Is it Dish Washer Safe? Automatically Answering "Yes/No" Questions using Customer Reviews

Daria Dzendzik
ADAPT Centre
Dublin City University
Dublin, Ireland
daria.dzendzik@adaptcentre.ie

Carl Vogel
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
vogel@tcd.ie

Jennifer Foster
ADAPT Centre
Dublin City University
Dublin, Ireland
jennifer.foster@adaptcentre.ie

Abstract

It has become commonplace for people to share their opinions about all kinds of products by posting reviews online. It has also become commonplace for potential customers to do research about the quality and limitations of these products by posting questions online. We test the extent to which reviews are useful in question-answering by combining two Amazon datasets, and focusing our attention on yes/no questions. A manual analysis of 400 cases reveals that the reviews directly contain the answer to the question just over a third of the time. Preliminary reading comprehension experiments with this dataset prove inconclusive, with accuracy in the range 50-66%.

1 Introduction

Consumers often carry out online research about a product before purchasing. This can take the form of reading consumer reviews and/or asking specific questions on online fora. In this paper we ask whether a question-answering (QA) system can utilize the information in consumer reviews when answering yes/no questions about a product.

We compile a dataset of questions about Amazon products together with consumer reviews of the same products, and manually analyse a sample of 100 questions from four domains. We find that the reviews contain the answer in only 45% of cases. In 36% of cases, the answer is directly expressed in at least one of the reviews, and 9% of the time, it is indirectly expressed. This suggests that reviews can *sometimes* be useful and so we go on to experiment with QA systems that use the reviews in addition to the question. We focus on yes/no questions. Being able to answer these is not only an indicator of whether reviews will be useful for other question types but is also a signal of how much comprehension is actually taking place.

In our preliminary experiments with three domains from this new dataset, we compare systems which attempt to answer a yes/no question based on the question alone to those that also use related reviews. We experiment with two methods for selecting relevant sentences from the reviews, and with various representations for encoding the questions and reviews including bag-of-words, word2vec (Le and Mikolov, 2014), ELMO (Peters et al., 2018), and BERT (Devlin et al., 2018). On the development set, our systems tend to outperform the chance baseline but not by a large margin – our development set results range from 50 to 66%. Over the three domains, we also find that the question-only systems tend to perform as well as and sometimes outperform those which also use the reviews, suggesting that separating the answers from the noise in these reviews is not straightforward.

2 Related Work

A number of studies have explored the use of customer reviews in retrieval and question answering. Using Amazon data, Yu et al. (2018) develop a framework which returns a ranked list of sentences from reviews or existing question-answer pairs for a given question. Xu et al. (2019) create a new dataset comprising Amazon laptop reviews and questions and Yelp restaurant reviews and questions, where reviews are used to answer questions in multiple-turn dialogue form. Bogdanova et al. (2017) and Bogdanova and Foster (2016)) do not use review data but also focus on QA over user-generated content, attempting to find similar questions or rank answers in user fora. We use the same Amazon data as Yu et al. (2018) but consider a wider set of domains (they consider only two), and attempt to directly answer yes/no questions. To the best of our knowledge, the novelty in

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 1–6
Minneapolis, Minnesota, June 3 - 5, 2019.

our work lies in trying to directly answer customer questions using user-generated reviews.

Unlike popular Reading Comprehension datasets such as MovieQA (Tapaswi et al., 2016) and SQuAD (Rajpurkar et al., 2018, 2016), which are created by crowdsourcing, we work with authentic user-generated data. This means that the data is collected from sources where users spontaneously created content for their own purposes. Since there is no guarantee that reviews contain text related to the question, there is no span data that can be reliably used to provide the answer. This, together with the considerable volume of review text, contributes to the difficulty of the task.

3 Data

We work with two Amazon datasets: the first, He and McAuley (2016),[1] is a collection of product reviews from 24 domains. The second, Wan and McAuley (2016); McAuley and Yang (2016), contains questions and answers about products from 21 domains. These two datasets have 17 domains in common and can be matched using the Amazon Standard Identification Number (ASIN).

In order to obtain data with reviews, questions and answers, we first select all those products which contain reviews and questions, focusing on yes/no questions. We observe that the majority of questions can be answered "Yes" (65-75% depending on the domain), so we balance the data by selecting an equal amount of yes/no questions. This results in 80391 questions about 40806 products – see Table 1 for more details.

All data is fully user-generated except the answer tags which are provided by McAuley and Yang (2016). An example of the combined data is shown below (we keep the original spelling).

> Reviews (R): *...I was a little surprised at how much time it took to assemble. There were alot of the smaller parts that I would have assumed pre-assembled that weren't...*[2]
> Question (Q): *Does it come assembled*
> Answer (A): *No, count on, at least an hour to assemble.* (Answer Tag (AT): *No*)

The authentic user-generated nature of this dataset makes it significantly different from other reading comprehension datasets. Table 2 shows a comparison with MovieQA and SQuAD2.0. The length of questions is almost the same (10-11.5 words) in all three datasets, although the number of instances (movie plots for MovieQA, topic for SQuAD and product for Amazon) is significantly bigger. Moreover, the average length of context in the Amazon data is unquestionably larger than in the MovieQA and SQuAD: 188 vs 35 and 5 sentences, and 3265 vs 728 and 117 words.

To better understand the nature of questions we carried out some additional analysis looking into question formulation. 21% of questions are formulated with more than one sentence (16–31% depending on the domain), more than 25% of questions (20686) start with the word *Does*, and more than 15% (>12000) with *Can*, *Is* or *Will*.

4 Do reviews contain the answers?

According to Kaushik and Lipton (2018) reading comprehension datasets are not studied enough in terms of difficulty. We conduct a manual analysis to better understand the relationship between questions and reviews, to assess the feasibility of using user reviews to answer user questions and to estimate an upper bound on system performance. 100 questions from four domains are analysed. We define seven classes of questions:

Easy: Questions are clearly answered in the reviews, e.g.

(1) R: *... I used two of these, one for each side of the bed.*
Q: *can this product be used if 2 bed rails are needed for one bed?* (AT: *Yes*)

Error: Questions where the answer tag contradicts the user-provided answer, e.g.

(2) Q: *Can you mount this upside down i.e. The receiver on top of the bumper?*
A: *I don't see why not, the is nothing preventing you.* (AT: *No*)

Indirect: Questions which can be indirectly answered by the review, e.g.

(3) R: *... it doesn't give an exact voltage and maxes out at 12.7 volts*
Q: *Can this be used to charge a 48v battery?* (AT: *No*)
Q: *Is this a good charger/jump starter for a 12v deep cell battery?* (AT: *Yes*)

[1] More details here: http://jmcauley.ucsd.edu/data/amazon/ – last verified (l.v.) 02/20119

[2] To save space we provide only part of user review

Domain	# P	Question			Review		
		#	# S	# W	#	# S	# W
Automotive	574	1113	1469	14158	7276	34112	618k
Baby	1105	2163	2793	26513	48835	281953	5083k
Beauty	1522	2763	3537	29105	39381	205000	3437k
Cell Phones & Accessories	2401	5711	6836	60946	72407	369241	6836k
Clothing Shoes & Jewellery	251	479	622	5166	4349	19815	310k
Electronics	13683	27877	35073	330340	691400	4130768	78242k
Grocery & Gourmet Food	758	1223	1549	12288	17436	85097	1417k
Health & Personal Care	3259	5833	7491	63520	93189	488411	8658k
Home & Kitchen	6527	12003	15580	138021	215194	1230269	21313k
Musical Instruments	227	399	505	4642	3150	15642	284k
Office Products	624	1269	1574	14047	10200	79444	1598k
Patio Lawn & Garden	352	637	851	7935	4576	35604	712k
Pet Supplies	1132	1945	2722	25428	37538	202237	3574k
Sports & Outdoors	3455	6699	8366	75405	90501	452578	7958k
Tools & Home Improvement	2619	5245	6883	65978	47491	270010	4983k
Toys & Games	1719	3205	3975	34301	39456	215718	3712k
Video Games	598	1827	2192	19902	32790	291642	6071k
Total	40806	80391	102018	927695	1455169	8407541	154m

Table 1: Balanced yes/no dataset statistics per domain: Number of products (**P**) which have yes/no questions, number of questions (**# Question**), count of sentences in questions (**S**), total number of words in questions (**W**), total number of reviews (**# Reviews**), all number of sentences in reviews, total number of words in reviews.

Dataset	# I	# T	# Q	AVG # W in Q	AVG # S in T	AVG # W in T
Amazon yes/no	40806	1455169	80391	11.50	188.49	3265.39
MovieQA	408	408	14944	9.34	35.26	727.91
SQuAD2.0	442	20239	142192	9.90	4.97	117.18

Table 2: Comparison of balanced yes/no dataset with MovieQA and SQuAD2.0: Number of instances (**I**): articles, movie plots, or products; Number of text passages (**T**): context, reviews, or plots; Average number of words in the question (**AVG W in Q**), sentences in the text (**AVG S in T**), and words in the text (**AVG W in T**)

Real-world: Questions where the review does not contain the answer but where an educated guess can be made using common sense or real-world knowledge, e.g.

(4) *Q: Can I use the cloth to clean the keys on my clarinet? (AT: Yes)*

(5) *Q: Has anyone traveled with this stroller on an airplane? (AT: No)*

Opinion: Questions which can be answered differently based on different reviews (6) or when the answer and review contain contradictory information (7). Often such questions ask for an opinion, so the answer depends on the user providing it, e.g.

(6) R: *...all in all these pans are worthless ...so many folks have had a horrid experience!!! ...At $15.00, it's a good pan for my purposes ...This pan is awesome for the price*[3]
Q: *is this item any good?* (AT: *No*)

(7) R: *...but it seems to get a little hot and makes a plastic noise under the sheet...*
Q: *does it make noise when baby moves around?*
A: *No not with a sheet on it.* (AT: *No*)

Unrelated: Questions which are asked not about the product but about service and delivery, e.g.

(8) Q: *Is there a warranty when you buy it from amazon?*

No answer: Questions which cannot be answered without additional information, i.e. reviews do not contain the required information.

The *indirect* and *real-world* classes can be considered to be difficult questions. However, in general, we believe that the *easy*, *indirect* and *real-world* question classes can be answered without resorting to guessing.

Detailed information is provided in Table 3. Around 53.5% of questions can be answered (36.5% are easy and 17% are difficult). Although it is difficult to conclude too much from this sample of 400, we can roughly estimate that the best performance we could expect from an automatic QA system would be around 77%. This means the

[3]The sentences are taken from different reviews.

Domain	Answerable			Guessing				Total (%)
	Easy	Indirect	Real-world	Opinion	No answer	Unrelated	Error	
Home & Kitchen	31	9	4	9	34	1	12	100 (0.83)
Beauty	37	8	6	7	33	3	6	100 (3.6)
Baby	44	11	8	11	21	1	4	100 (4.6)
Clothing Shoes & Jewellery	34	8	14	10	23	1	10	100 (20.9)
Total	146	36	32	37	111	6	32	400

Table 3: Selection of 100 questions from 4 domains for manual analysis. The last column contains the percentage of the analysed questions from each domain (eg. 100 is 4.6% of the Baby question data, 3.6% of Beauty, etc.).

system answers all answerable questions correctly (53.5%) and guesses half of those questions which cannot be answered (23.25%).

5 Preliminary Experiments

5.1 Approach

In order to establish some baselines on this dataset and task, we carry out preliminary binary classification experiments with a sample of the domains. There are three aspects to the systems we evaluate:

Text Representations To represent questions and reviews we experiment with simple *Bag of words (BOW)*, *word2vec* (Le and Mikolov, 2014), *Deep contextualized word representations (ELMO)* (Peters et al., 2018)[4] and *Bidirectional Encoder Representations from Transformers (BERT)* (Devlin et al., 2018).[5]

Review Filtering The reviews in our dataset are long compared to the answer passages used in other reading comprehension tasks – see Tables 1 and 2. Pascanu et al. (2012) report that long sequences are hard to process from both time and resource perspectives with sequence-to-sequence models. Therefore, rather than using the full text of the reviews, we use string similarity to select only those sentences that are likely to be relevant to the question. We base our selection method on our previous work which achieved state-of-the-art performance on the MovieQA reading comprehension task (Dzendzik et al., 2017; Tapaswi et al., 2016). Review sentences are compared to the question using cosine similarity of tf-idf representations, bag of words overlap, character n-grams, and window slide. Two sets of sentences are extracted. The first one is based on sentence union, in other words, all sentences which have been marked as relevant by any of the metrics are selected. The second one is based on sentence intersection, i.e. only sentences which have been marked as relevant by more than one similarity metric are selected.

Binary Classifier Following our previous work (Dzendzik et al., 2017) we use logistic regression with the bag-of-word, ELMO and word2vec representations. In the BERT experiments we add a softmax classification layer on top of the final hidden state of the transformer.

5.2 Experimental Setup

We compare systems which use the review and question text to systems which just use the question text. We select three of the four domains used for manual analysis.We exclude *Clothing Shoes & Jewellery* due to the small number of questions.

Table 4 represents the number of questions in the training, development and test set and the ratio of "Yes" and "No" questions in each of them.[6] The evaluation metric is accuracy.

For the bow and word2vec experiments, we normalize the text and remove stop words. We use a pre-trained Google News word2vec model. Every text is encoded as a sum of its word vectors and normalized. For the ELMO representation we average three layers of ELMO output and represent a sentence as concatenation of its words vectors.

In the question-only BERT experiment, we perform single-sentence classification (Devlin et al., 2018, Fig. 3b). In the experiments where the reviews are used, we perform sentence[7] pair classification where the question is the first sentence and the review text the second (Devlin et al., 2018, Fig. 3a). We use the pretrained models B_{base} and B_{large}. Both of them are uncased.

[4] We use https://github.com/allenai/allennlp – l.v. 04/2019

[5] We use https://github.com/huggingface/pytorch-pretrained-BERT – l.v. 04/2019

[6] Although data is balanced, we divide the dataset by products so some fluctuation between the percentage of "Yes" and "No" questions is possible.

[7] We use the word "sentence" in the same way as Devlin et al. (2018)

Domain	Training			Development			Test		
	All	Yes %	No %	All	Yes %	No %	All	Yes %	No %
Home & Kitchen	8502	50.02	49.98	1688	50.00	50.00	1813	49.92	50.08
Beauty	1965	49.21	50.79	383	51.17	48.83	415	52.77	47.23
Baby	1542	50.26	49.74	300	48.00	52.00	321	50.78	49.22

Table 4: Domain split of the training, development and test sets using the number of questions and the ratio of "Yes" and "No" questions in each of them.

Method	Baby			Beauty			Home and Kitchen		
	Q Only	Q + Review		Q Only	Q + Review		Q Only	Q + Review	
		Intersection	Union		Intersection	Union		Intersection	Union
LR bow	65.33	58.33	59.66	62.14	59.53	59.01	58.17	55.27	55.50
LR w2 v	59.33	60.67	59.66	57.44	55.09	56.40	59.03	55.69	57.17
LR elmo	53.33	56.99	56.99	60.83	60.57	55.35	52.37	49.17	48.93
B_{base}	**65.66**	55.67	60.00	**64.75**	50.91	60.05	**61.67**	**64.04**	63,21
B_{large}	52.00	**63.00**	48.00	48.82	62.92	**64.23**	50.00	62.20	63.68

Table 5: Results on development set of Logistic Regression (LR) applied to bag of word (**bow**), word2vec (**w2v**) and ELMO (**elmo**) representations, and BERT models (**B base** and **B large**) for 3 domains. The best question-only (**Q only**) and question+review (**Q+Review**) systems are in bold.

Model	Data	Accuracy
	Baby	
B_{base}	Question Only	64.17
$B_{large}Inter$	Question+Review	49.22
	Beauty	
B_{base}	Question Only	64.41
$B_{large}Union$	Question+Review	47.22
	Home and Kitchen	
B_{base}	Question Only	58.57
$B_{base}Inter$	Question+Review	60.23

Table 6: Results on test set with best-scoring question and review+question systems on development set.

5.3 Results

The development set results are shown in Table 5. Apparently, questions themselves provide some information and help find the correct answer: two out of three domains show the best performance using the question only. Only the *Home and Kitchen* domain shows better performance with the question+review systems. When selecting sentences from the reviews, there is no clear winner between the intersection and union methods. It varies according to method and domain.

BERT, the base and large models, perform better then logistic regression on the development set so we apply the best question-only and question+review models to the test set (Table 6). Performance drops below chance for two domains. It remains to be seen why we are seeing this unstable performance.

6 Conclusions

We introduce a fully user-generated reading comprehension dataset by composing two existing datasets into a new one designed to addresses yes/no questions about products using reviews. All data in this work is substantial and comes from real users. We provide a preliminary analysis of data and show that reviews can, to some extent, be used to answer yes/no questions .

We build several baseline systems. Although performance does not reach our estimated upper bound of 77%, our results show that they are doing more than mere majority classification. The relatively good performance of the question-only systems leads us to believe that the systems are applying closed-world assumptions by associating terms in the training set questions with terms in the test set questions.

Each of the components of our systems can be replaced or improved. Our immediate next step is to investigate more closely the part of the system which selects relevant sentences from the reviews.

Acknowledgments

This research is supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and the European Regional Development Fund. We thank Koel Dutta Chowdhury, Andrew Dunne, Dimitar Shterionov, Eva Vanmassenhove and Henry Elder for discussions and support.

References

Dasha Bogdanova and Jennifer Foster. 2016. This is how we do it: Answer reranking for open-domain how questions with paragraph vectors and minimal feature engineering. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1290–1295.

Dasha Bogdanova, Jennifer Foster, Daria Dzendzik, and Qun Liu. 2017. If you can't beat them join them: Handcrafted features complement neural nets for non-factoid answer reranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 121–131, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Daria Dzendzik, Carl Vogel, and Qun Liu. 2017. Who framed Roger Rabbit? answering questions about movie plot. The Joint Video and Language Understanding Workshop: MovieQA and The Large Scale Movie Description Challenge (LSMDC), at ICCV 2017, 23th of October, Venice, Italy.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5010–5015.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 625–635.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4631–4640.

Mengting Wan and Julian McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *ICDM*, pages 489–498. IEEE.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Review conversational reading comprehension. *CoRR*, abs/1902.00821.

Qian Yu, Wai Lam, and Zihao Wang. 2018. Responding e-commerce product questions via exploiting QA collections and reviews. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2192–2203.