. Amsler, S. Ebling, Benchmarking Automated Review Response Generation for the Hospitality Domain, in: Proceedings of Workshop on Natural Language Processing in E-Commerce, Barcelona, 2020, pp. 43–52. URL: <https://aclanthology.org/2020.ecomnlp-1.5>.
- [10] H. Khayrallah, P. Koehn, On the Impact of Various Types of Noise on Neural Machine Translation, in: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Melbourne, 2018, pp. 74–83. URL: <https://aclanthology.org/W18-2709>. doi:10.18653/v1/W18-2709.
- [11] R. Sennrich, M. Volk, MT-based Sentence Alignment for OCR-generated Parallel Texts, in: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers, Denver, 2010. URL: <https://aclanthology.org/2010.amta-papers.14>.
- [12] B. Thompson, P. Koehn, Vecalign: Improved Sentence Alignment in Linear Time and Space, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 2019, pp. 1342–1348. URL: <https://www.aclweb.org/anthology/D19-1136>. doi:10.18653/v1/D19-1136.
- [13] W. Xu, C. Callison-Burch, C. Napoles, Problems in Current Text Simplification Research: New Data Can Help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://doi.org/10.1162/tacl_a_00139. doi:10.1162/tacl_a_00139.
- [14] C. Jiang, M. Maddela, W. Lan, Y. Zhong, W. Xu, Neural CRF Model for Sentence Alignment in Text Simplification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 7943–7960. URL: <https://aclanthology.org/2020.acl-main.709>. doi:10.18653/v1/2020.acl-main.709.
- [15] R. Barzilay, N. Elhadad, Sentence Alignment for Monolingual Comparable Corpora, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 2003, pp. 25–32. URL: <https://aclanthology.org/W03-1004>.
- [16] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, 2015, pp. 392–395. URL: <https://aclanthology.org/W15-3049>. doi:10.18653/v1/W15-3049.
- [17] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 2019, pp. 3980–3990. URL: <https://www.aclweb.org/anthology/D19-1410>. doi:10.18653/v1/D19-1410.
- [18] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46. doi:<https://doi.org/10.1177/001316446002000104>.
- [19] J. Carletta, Assessing agreement on classification tasks: the kappa statistic, CoRR cmp-lg/9602004 (1996). URL: <http://arxiv.org/abs/cmp-lg/9602004>.

A. Appendix

	resp0	resp1	resp2	resp3	resp4
rev0	0,26	0,03	0,06	0,25	0,20
rev1	0,23	0,41	0,17	0,10	0,24
rev2	0,17	0,23	0,17	0,15	0,29
rev3	0,24	0,11	0,44	0,24	0,11
rev4	0,14	0,06	0,09	0,07	0,06
rev5	0,11	0,16	0,22	0,37	0,20
rev6	0,29	0,23	0,42	0,38	0,31

Figure 4: Similarity matrix for the review-response document in Table 1, computed with a sentence embeddings based approach using the *nli-mpnet-base-v2* model. The alignments annotated by a human annotator are written in bold.

Review

- 0 On Way to Excellence —SEP—
- 1 In midst of renovations, but really did not hear much or was disturbed.,
- 2 Would love it if the M Lounge remained on first floor - ideal location!,
- 3 Concern has to do with F&B and housekeeping.,
- 4 Food service items remained in hall for DAYS!,
- 5 Housekeeping carts blocked movement in hallway - why can't Marriott be cutting edge and develop or adopt Asian hidden housekeeping?,
- 6 Finally, it seems as though if the hotel strategically places tip envelopes for housekeeping staff why not for other employees?

Response

- 0 Dear Jim H, Thanks for your review, and feedback after your recent stay.,
- 1 We are happy to know you enjoyed your stay, and that our renovations at the hotel did not disturb you at all!,
- 2 We do apologize for the missteps with the housekeeping department, and appreciate your comments.,
- 3 We do our best to extend the Marriott standards, and quality our guests have come to expect, and have taken note of your feedback for our team to review.,
- 4 We appreciate your review, and hope you will return to enjoy all of the new renovations being done at the hotel!

Table 1

An example of a sentence-split review-response document