Effectiveness of Data Augmentation to Identify Relevant Reviews for Product Question Answering

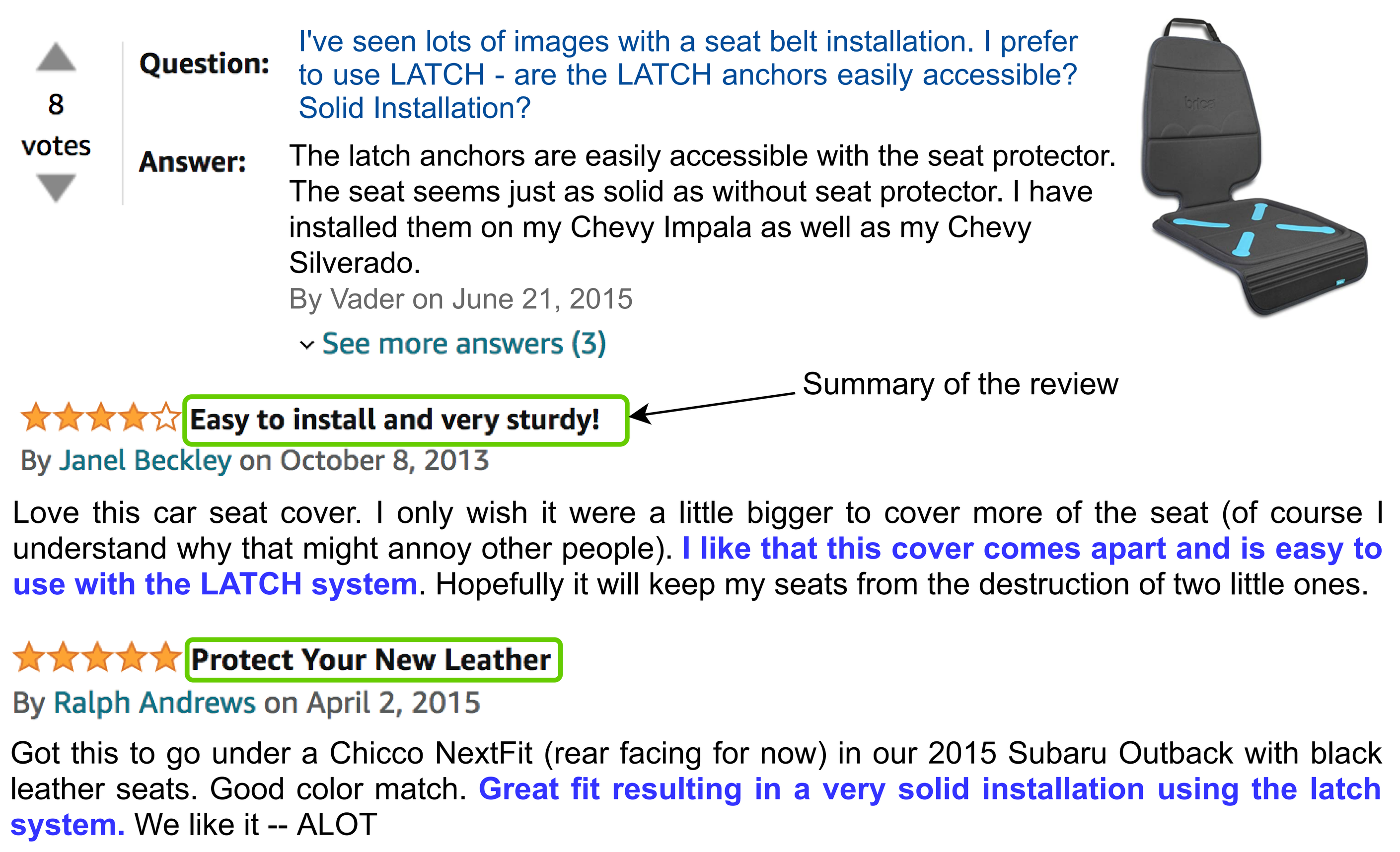
Kalyani Roy, Avani Goel, Pawan Goyal

Dept. of CSE, IIT Kharagpur, India

1 Objective

- With the rapid growth of e-commerce, there is a need for providing automated answers to customer posted questions.
- A potential solution would be to automatically identify the answer to the new question from the already posted reviews.
- We attempt to improve the review ranking task with a better transformer-based model and Data Augmentation technique using T5.

2 Example Product Snapshot



The plausible answers to the question from the reviews are highlighted in **blue**. The texts inside the **green** rectangles are the summaries of the reviews.

3 Problem Statement

Given a question Q about a product P and a set of reviews $R = \{r_1, r_2, r_3, \dots\}$ for that question, our aim is to provide a ranked list of reviews $R' = \{r_1', r_2', r_3', \dots\}$ where r_i' are ranked in order of decreasing relevance with question Q.

4 Synthetic Training Data Creation

- Due to lack of a labeled dataset which contains question along with its relevant reviews, we resort to synthetic training settings
- We use **Amazon Question Answer Dataset** and **Amazon Review Dataset**.
- For every question Q, we select the positive response A_p with the most helpful votes. We randomly select an answer from a different question of the same product as the negative answer A_n.
- By utilizing Q, A_p, and A_n, we train a classifier with BERT that predicts whether an answer is relevant to a given question or not.
- Initially, we choose the top 100 review sentences using BM25. We refine this list with the trained classifier, and we take the top 10 reviews as the set R for each question.
- **QAR Dataset** is our synthetic dataset that consists of Q, R, A_p, and A_n.

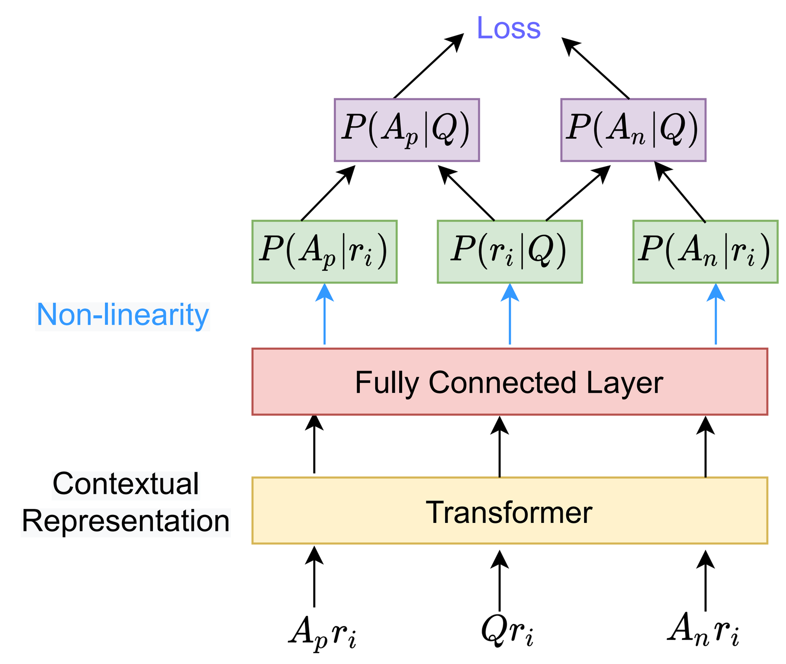
5 Data Augmentation

- We consider the review of a product as "context" and the summary as the "answer" and attempt to generate question with it.
- **FT1 model** : The T5 model that is already fine-tuned on the question generation task on the SQuAD dataset.
  FT2 model : We use FT1 to further fine-tune the model on the SubjQA dataset .
- We employ both FT1 and FT2 to generate new questions. We discard the questions containing possessive pronouns, and questions that do not end with '?'.
- We split the review into sentences to form the review list R, we take the review summary as our positive answer A_p. We combine this generated data with QAR Dataset to get **QAR-aug Dataset**.

Review: Since it's a long hose there was a bit of kink that I didn't notice. Turned the water on and it instantly blew a large hole in the hose, rendering it useless. The plastic is pretty thin, so be warned to carefully check for any kinks.
Summary: On 3rd time used it had a huge hole in it.
Gen. Que. with FT1: What was the problem with the hose?

Review: This is a very cute decal for my baby girl's room, but it is not as tall as the picture depicts. I would suggest putting it behind the baby bed or dresser to hide how short it really is.
Summary: Cute, but not as big as in the photo.
Gen. Que. with FT2: How is the decal?

6 Modelling Question Review Relevance



- Our base model simultaneously learns relevance functions between 'question and review' and 'review and answer'

$$P(r_i|Q) = softmax(W^T\ Tran(Qr_i))$$
$$P(A_p|r_i) = \sigma(W^T\ Tran(A_p r_i))$$
$$P(A_n|r_i) = \sigma(W^T\ Tran(A_n r_i))$$

W is a learnable matrix, and $Tran(XY)$ denotes the representation of the paired sentences X and Y using transformer.

- We score an answer by combining the learnt relevance functions :

$$P(A_p|Q) = \sum_{r_i} P(A_p|r_i)\ P(r_i|Q)$$
$$P(A_n|Q) = \sum_{r_i} P(A_n|r_i)\ P(r_i|Q)$$

- The objective of the model is to rank positive answer higher than the negative answer.

$$loss = max(0, P(A_p|Q) - P(A_n|Q) - \delta)$$

- At the time of inference, we use the learned relevance function between 'question and review' to rank the reviews related to a question.

7 Experiments

- For evaluating the models, we use the annotated dataset from RIKER [2]

	nDCG	BM25	RIKER [2]	Bert-RR [1]		Deberta-RR	
	(%)	-	-	×	✓	×	✓
Baby	@1	42.08	-	55.00	62.08	59.17	**64.16**
	@3	38.64	-	57.95	61.32	60.70	**63.28**
	@5	44.34	-	61.60	63.79	64.71	**66.85**
	@10	52.76	64.80	66.95	69.31	67.51	**70.78**
Tools & Home	@1	37.50	-	40.83	**46.25**	41.25	44.17
	@3	38.44	-	45.32	**45.82**	44.14	45.46
	@5	38.36	-	45.76	46.60	46.68	**46.90**
	@10	43.81	45.12	49.67	50.69	50.13	**50.75**
Patio Lawn & Garden	@1	31.25	-	45.00	49.16	48.33	**50.00**
	@3	34.70	-	44.46	47.13	48.96	**52.11**
	@5	36.40	-	46.88	50.52	49.99	**52.35**
	@10	44.04	55.91	55.01	58.30	57.17	**58.99**
Average	@1	36.94	-	46.94	52.50	49.58	**52.78**
	@3	37.26	-	49.24	51.42	51.27	**53.62**
	@5	39.70	-	51.41	53.64	53.79	**55.37**
	@10	46.87	55.28	57.21	59.43	58.27	**60.17**

Table 1:Performance of all the models in three categories. The cross and the checkmark symbols indicate that the model is trained with the QAR dataset and the augmented dataset QAR-aug, respectively.

8 Conclusion

- We utilize transformer-based models to provide relevant reviews to a new question.
- We present a data augmentation technique by fine-tuning the T5 model to generate new questions from customer reviews.
- Experimental results show substantial improvements over the existing approaches using the data augmentation technique.

References

[1] Zhang, S., Lau, J.H., Zhang, X., Chan, J., Paris, C.: Discovering relevant reviews for answering product-related queries. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 1468–1473 (2019)

[2] Zhao, J., Guan, Z., Sun, H.: Riker: Mining rich keyword representations for interpretable product question answering. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1389–1398 (2019)

Dataset: `https://github.com/kalyani-roy/DARR`