SUBJQA: A Dataset for Subjectivity and Review Comprehension

Johannes Bjerva[1,3*] **Nikita Bhutani**[2*]
Behzad Golshan[2]   **Wang-Chiew Tan**[2]   **Isabelle Augenstein**[1]

[1]Department of Computer Science, University of Copenhagen
[2]Megagon Labs, Mountain View
[3]Department of Computer Science, Aalborg University
jbjerva@cs.aau.dk, augenstein@di.ku.dk
{nikita,behzad,wangchiew}@megagon.ai

Abstract

Subjectivity is the expression of internal opinions or beliefs which cannot be objectively observed or verified, and has been shown to be important for sentiment analysis and word-sense disambiguation. Furthermore, subjectivity is an important aspect of user-generated data. In spite of this, subjectivity has not been investigated in contexts where such data is widespread, such as in question answering (QA). We develop a new dataset which allows us to investigate this relationship. We find that subjectivity is an important feature in the case of QA, albeit with more intricate interactions between subjectivity and QA performance than found in previous work on sentiment analysis. For instance, a subjective question may or may not be associated with a subjective answer. We release an English QA dataset (SUBJQA) based on customer reviews, containing subjectivity annotations for questions and answer spans across 6 domains.

1 Introduction

Subjectivity is ubiquitous in our use of language (Banfield, 1982; Quirk et al., 1985; Wiebe et al., 1999; Benamara et al., 2017), and is therefore an important aspect to consider in Natural Language Processing (NLP). For example, subjectivity can be associated with different senses of the same word. BOILING is objective in the context of *hot water*, but subjective in the context of a person *boiling with anger* (Wiebe and Mihalcea, 2006). The same applies to sentences in discourse contexts (Pang and Lee, 2004). While early work has shown subjectivity to be an important feature for low-level tasks such as word-sense disambiguation and sentiment analysis, subjectivity in NLP has not been explored in many contexts where it is prevalent.

In recent years, there has been renewed interest in areas of NLP for which subjectivity is important, and a specific topic of interest is question answering (QA). This includes work on aspect extraction (Poria et al., 2016), opinion mining (Sun et al., 2017) and community QA (Gupta et al., 2019). Many QA systems are based on representation learning architectures (e.g. Devlin et al. (2018); Radford et al. (2018)) that are typically trained on factual, encyclopedic knowledge such as Wikipedia or books. It is unclear if these architectures can handle subjective statements such as those that appear in reviews.

The interactions between QA and subjectivity are even more relevant today as users' natural search criteria in many domains, including products and services, have become increasingly subjective. According to McAuley and Yang (2016a), around 20% of product queries were labeled as being 'subjective' by workers. Their questions can often be answered by online customer reviews, which tend to be highly subjective as well. Although QA over customer reviews has gained traction recently with the availability of new datasets and architectures (Grail and Perez, 2018; Gupta et al., 2019; Fan et al., 2019; Xu et al., 2019b; Li et al., 2019), these are agnostic with respect to how subjectivity is expressed in the questions and the reviews. Furthermore, the datasets are either very small (< 2000 questions) or have target-specific question types (e.g., yes-no). Most QA datasets and systems focus on answering questions over factual data such as Wikipedia articles and News (Joshi et al., 2017; Trischler et al., 2017; Rajpurkar et al., 2018; Abdou et al., 2019; Reddy et al., 2019). In this work, on the other hand, we focus on QA over subjective data from reviews on product and service websites.

In this work, we investigate the relation between subjectivity and question answering (QA) in the context of customer reviews. As no such QA

---

*JB and NB contributed equally to this work.

dataset exists, we construct SUBJQA.[1] In order to capture subjectivity, our data collection method builds on recent developments in opinion extraction and matrix factorisation, instead of relying on linguistic similarity between questions and reviews (Gupta et al., 2019). SUBJQA includes over 10,000 English examples spanning 6 domains that cover both products and services. We find a large percentage of subjective questions and answers in SUBJQA, as 73% of the questions are subjective and 74% of the answers are subjective. Experiments show that existing QA systems trained to find factual answers struggle with subjective questions and reviews. For instance, fine-tuning BERT (Devlin et al., 2018), a state-of-the-art QA model, yields 92.9% F_1 on SQuAD (Rajpurkar et al., 2016), but only achieves a mean score of 74.1% F_1 across the domains in SUBJQA.

To demonstrate the importance of subjectivity in QA, we develop a subjectivity-aware QA model that extends an existing QA model in a multi-task learning paradigm. It is trained to predict the subjectivity label and answer span simultaneously, and does not require subjectivity labels at test time. This QA model achieves 76.3% F_1 on an average over different domains of SUBJQA.

Contributions (i) We release a challenging QA dataset with subjectivity labels for questions and answers, spanning 6 domains; (ii) We investigate the relationship between subjectivity and QA; (iii) We develop a subjectivity-aware QA model; (iv) We verify the findings of previous work on subjectivity, using recent NLP architectures.

2 Subjectivity

Written text, as an expression of language, contains information on several linguistic levels, many of which have been thoroughly explored in NLP.[2] For instance, both the semantic content of text and the (surface) forms of words and sentences, as expressed through syntax and morphology, have been at the core of the field for decades. However, another level of information can be found when trying to observe or encode the so-called *private states* of the writer (Quirk et al., 1985). Examples of private states include the opinions and beliefs of a writer, and can concretely be said to not be avail-
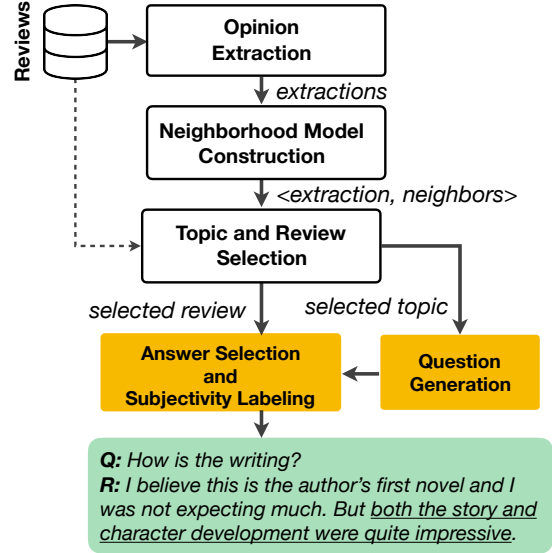

Figure 1: Our data collection pipeline

able for verification or objective observation. It is this type of state which is referred to as *subjectivity* (Banfield, 1982; Banea et al., 2011).

Whereas subjectivity has been investigated in isolation, it can be argued that subjectivity is only meaningful given sufficient context. Regardless, much previous work has focused on annotating words (Heise, 2001), word senses (Durkin and Manning, 1989; Wiebe and Mihalcea, 2006; Akkaya et al., 2009), or sentences (Pang and Lee, 2004), with notable exceptions such as Wiebe et al. (2005); Banea et al. (2010), who investigate subjectivity in phrases in the context of a text or conversation. There is limited work that studies subjectivity using architectures that allow for contexts to be incorporated efficiently (Vaswani et al., 2017).

As subjectivity relies heavily on context, and we have access to methods which can encode such context, what then of access to data which encodes subjectivity? We argue that in order to fully investigate research questions dealing with subjectivity in contexts, a large-scale dataset is needed. We choose to frame this as a QA dataset, as it not only offers the potential to investigate interactions in a single contiguous document, but also allows interactions *between* contexts, where parts may be subjective and other parts may be objective. Concretely, one might seek to investigate the interactions between an objective question and a subjective answer, or vice-versa.

3 Data Collection

We found two limitations of existing datasets and collection strategies that motivated us to create a new QA dataset to understand subjectivity in QA. First, data collection methods (Gupta et al., 2019; Xu et al., 2019b) often rely on the linguistic similarity between the questions and the reviews (e.g. information retrieval). However, subjective questions may not always use the same words or phrases as the review. Consider the examples below. The answer span '*vegan dishes*' is similar to the question Q_1. The answer to the more subjective question Q_2, however, is linguistically quite different.

Example 1 Q_1: Is the restaurant vegan friendly?
Review: [...] many **vegan dishes** on its menu.

Q_2: Does the restaurant have a romantic vibe?
Review: Amazing selection of wines, **perfect for a date night**.

Secondly, existing review-based datasets are small and not very diverse in terms of question topics and types (Xu et al., 2019a; Gupta et al., 2019). We, therefore, consider reviews about both products and services from 6 different domains, namely TripAdvisor, Restaurants, Movies, Books, Electronics and Grocery. We use the data of Wang et al. (2010) for TripAdvisor, and Yelp[3] data for Restaurants. We use the subsets for which an open-source opinion extractor was available (Li et al., 2019). We use the data of McAuley and Yang (2016b) that contains reviews from product pages of Amazon.com spanning multiple categories. We target categories that had more opinion expressions than others, determined by an opinion extractor.

Figure 1 depicts our data collection pipeline which builds upon recent developments in opinion extraction and matrix factorization. An opinion extractor is crucial to identify subjective or opinionated expressions, which other IR-based methods cannot. On the other hand, matrix factorization helps identify which of these expressions are related based on their co-occurrence in the review corpora, instead of their linguistic similarities. Relying on opinions instead of factual information to construct the dataset makes it inherently subjective. Furthermore, pairing questions and reviews based on related opinion expressions provides us some control over the diversity in the dataset. To the best of our knowledge, we are the first to explore such a method to construct a subjective QA dataset.

---

[3] https://www.yelp.com/dataset

Given a review corpus, we extract opinions about various aspects of the items being reviewed (*Opinion Extraction*). Consider the following review snippets and extractions.

Example 2 Review: [...] character development was quite impressive.
e_1:‹'*impressive*', '*character development*'›
Review: 3 stars for good power and good writing.
e_2:‹'*good*', '*writing*'›

In the next (*Neighborhood Model Construction*) step, we characterize the items being reviewed and their subjective extractions using latent features between two items. In particular, we use matrix factorization techniques (Riedel et al., 2013) to construct a neighborhood model N via a set of weights $w_{e,e'}$, where each corresponds to a directed association strength between extraction e and e'. For instance, e_1 and e_2 in Example 2 could have a similarity score 0.93. This neighborhood model forms the core of data collection. We select a subset of extractions from N as *topics* (*Topic Selection*) and ask crowd workers to translate them to natural language questions (*Question Generation*). For each topic, a subset of its neighbors from N and reviews which mention them are selected (*Review Selection*). In this manner, question-review pairs are generated based on the neighborhood model.

Finally, we present each question-review pair to crowdworkers who highlight an answer span in the review. Additionally, they provide subjectivity scores for both the questions and the answer span.

3.1 Opinion Extraction

An opinion extractor processes all reviews and finds extractions ‹X,Y› where X represents an opinion expressed on aspect Y. Table 1 shows sample extractions from different domains. We use OpineDB (Li et al., 2019), a state-of-the-art opinion extractor, for restaurants and hotels. For other domains where OpineDB was not available, we use the syntactic extraction patterns of Abbasi Moghaddam (2013).

3.2 Neighborhood Model Construction

We rely on matrix factorization to learn dense representations for items and extractions, and identify similar extractions. Such a method has been shown to be effective in building knowledge bases (Bhutani et al., 2020). As depicted in Figure 2, we organize the extractions into a matrix M where each row i corresponds to an item being

Domain	Opinion Span	Aspect Span
Restaurants	huge	lineup
Hospitality	no free	wifi
Books	hilarious	book
Movies	not believable	characters
Electronics	impressive	sound
Grocery	high	sodium level

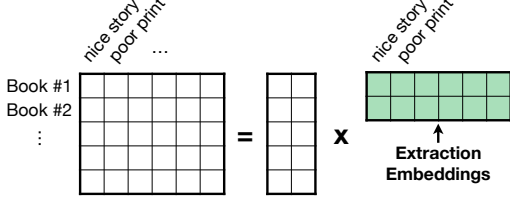Table 1: Example extractions from different domains

Figure 2: Learning representations of extractions via non-negative matrix factorization

reviewed and each column j corresponds to an extraction. The value M_{ij} denotes the frequency of extraction e_j in reviews of item x_i. Given M and a latent feature model F, we obtain extraction embeddings using non-negative matrix factorization. Concretely, each value M_{ij} is obtained from the dot product of two extractions of size K^F:

$$M_{ij}^F = \sum_k^{K^F} x_{i,k} e_{j,k}. \qquad (1)$$

For each extraction, we find its *neighbors* based on the cosine similarity of their embeddings.[4]

3.3 Topic and Review Selection

We next identify a subset of extractions to be used as *topics* for the questions. In order to maximize the diversity and difficulty in the dataset, we use the following criteria developed iteratively based on manual inspection followed by user experiments.

1. **Cosine Similarity:** We prune neighbors of an extraction which have low cosine similarity (< 0.8). Irrelevant neighbors can lead to noisy topic-review pairs which would be marked *non-answerable* by the annotators.

2. **Semantic Similarity:** We prune neighbors that are linguistically similar (> 0.975 similarity)[5] as they yield easy topic-review pairs.

3. **Diversity**: To promote diversity in topics and reviews, we select extractions which have many (> 5) neighbors.

---

[4] Details about hyper-parameters are included in the Appendix.
[5] Using GloVe embeddings provided by spaCy.

4. **Frequency**: To ensure selected topics are also popular, we select a topic if: a) its frequency is higher than the median frequency of all extractions, and b) it has at least one neighbor that is more frequent than the topic itself.

We pair each topic with reviews that mention one of its neighbors. The key benefit of a factorization-based method is that it is not only based on linguistic similarity, and forces a QA system to understand subjectivity in questions and reviews.

3.4 Question Generation

Each selected topic is presented to a human annotator together with a review that mentions that topic. We ask the annotator to write a question about the topic that can be answered by the review. For example, ‹‘good’, ‘writing’› could be translated to "Is the writing any good?" or "How is the writing?".

3.5 Answer-Span and Subjectivity Labeling

Lastly, we present each question and its corresponding review to human annotators (crowdworkers), who provides a subjectivity score to the question on a 1 to 5 scale based on whether it seeks an opinion (e.g., "How good is this book?") or factual information (e.g., "is this a hard-cover?"). Additionally, we ask them to highlight the shortest answer span in the review or mark the question as unanswerable. They also provide subjectivity scores for the answer spans. We provide details of our neighborhood model construction and crowdsourcing experiments in the Appendix.

4 Dataset Analysis

In this section, we analyze the questions and answers to understand the properties of our SUBJQA dataset. We present the dataset statistics in Section 4.1. We then analyze the diversity and difficulty of the questions. We also discuss the distributions of subjectivity and answerability in our dataset. Additionally, we manually inspect 100 randomly chosen questions from the development set in Section 4.3 to understand the challenges posed by subjectivity of the questions and/or the answers.

4.1 Data Statistics

Table 2 summarizes the number of examples we collected for different domains. To generate the train, development, and test splits, we partition the topics into training (80%), dev (10%) and test (10%) sets. We partition the questions and reviews based on the partitioning of the topics.

Domain	Train	Dev	Test	Total
TripAdvisor	1165	230	512	1686
Restaurants	1400	267	266	1683
Movies	1369	261	291	1677
Books	1314	256	345	1668
Electronics	1295	255	358	1659
Grocery	1124	218	591	1725

Table 2: No. of examples in each domain split.

Domain	Review len	Q len	A len	% answerable
TripAdvisor	187.25	5.66	6.71	78.17
Restaurants	185.40	5.44	6.67	60.72
Movies	331.56	5.59	7.32	55.69
Books	285.47	5.78	7.78	52.99
Electronics	249.44	5.56	6.98	58.89
Grocery	164.75	5.44	7.25	64.69

Table 3: Domain statistics. Len denotes n tokens.

4.2 Difficulty and Diversity of Questions

Table 3 shows that reviews in different domains tend to vary in length. Answer spans tend to be 6-7 tokens long, compared to 2-3 tokens in SQuAD.

Table 4 shows the number of distinct questions and topics in each domain. On average we collected 1500 questions covering 225 aspects. We also automatically categorize the boolean questions based on a lexicon of question prefixes. Unlike other review-based QA datasets (Gupta et al., 2019), SUBJQA contains more diverse questions, the majority of which are not yes/no questions. The questions in SUBJQA are also linguistically varied, as indicated by the trigram prefixes of the questions (Figure 3). Most of the frequent trigram prefixes in SUBJQA (e.g., *how is the*, *how was the*, *how do you*) are almost missing in SQuAD and Gupta et al. (2019). The diversity of questions in SUBJQA demonstrate challenges unique to the dataset.

4.3 Data Quality Assessment

We randomly sample 100 answerable questions to manually categorize them according to their reasoning types. Table 5 shows the distribution of the

Domain	# questions	# aspects	% boolean Q
TripAdvisor	1411	171	16.13
Restaurants	1553	238	17.29
Movies	1556	228	15.56
Books	1517	231	16.90
Electronics	1535	314	14.94
Grocery	1333	163	14.78

Table 4: Diversity of questions and topics

Figure 3: The distribution of prefixes of questions. The outermost ring shows unigram prefixes (e.g., 57.9% questions start with *how*). The middle and innermost rings correspond to bigrams and trigrams, respectively.

reasoning types and representative examples. As expected, since a large fraction of the questions are subjective, they cannot be simply answered using a keyword-search over the reviews or by paraphrasing the input question. Answering such questions requires a much deeper understanding of the reviews. Since the labels are crowdsourced, a small fraction of the answer spans are noisy.

We also categorized the answers based on answer-types. We observed that 64% of the answer spans were independent clauses (e.g., *the staff was very helpful and friendly*), 25% were noun phrases (e.g., *great bed*) and 11% were incomplete clauses/spans (e.g., *so much action*). This supports our argument that subjective questions often cannot be answered simply by an adjective or noun phrase.

We rely on an automatically constructed opinion KB and collect labels for answer spans and subjectivity for all the question-review pairs. As crowdworkers would label any spurious question-review pair as 'unanswerable', this might increase the number of negative examples in the dataset. However, such examples are both much more difficult and more valuable than randomly paired questions and reviews. We found that the KBs used in our dataset collection achieved 35-50% mean precision on various domains. We also found that approximately 48% of a random set of 50 question-review pairs from SubjQA were marked unanswerable because of the unrelated opinion pairs in the KB.

Reasoning	Percent.	Example
Lexical	18%	Q: How small was the hotel bathroom? R: ...**Bathroom on the small** side with older fixtures...
Paraphrase	28%	Q: How amazing was the end? R: ...The **ending was absolutely awesome**, it makes the experience not so ...
Indirect	43%	Q: How was the plot of the movie? R: ...simply because there's so much going on, **so much action**, so many complex ..
N/A	11%	Q: How do you like the episode? R: **For a show that I think was broadcast in HighDef**, it seems impossible that the...

Table 5: Types of reasoning required for the various domains. N/A indicates noisy or incorrect spans.

4.4 Answerability and Subjectivity

The dataset construction relies on a neighborhood model generated automatically using factorization. It captures co-occurrence signals instead of linguistic signals. Consequently, the dataset generated is not guaranteed to only contain answerable questions. As expected, about 65% of the questions in the dataset are answerable from the reviews (see Table 7). However, unlike Gupta et al. (2019), we do not predict answerability using a classifier. The answerability labels are provided by the crowd-workers instead, and are therefore more reliable.

Table 7 shows the subjectivity distribution in questions and answer spans across different domains. A vast majority of the questions we collected are subjective, which is not surprising since we selected topics from opinion extractions. A large fraction of the subjective questions (~70%) were also answerable from their reviews.

We also compare the subjectivity of questions with the subjectivity of answers. As can be seen in Table 6, the subjectivity of an answer is strongly correlated with the subjectivity of the question. Subjective questions often have answers that are also subjective. Similarly, factual questions, with few exceptions, have factual answers. This indicates that a QA system must understand how subjectivity is expressed in a question to correctly find its answer. Most domains have 75% subjective questions on average. However, the BERT-QA model fine-tuned on each domain achieves 80% F1 on subjective questions in movies and books, but only achieves 67-73% F1 on subjective questions in grocery and electronics. Future QA systems for user-generated content, such as for customer support, should therefore model subjectivity explicitly.

	subj. Q	fact. Q
subj. A	79.8%	1.31%
fact. A	1.29%	17.58%

Table 6: Subjectivity distribution in SUBJQA.

Domain	% subj. Q	% answerable	% subj. A
TripAdvisor	74.49	83.20	75.20
Restaurants	76.11	65.72	76.29
Movies	74.41	62.09	74.59
Books	75.77	58.86	75.35
Electronics	69.80	65.37	69.98
Grocery	73.21	70.22	73.15

Table 7: Statistics on subjective Q, answerability, and subjective A per domain in SUBJQA.

4.5 Comparison with other datasets

Although this is not the first work to design a review-based QA dataset, no prior QA dataset targets the understanding of subjectivity in reviews. AmazonQA (Gupta et al., 2019) is one of the largest review-based QA dataset, but differs from SubjQA in three core ways: i) they lack subjectivity labels; ii) they are constructed using retrieval methods; and iii) they do not highlight answer spans in the reviews and instead provide an answerability label for the question given a set of reviews. These differences make it difficult for a fair comparison with their dataset. Some of the distinguishing characteristics of the two datasets include: (a) AmazonQA reports 61% of the questions are answerable with 72% precision (based on a classifier). In comparison, 65% of the questions in SubjQA are answerable (based on human labels); (b) We provide opinion annotations and subjectivity labels for the questions, making it easier for researchers to target specific opinions and subjective text; and (c) 58% of questions in SubjQA start with 'how'. In comparison, most of the question prefixes (e.g., 'will

	IMDB	SUBJQA
Word-based (TEXTBLOB)	61.90	57.50
BERT fine-tuned	88.20	62.77

Table 8: Subjectivity prediction accuracies on IMDB data (Pang and Lee, 2004) and our dataset (SUBJQA).

this work', 'does it come with' etc.) in AmazonQA indicate product queries rather than opinion-based queries.

5 Subjectivity Modelling

We now turn to experiments on subjectivity, first investigating claims made by previous work, and whether they still hold when using recently developed architectures, before investigating how to model subjectivity in QA.

5.1 Subjectivity in Sentiment Analysis

Pang and Lee (2004) have shown that subjectivity is an important feature for sentiment analysis. Sorting sentences by their estimated subjectivity scores, and only using the top n such sentences, allows for a more efficient and better-performing sentiment analysis system, than when considering both subjective and objective sentences equally. We first investigate whether the same findings hold true when subjectivity is estimated using transformer-based architectures.[6] We predict the subjectivity of a sentence by passing its representation through a feed-forward neural network. We compare this with using subjectivity scores of TEXTBLOB[7], a sentiment lexicon-based method, as a baseline.

We evaluate this setup on subjectivity data from Pang and Lee (2004)[8] and the subjectivity labels made available in SUBJQA. Unsurprisingly, a contextually-aware classifier vastly outperforms a word-based classifier, highlighting the importance of context in subjectivity analysis (see Table 8). Interestingly, however, predicting subjectivity in SUBJQA is more challenging than in IMDB - we hypothesise that this is because SUBJQA spans multiple domains.

We further investigate if our subjectivity classifier is beneficial for sentiment analysis. We replicate Pang and Lee (2004), and perform sentiment analysis using the top N subjective and objective

---

[6] Using the BERT-base uncased model.
[7] https://textblob.readthedocs.io/
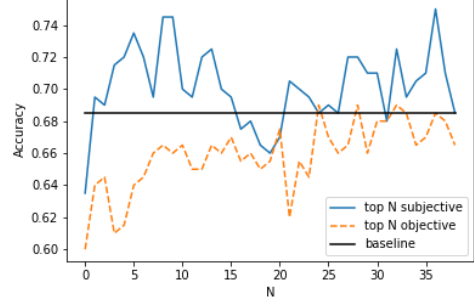[8] http://www.cs.cornell.edu/people/pabo/movie-review-data/

Figure 4: Sentiment Analysis accuracy using top N subj. sentences (blue), top N fact. sentences (orange dashed), compared to the all sentences baseline (black).

sentences based on our system. Figure 4 shows that giving a contextually-aware sentiment analysis model access to N subjective sentences improves performance, as compared to using all sentences, or using N objective sentences.

5.2 Subjectivity-Aware QA Model

Given the importance of subjectivity in other NLP tasks, we investigate whether it is also an important feature for QA in SUBJQA, with a relatively simple model. We approach this by implementing a subjectivity-aware QA model, as an extension of our baseline models FastQA (Weissenborn et al., 2017) in a multitask learning (MTL) paradigm (Caruana, 1997), as this approach has been shown to be useful for learning cross-task representations (Bjerva, 2017a,b; Bingel and Søgaard, 2017; de Lhoneux et al., 2018; Augenstein et al., 2018, 2019; Ruder et al., 2019). One concrete advantage of using MTL is that we do not need to have access to subjectivity labels at test time, as would be the case if we required subjectivity labels as a feature for each answer span. Each input paragraph is encoded with a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) over a sequence of word embeddings and contextual features ($\tilde{X}$). This encoding, H', is passed through a hidden layer and a non-linearity:

$$H' = \text{Bi-LSTM}(\tilde{X}) \qquad (2)$$

$$H = \tanh(BH'^{\top}) \qquad (3)$$

We extend this implementation by adding two hidden layers of task-specific parameters (W^n) associated with a second learning objective:

$$S' = \text{ReLU}(W^1 H) \qquad (4)$$
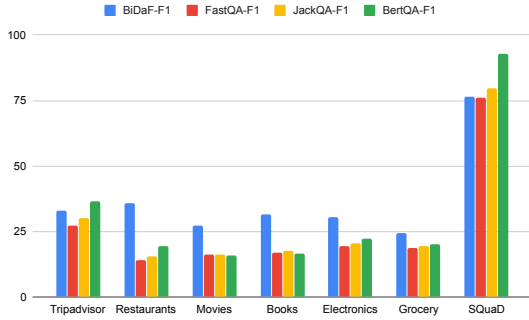
$$S = \text{softmax}(W^2 S') \qquad (5)$$

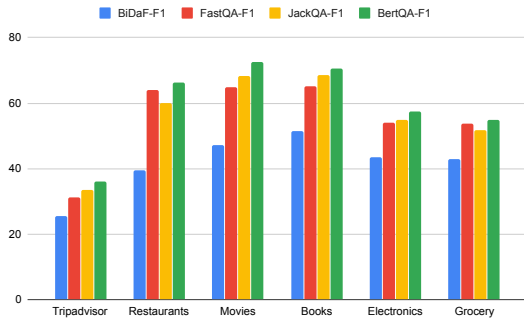Figure 5: F1 scores of pre-trained out-of-the-box models on different domains in SUBJQA.

Figure 6: Gain in F1 with models fine-tuned on different domains over the pre-trained model.

In training, we randomly sample between the two tasks (QA and Subjectivity classification).

5.3 Baselines

We use four pre-trained models to investigate how their performances on SUBJQA compare with a factual dataset, SQuAD (Rajpurkar et al., 2016), created using Wikipedia. Specifically, we evaluate BiDaF (Seo et al., 2017), FastQA (Weissenborn et al., 2017), JackQA (Weissenborn et al., 2018)[9] and BERT (Devlin et al., 2018),[10] all pre-trained on SQuAD. Additionally, we fine tune the models on each domain in SUBJQA.

Figure 5 shows the F1 scores of the pre-trained models. We report the Exact match scores in Appendix A.3. Pre-trained models achieve F1 scores as high as 92.9% on the SQuAD. On the other hand, the best model achieves an average F1 of 30.5% across all domains and 36.5% F1 at best on any given domain in SUBJQA. The difference in performance can be attributed to both differences in domain (Wikipedia vs. customer reviews) and how subjectivity is expressed across different domains.

---

[9] https://github.com/uclnlp/jack
[10] BERT-Large, Cased (Whole Word Masking)

Figure 6 shows the absolute gains in F1 scores of models fine-tuned on specific domains, over the pre-trained model. After fine-tuning on each domain, the best model achieves an average F1 of 74.1% across the different domains. While fine-tuning significantly boosts the F1 scores in each domain, they are still lower than the F1 scores on the SQuAD dataset. We argue that this is because the models are agnostic about subjective expressions in questions and reviews. To validate our hypothesis, we compare the gain in F1 scores of the BERT model on subjective questions and factual questions. We find that the difference in F1 gains is as high as 23.4% between factual and subjective questions. F1 gains differ by as much as 23.0% for factual vs. subjective answers.

5.4 Subjectivity-Aware Modeling

After fine-tuning over each domain in the MTL setting, the subjectivity-aware model achieves an average F1 of 76.3% across domains (Table 9). Incorporating subjectivity in the model thus boosts performance across all domains and both metrics. Although there are gains also for subjective questions and answers, it is noteworthy that the highest gains can be found for factual questions and answers. This is likely because existing techniques already are tuned for factual questions.

5.5 Error Analysis

We perform an error analysis based on the predictions of the BERT model fine-tuned for each domain, investigating a sample of 10 questions for which the model's predictions are incorrect. This reveals that the model often predicts a question to be unanswerable, although a ground truth answer exists. Interestingly, most of these questions are subjective. A detailed examination reveals that answers to these questions are relatively indirect, hence it is not surprising that the model struggles. For instance, given the question *"How slow is the internet service?"*, the correct answer is *"Don't expect to always get the 150Mbps"*.

6 Related Work

We are witnessing a substantial rise in user-generated content, including subjective information ranging from personal experiences to opinions about specific aspects of a product. However, subjectivity has largely been studied in the context of sentiment analysis (Hu and Liu, 2004) and opinion

	Fact. A		Subj. A		Fact. Q		Subj. Q		Overall	
	F1	E	F1	E	F1	E	F1	E	F1	E
Tripadvisor	17.50	20.88	1.28	7.43	18.85	21.60	1.16	7.37	1.01	7.42
Restaurants	10.36	12.38	8.37	11.49	13.85	15.77	8.19	11.07	5.71	8.65
Movies	14.49	14.63	5.17	8.02	14.27	14.41	5.44	8.28	3.08	5.84
Books	13.95	14.10	7.18	9.82	14.68	14.83	7.05	9.68	4.06	6.67
Electronics	14.15	18.70	0.28	7.06	13.29	18.22	0.40	7.24	-0.01	7.26
Grocery	9.69	11.74	-0.16	3.75	10.71	12.32	-0.48	3.41	-1.57	2.20
Average	13.35	15.40	3.69	7.93	**14.27**	**16.19**	3.63	7.84	2.05	6.34

Table 9: MTL gains/losses over the fine-tuning condition (F1 and Exact match), across subj./fact. QA.

mining (Blair-Goldensohn et al., 2008), with a focus on text polarity. There is renewed intereste in incorporating subjective opinion data into general data management systems (Li et al., 2019; Kobren et al., 2019) and for querying subjective data.

In this work, we revisit subjectivity in the context of review QA. Yu et al. (2012); McAuley and Yang (2016b) also use review data, as they leverage question types and aspects to answer questions. However, no prior work has modeled subjectivity explicitly using end-to-end architectures.

Furthermore, none of the existing review-based QA datasets are targeted at understanding subjectivity. This can be attributed to how these datasets are constructed. Large-scale QA datasets, such as SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), CoQA (Reddy et al., 2019), MLQA (Lewis et al., 2020) are based on factual data. We are the first to attempt to create a review-based QA dataset for the purpose of understanding subjectivity. Recent work has corroborated our findings on the benefits of modelling subjectivity QA, and highlights the differences in the distributions of hidden representation between SUBJQA and the factual SQuAD data (Muttenthaler et al., 2020).

7 Conclusion

In this paper we investigate subjectivity in QA by leveraging end-to-end architectures. We release SUBJQA, a question-answering dataset which contains subjectivity labels for both questions and answers. The dataset allows i) evaluation and development of architectures for subjective content, and ii) investigation of subjectivity and its interactions in broad and diverse contexts. We further implement a subjectivity-aware model and evaluate it, along with 4 strong baseline models. We hope this dataset opens new avenues for research on querying subjective content, and into subjectivity in general.

References

Samaneh Abbasi Moghaddam. 2013. *Aspect-based opinion mining in online reviews*. Ph.D. thesis, Applied Sciences: School of Computing Science.

Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard. 2019. X-WikiRE: A large, multilingual resource for relation extraction as machine comprehension. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 265–274, Hong Kong, China. Association for Computational Linguistics.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 190–199.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *EMNLP*. Association for Computational Linguistics.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces. 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 1896–1906. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd international conference on computational linguistics*, pages 28–36. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2011. Multilingual sentiment and subjectivity analysis. *Multilingual natural language processing*, 6:1–19.

Ann Banfield. 1982. *Unspeakable Sentences: The sentence representing non-reflective consciousness and the absence of the narrator*. Routledge.

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Nikita Bhutani, Aaron Traylor, Chen Chen, Xiaolan Wang, Behzad Golshan, and Wang-Chiew Tan. 2020. *Sampo: Unsupervised Knowledge Base Construction for Opinions and Implications*.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Johannes Bjerva. 2017a. *One Model to Rule them all – Multitask and Multilingual Modelling for Lexical Analysis*. Ph.D. thesis, University of Groningen.

Johannes Bjerva. 2017b. Will my auxiliary tagging task help? estimating auxiliary tasks effectivity in multi-task learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 216–220, Gothenburg, Sweden. Association for Computational Linguistics.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28 (1):41–75.

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

Kevin Durkin and Jocelyn Manning. 1989. Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18(6):577–612.

Miao Fan, Chao Feng, Mingming Sun, Ping Li, and Haifeng Wang. 2019. Reading customer reviews to answer product-related questions. In *Proceedings of the 2019 SIAM International Conference on Data Mining, SDM 2019, Calgary, Alberta, Canada, May 2-4, 2019*, pages 567–575.

Quentin Grail and Julien Perez. 2018. ReviewQA: a relational aspect-based opinion reading dataset. *CoRR*, abs/1810.12196.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4996–5002.

David R Heise. 2001. Project Magellan: Collecting cross-cultural affective meanings via the internet. *Electronic Journal of Sociology*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611.

Ari Kobren, Pablo Bario, Oksana Yakhnenko, Johann Hibschman, and Ian Langmore. 2019. Constructing High Precision Knowledge Bases with Subjective and Factual Attributes. *arXiv preprint arXiv:1905.12807*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *ACL 2020, arXiv preprint arXiv:1910.07475*.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997, Brussels, Belgium. Association for Computational Linguistics.

Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Y. Halevy, Vivian Li, and Wang-Chiew Tan. 2019. Subjective Databases. *PVLDB*, 12(11):1330–1343.

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

Julian J. McAuley and Alex Yang. 2016b. 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.

Lukas Muttenthaler, Isabelle Augenstein, and Johannes Bjerva. 2020. Unsupervised Evaluation for Question Answering with Transformers. In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278.

Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowl.-Based Syst.*, 108:42–49.

R. Quirk, Greenbaum S., Leech G., and Svartvik J. 1985. *A Comprehensive Grammar of the English Language*. Longman, New York.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners.

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.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *TACL*, 7:249–266.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 74–84.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent Multi-task Architecture Learning. In *In Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 783–792.

Dirk Weissenborn, Pasquale Minervini, Isabelle Augenstein, Johannes Welbl, Tim Rocktäschel, Matko Bosnjak, Jeff Mitchell, Thomas Demeester, Tim Dettmers, Pontus Stenetorp, and Sebastian Riedel. 2018. Jack the Reader - A Machine Reading Framework. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 25–30.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. FastQA: A Simple and Efficient Neural Architecture for Question Answering. *CoRR*, abs/1703.04816.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019a. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *NAACL-HLT (1)*, pages 2324–2335. Association for Computational Linguistics.

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

Jianxing Yu, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Answering Opinion Questions on Products by Exploiting Hierarchical Organization of Consumer Reviews. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 391–401.

A Appendices

A.1 Neighborhood Model Construction

For constructing the matrix for factorization, we focus on frequently reviewed items and frequent extractions. In particular, we consider items which have more than 10,000 reviews and extractions that were expressed in more than 5000 reviews. Once the matrix is constructed, we factorize it using non-negative factorization method using 20 as the dimension of the extraction embedding vector.

In the next step, we construct the neighborhood model by finding top-10 neighbors for each extraction based on cosine similarity of the extraction and the neighbor. We further select topics from the extractions, and prune the neighbors based on the criteria we described earlier. Table 10 shows example extraction and their neighbors discovered using the neighborhood model.

Domain	Extraction	Neighbor
Movies	(satisfying, ending)	(good, script)
	(believable, acting)	(interesting, movie)
Electronics	(responsive, key)	(nice, keyboard)
	(easy to navigate, menu)	(simple, interface)
Books	(graphic, violence)	(disturbing, book)
	(interesting, twist)	(unpredictable, story)
Grocery	(healthy, snack)	(high, protein)
	(easy to follow, direction)	(quick, preparation)
Tripadvisor	(excellent, service)	(amazing, hotel)
	(good, drinks)	(great, bar)
Restaurants	(great, meal)	(good, restaurant)
	(fast and friendly, service)	(quick, food)

Table 10: Examples extraction and their neighbors

A.2 Additional Dataset Examples

Table 11 shows more examples of question and their review snippets. Answer spans have been underlined. As can be seen, many questions and answer spans in our dataset are subjective.

A.3 Additional Experimental Results

Figure 7 shows the exact scores achieved by the pre-trained out-of-the-box models on various domains in SUBJQA. Figure 8 shows the exact scores of the models fine-tuned on each domain in SUBJQA.
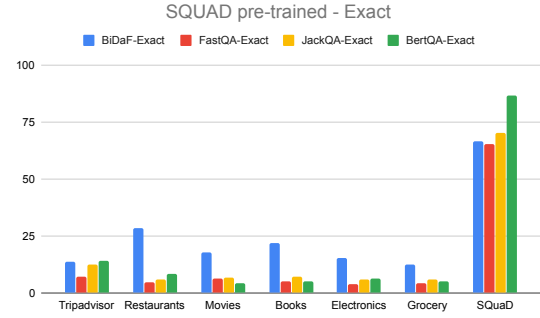

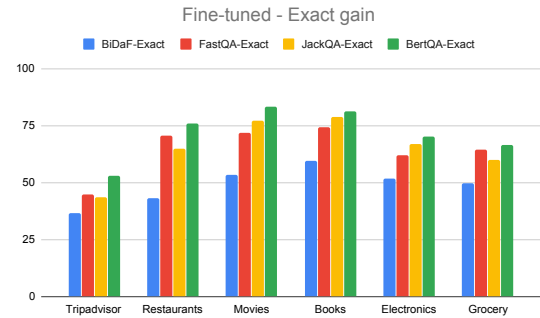Figure 7: Exact scores of pre-trained out-of-the-box models on different domains.

Figure 8: Gain in Exact scores with models fine-tuned on different domains.

A.4 Crowdsourcing Details

We used the Appen platform [11] to obtain labels. The platform provides quality control by showing the workers 5 questions at a time, out of which one is labeled by the experts. A worker who fails to maintain 70% accuracy is kicked out by the platform and his judgements are ignored.

Figure 9 illustrates the instructions that were shown to the crowdworkers for the question generation task. Figure 10 shows the interface for the answer-span collection and subjectivity labeling tasks. The workers assign subjectivity scores (1-5) to each question and the selected answer span. They can also indicate if a question cannot be answered from the given review. Before finalizing this interface, we experimented with multiple design variations on a small subset. This included collecting binary labels for subjectivity instead of scores, simultaneously collecting different labels vs in different tasks etc. To ensure good quality labels, we each worker 5 cents per annotation.

---

[11] https://appen.com

Domain	Question	Review
Movies	Is the main character a good actor?	Beautifully written series depicting the lives of criminals in Kentucky and thedeputies in the US Marshall Service who attempt to stop the criminal elementin the community.The acting is outstanding and the <u>cast bring the characters to life.</u>
	Can we enjoy the movie along with family?	<u>An outstanding romantic comedy</u>, 13 Going on 30, brings to the screen exactly what the title implies: the story of a 13-year old girl who has her wish fulfilled and wakes up seven years later in the body of her 30-year old self!
Electronics	Why is the camera of poor quality?	Item like the picture, fast deliver 3 days well packed, good quality for the price. The camera is decent (as phone cameras go), <u>There is no flash</u> though.
	How big is the unit?	It's a great product, especially for the money.Good Battery life, totally useable for a full movie. Storage capacity. <u>32GB</u> is a great point for the price. Speakers. Screen - bright, clear and HD resolution.
Books	Is the plot line good enough?	<u>The book got me hooked almost immediately</u> and then I got to the end and realized that there is another book after this one. I was unaware of this dilemma but its so good I did not care. Characters and dialogue are good but I liked the movie better.
	What is the most exciting part of the story?	Yann Martel's Life of Pi is a wondrous novel–there is much to wonder and marvel at. The story is simple, yet complex at the same time and can be read on many levels. <u>The novel ends with a philisophical bang, which did blow me away.</u>
Grocery	Does this coffee taste good?	While I don't consider myself a coffee snob, I know what I like. Actually, <u>it tasted on the light side of light</u>. I prefer a dark bold flavor but don't mind a good medium roast. The package also said Exotic Floral and Berry.
	Is the sauce tasty?	I was hoping that this sauce would be a little more consistent and thick that it is. <u>The taste is a bit sharp and perhaps it's just not in my palette</u>, but I'll stick with a homeade ranchero until I find one that is a quick retail replacement.
Tripadvisor	How was the attention of the staff?	The Handlery Union Square Hotel offers great rooms in the centre of San Francisco for a very nice price! Excellent value for money. <u>Friendly personnel</u> and top location! Highly recommended!
	How is the parking?	$40.00 a night; not the room - parking rate per night.Got a room with no view at the end of the hall way, nasty smell like very old, damp room. Front desk clark was not friendly, when raised concern about parking rate, he compared that with NYC parking. The <u>parking rate is ridiculously high</u>; $40 a night!!!!!!!!!!!
Restaurants	Is the price economical?	This place is the best Pho Place in the area. If you are too lazy to drive to Pho Dau Bo then come here for your hangover cure. <u>The prices are a little bit on the high side</u> but that is simply a reflection of the neighbourbood the restaurant is located in.
	Was it a good place for late night snacking?	<u>Lots of different beer choices and liter-ful mugs</u> that will satiate your European beer craving. Social seating on long benches (no individual tables) filters out the snobs and the social invalids, so everyone has a great time!

Table 11: Examples questions and review snippets. Answer spans are underlined.

Overview

You will be shown a review about an item (a hotel, a product or ...) that discusses a particular **aspect** of the item. Your task is to write a natural question <u>about the given aspect</u> such that the question can be answered (at least partially) with the review

---

Rules & Tips

(1) Make sure to write a natural and <u>a grammatically-correct</u> question.

The questions should make sense on their own. Questions like "what about the price?" would not be acceptable.

(2) If you fail a test question, and you think your question is valid please contest and we would review it.

(3) Make sure your question has the aspect in it.

If the aspect is "wine tasting", then "wine tasting" should appear in the question.

(4) Avoid the following simple template: how is <<aspect>>? or how is the <<aspect>>?

For example, "how is view?" or "how is the service?" would be invalid answers.

These questions may be valid, but they would not be considered as acceptable answers.

(5) Your question should start with a question word (i.e., how, when, is, where, ...) and it should end with <u>question</u> mark --> <u>?</u>

(6) Please only write questions in English.

(7) You are encouraged to write questions in your own words without copying phrases from the review. Avoid simply rephrasing the review into a question.

For example, converting "the service is great" to "is the service great?" would not be acceptable.

Figure 9: The instructions shown to crowdworkers for the question writing task.

Figure 10: The interface for the answer-span collection and subjectivity labeling tasks.