Similarity or deeper understanding? Analyzing the TED-Q dataset of evoked questions

Matthijs Westera
Universitat Pompeu Fabra / Leiden University
matthijs.westera@gmail.com

Jacopo Amidei
The Open University
jacopo.amidei@open.ac.uk

Laia Mayol
Universitat Pompeu Fabra
laia.mayol@upf.edu

Abstract

We take a close look at a recent dataset of TED-talks annotated with the questions they implicitly evoke, TED-Q (Westera et al., 2020). We test to what extent the relation between a discourse and the questions it evokes is merely one of similarity or association, as opposed to deeper semantic/pragmatic interpretation. We do so by turning the TED-Q dataset into a binary classification task, constructing an analogous task from explicit questions we extract from the BookCorpus (Zhu et al., 2015), and fitting a BERT-based classifier alongside models based on different notions of similarity. The BERT-based classifier, achieving close to human performance, outperforms all similarity-based models, suggesting that there is more to identifying true evoked questions than plain similarity.

1 Introduction

Asking explicit questions is essential for resolving information gaps, maintaining common ground, and fostering a shared view of the discourse topics. Understanding *implicit* questions is crucial for the same reasons, and moreover underlies our perception of discourse coherence (e.g., Van Kuppevelt (1995)): a discourse is coherent if each utterance addresses a question implicitly evoked by what came before. Despite their theoretical significance and their potential relevance to many Natural Language Processing (NLP) tasks, the questions implicitly evoked by a discourse have not received much attention in NLP, where the typical questions sought after are, rather, comprehension-testing questions – although this has begun to change, see Section 2.

Recently a crowdsourced dataset became available aimed at making implicit questions explicit: the TED-Q dataset (Westera et al., 2020). We test to what extent the relation between a discourse and the questions it evokes is merely one if similarity or association, as opposed to deeper semantic/pragmatic interpretation. We do so by constructing a classification task from TED-Q where the goal is to predict, given a preceding discourse, whether the given question was evoked at that point or not. We compare human performance, a BERT-based classifier and four models based on different notions of similarity. In order to, moreover, investigate how the implicit questions of TED-Q compare to explicit questions from existing (unannotated) corpora, which are far easier to come by, we construct an analogous task from written dialogue in the existing BookCorpus (Zhu et al., 2015).[1]

2 Background

Questions have long been an important topic in NLP, but mostly in their role of comprehension-testing questions, i.e., questions that are *answered* by the discourse, rather than unanswered questions that it *evokes*, as in the current work. Recently this has begun to change. For example, in terms of available datasets, the QuAC ('Question Answering in Context') dataset (Choi et al., 2018) consists of 100K questions asked in unscripted dialogue about a given topic; Rao and Daumé III (2018) present a dataset of

[1] Scripts for creating the tasks, models and analyses are available at https://github.com/amore-upf/ted-q_eval. We scraped the BookCorpus from smashwords.com using the script at https://github.com/soskek/bookcorpus.

75K clarification questions collected from StackExchange; Riester (2019) attempts expert annotation of implicit and explicit questions in their role of structuring discourse (so-called Questions Under Discussion, see below); Pyatkin et al. (2020) crowdsource question-answer pairs as a more natural way of annotating discourse structure (QADiscourse); and Ko et al. (2020) present Inquisitive, a dataset of 19K questions that are evoked by news texts, crowdsourced using a method very similar to the TED-Q dataset on which we will rely (see Section 3 for an overview of TED-Q). On the modeling side, exceptions to the field's focus on comprehension questions are Rao and Daumé III (2019)'s model of clarification questions, Qi et al. (2020)'s informative question generation based on QuAC, and the approach in Schricker and Scheffler (2019) towards ranking questions evoked by the preceding discourse, a task which they construct from annotations in the style of Riester (2019).

In the more theoretical linguistics literature, the notion of an evoked question relates closely to the notion of Question Under Discussion or QUD (Von Stutterheim and Klein, 1989; Van Kuppevelt, 1995; Roberts, 1996). In QUD-based approaches, discourse is assumed to consists of a number of moves, where each move addresses a specific implicit or explicit QUD, and the QUDs are related and together form a structured whole, e.g., a tree with super-questions and sub-questions, where one discourse move addresses a question left open (or evoked) by another. Despite its centrality in theoretical semantics and pragmatics for describing discourse structure, the notion of QUD has not yet gained traction in computational linguistics and NLP. This is in part because of data scarcity: whereas discourse structure annotations in the relational approach have long been available (e.g., Mann and Thompson (1988); Webber et al. (2019)), QUD annotation has only recently started (Riester, 2019). Because evoked questions are potential QUDs for an upcoming discourse move (Onea, 2013; Westera et al., 2020), we hope that the current work on evoked questions can help empirically ground theoretical work on QUDs.

The type of research question that we address – whether a given phenomenon, in our case evoked questions, reflects patterns that are more sophisticated than mere similarity – has been considered in relation to many other phenomena in the literature, where similarity-based baselines commonly serve either model comparison or, as in the current paper, dataset evaluation.

3 Constructing the tasks

The TED-Q dataset consists of 6 TED talks, annotated throughout with the questions evoked by the talk up to a certain point. 111 crowd-sourced annotators read excerpts of the talks and entered free-form questions that were evoked by the excerpt up to that point (and not yet answered). This resulted in 2421 evoked questions, with on average 5 questions after every sentence. After entering a question the continuation of the excerpt was revealed, and annotators were asked to rate on a 5-point scale to what extent the question had been answered by the continuation (ANSWERED, used for analysis below).

We take each evoked question from TED-Q with its preceding context as a positive item, and combine the same context with a random nearby evoked question as a negative item. We chose our context size to be 3 sentences, guaranteeing that the context would include at least the context seen by the crowd worker of TED-Q who posed the question.[2] For the negative items we sampled evoked questions from between 3 and 4 sentences away in either direction, a range on which we converged through manual inspection, aimed at maximizing the challenge (nearby questions will be more similar) without making the task entirely impossible. This resulted in 4824 items, half of which positive.

We construct an analogous task from the BookCorpus, which consists of around 11000 books of fiction in plain text. We extracted 2 million multi-turn dialogues by extracting quotations, assuming that quotations within 200 characters of each other are part of a single dialogue, and that quotations on the same line constitute a single dialogue turn. For any turn that begins with a question and is preceded by a non-question, we took the question with its preceding context as a positive item, and the same context with a random nearby question as a negative item, using the same context size and negative question range as for TED-Q. This resulted in 3.8 million items, half of which are positive.

We divide each task into a train and test set (80%-20%) for the final evaluation. Contexts can overlap

[2] A smaller context generally makes the task easier, in the sense of yielding higher scores. Models trained on a smaller context tended to perform a lot worse when given the full (3 sentence) contexts, but not vice versa.

	Models trained on TED-Q:					
	LEMMAOVERLAP	GLEU	MEANCOS	BERTSCORE	ALLSIMS	BERT
TED-Q	0.47 (60%)	0.24 (32%)	0.14 (49%)	0.33 (30%)	0.46 (52%)	**0.55** (61%)

Table 1: Main results: MCC and percentage of errors that are false positives.

considerably (e.g., for TED-Q, every sentence has evoked questions, and every context has three sentences). We avoid overlap between train and test set by removing any context from the test set that has a neighboring (within 3 sentences) context in the training set. To minimize data loss we sample the train and test partitions not by single items, but by sampling contiguous subsets of items of 20 subsequent contexts each. We use Matthew's Correlation Coefficient (MCC) (Matthews, 1975) as our main metric (see Chicco and Jurman (2020) for arguments in favor of MCC in binary classification).

4 Models

To test whether TED-Q's evoked questions reflect patterns that are more sophisticated than mere similarity, we fit two kinds of models: random decision forests based on various notions of similarity, and the neural network transformer BERT (Devlin et al., 2019).[3] We use four similarity notions: (i) LEMMAOVERLAP, the proportion of question lemmata that are also found in the context (using the SpaCy lemmatizer; `www.spacy.io`); (ii) GLEU (Wu et al., 2016), a slightly more syntax-aware measure of surface form similarity based on matching n-grams (we use the sentences in the context as separate references); (iii) MEANCOS, mean cosine similarity between words in the question and words in the context (crossing all pairs), using GloVe vectors (Pennington et al., 2014); and (iv) BERTSCORE (Zhang et al., 2019), which measures the match between tokens of the question and the context based on their contextualized semantic similarity according to pre-trained BERT (we use its F_1-score, again using the sentences in the context as separate references). The first two concern word choice and some syntax; the latter two are more semantic/pragmatic. Lastly, ALLSIMS combines all four notions in a single random forest.

For each model we ran a hyperparameter grid search with 5-fold cross-validation on the training set. For the random forest we varied the number of estimators (100, 200, 300), maximum tree depth (1, 2, 4, 8, 16) and minimal items per leaf (1, 2, 4, 6, 8, 16). Observing only little difference (also between models), we chose 300 estimators with maximum depth 4, and minimally 8 items per leaf for all random forests. For BERT we varied the warm-up ratio (0.03 (default), 0.1), learning rate (4e-4, 4e-5 (default), 2e-5, 1e-5, 4e-6), and weight decay (0 (default), 1e-7, 1e-6, 1e-5), finding the best and most stable results for warm-up ratio 0.1, learning rate 2e-5 and zero weight decay.

5 Results and discussion

Main results Table 1 compares the six models on the TED-Q test set, showing MCC and percentage of errors that are false positives. The BERT classifier outperforms all similarity-based models as well as their combination ALLSIMS. With regard to our main research question, this suggests that there is more to identifying true evoked questions than plain similarity, or even a combination of similarity notions.

Analyzing BERT To analyze BERT's behavior, two of us independently completed a portion of the task, resulting in 195 items with human annotations, reaching MCC scores which are similar to BERT's: 0.55 and 0.60. Annotation reliability by unweighted Cohen's κ (Cohen, 1960) reached 0.66, or a substantial level of agreement according to Landis and Koch (1977).[4] On these 195 items, BERT makes 32 errors (21 false positives; 11 false negatives), 12 of which were also incorrectly annotated by both annotators (either because they fit but they were actually non-evoked in TED-Q, or because, despite being actually evoked in TED-Q, they were somewhat vague). Of the 21 false positives, 10 were correctly

[3] We use Python's Scikit-learn (Pedregosa et al., 2011) for the random forests, and BERT-base-cased via the SimpleTransformers (`https://github.com/ThilinaRajapakse/simpletransformers`) interface to Transformers (`https://github.com/huggingface/transformers`).

[4] We used *irrCAC* provided by the *R* software (`https://rdrr.io/cran/irrCAC/`), more specifically `conger.kappa.raw()` unweighted, which reduces to Cohen's κ for the case with two annotators.

Models trained on BookCorpus:						
tested on	LEMMAOVER.	GLEU	COSSIM	BERTSCORE	ALLSIMS	BERT
TED-Q	**0.46** (63%)	0.28 (9%)	0.12 (25%)	0.29 (54%)	0.42 (72%)	0.43 (84%)
BookCorpus	0.18 (13%)	0.06 (27%)	0.03 (21%)	0.14 (32%)	0.17 (30%)	**0.38** (44%)

Table 2: Results using the BookCorpus: MCC and percentage of errors that are false positives.

annotated by both annotators. These 10 are mostly questions that easily fit in multiple places (e.g. "What would be an example of this?") or questions that are related to the context but not quite in the right way (e.g. in a context mentioning eclipses but not their effects "Do eclipses have any other effects?"). Of the 11 false negatives, 4 were correctly annotated by both annotators; 3 out of these 4 are again quite general and would fit in many contexts (e.g. 'What exactly is being talked about?'). See the supplementary material for more examples (1-15).

We also looked into BERT's errors by using another source of human judgments: TED-Q's existing ANSWERED ratings (see Section 2). We find no significant difference in ANSWERED between correctly and incorrectly classified items (t(872)=-0.29, p=.76), but a difference between false positives and false negatives (for BERT, COSSIM, BERTSCORE and ALLSIMS): questions of false positives are significantly more answered (BERT: t(194)=-2.13, p=.03; COSSIM: t(372)=-2.30, p=.01; BERTSCORE: t(295)=-2.77, p=.005; ALLSIMS t(235)=-2.68, p=.007). We take this to reflect that questions that are subsequently answered tend to fit the discourse better than questions that remain unanswered, to such an extent that they can seem to be evoked even by neighboring contexts (hence false positives).

Comparing the similarity models As for the similarity-based models, what is arguably the simplest notion of similarity performs best: LEMMAOVERLAP. The slightly more syntactic notion, GLEU, is a lot worse, reflecting that although questions and the preceding context (which are mostly assertions) may share some content words, they tend to have quite a different structure – this is due for instance to subject-verb inversion in questions, but also due to the fact that the relation between question and context is not one of answerhood (contrary to, e.g., question answering tasks).[5] The more semantic notion of lexical similarity, MEANCOS, is surprisingly bad: apparently surface form matching (LEMMAOVERLAP) is more reliable than semantic matching (say, association).[6] Comparing LEMMAOVERLAP and BERTSCORE, our best similarity-based models, shows similar precision on the positive items while the former wins on recall (0.78 vs. 0.52). (For some examples see (16-20) in the supplementary material.)

The foregoing findings may reflect one (or both) of two aspects of the TED-Q dataset. One is the more general linguistic tendency for interlocutors to literally re-use each other's words where possible rather than variations and synonyms (Pickering and Garrod (2004)), which likely plays a role in TED-Q as well. The second is the possibility that reusing words from the context where possible is a cognitively easy way of completing the TED-Q crowdsource task: simply repeat some words from the context and put a question word in front of it (note that the TED-Q elicitation task prevented workers from directly copy-pasting a selection from the context). However, the comparison to explicit questions, discussed next, where LEMMAOVERLAP is likewise the best notion of similarity, provides some evidence against it being a mere crowdsourcing artifact.

Comparison to explicit questions in BookCorpus Table 2 shows analogous results for models trained on the BookCorpus-derived task, showing evaluation results on both TED-Q and BookCorpus. This shows that the BookCorpus dialogues make for a more challenging task overall (lower scores in the second row compared to the first row, and compared also to Table 1), and that the superiority of BERT over the similarity models is greater for the BookCorpus than for TED-Q in Table 1 (i.e., when training on TED-Q). This is expected: a main reason for making a question explicit is that it is insufficiently predictable from the context – as otherwise it could have been left implicit. Explicit

[5]TED-Q annotators were instructed that the question they entered should not be already answered by the context, but, rather, evoked by it.

[6]Using maximal mean cosine to any of the context's sentences, instead of the overall mean, offered no improvement.

questions are therefore expected to be more difficult to model computationally than the implicit questions of TED-Q. However, it may also in part reflect that LEMMAOVERLAP corresponds to an easy strategy for crowdsource workers, as mentioned above.

Lastly, Table 2 shows, in the first row, that training on the BookCorpus can prepare the models for the TED-Q task, to varying degrees. Indeed, the performance of LEMMAOVERLAP is the same as when trained on TED-Q (Table 1), and the other similarity models are very close as well. Comparatively, the BERT classifier suffers more, going down from 0.55 to 0.43, with a substantial error increase especially in false positives (from 61% to 84%). This suggests that notions of similarity generalize better between implicit and explicit questions (a consequence of their simplicity), whereas BERT, being the more powerful model, is better able to attune to the specifics of TED-Q's evoked questions vs. explicit questions in dialogue. The increase in false positives suggests that, on the whole, the TED-Q questions fit more readily in nearby contexts than the questions in the BookCorpus.

6 Conclusion and outlook

The TED-Q dataset appears to give us a decent window on the questions implicitly evoked by a text: the task we constructed cannot be solved by means of similarity alone, while the BERT-based classifier comes close to expert human performance (though human performance was evaluated for only a small subset of the data). Moreover, we observed some theoretically interesting patterns in the data through error analysis and by means of a comparison to explicit questions from the BookCorpus. For instance, the most successful similarity notion was plain LEMMAOVERLAP, reflecting that evoked questions tend to reuse material from the discourse – where the fact that we see the same for the BookCorpus data speaks against this being a mere crowdsourcing artifact.

The data we analyzed is richer than we could cover in this paper. For instance, BERT and LEMMAOVERLAP are much better on wh-questions (MCC 0.61/0.53, respectively) than on polar questions (0.30 and 0.17, respectively), suggesting a different kind of relationship to the context, which we hope to investigate in the future. It would also be good to collect a more comprehensive human benchmark on the tasks, expand our error analysis and consider a wider range of theoretically inspired models.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154) and from the Spanish State Research Agency (AEI) and the European Regional Development Fund (FEDER, UE) (project PGC2018-094029-A-I00). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains. The second author would like to thank the Santander bank for the Mobility Scholarship Funding and the welcoming hospitality of the COLT research group at the Universitat Pompeu Fabra.

References

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Wei-Jen Ko, Te-Yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of EMNLP*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Edgar Onea. 2013. Potential questions in discourse and grammar. *Manuscript, University of Göttingen*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. Qadiscourse–discourse relations as qa pairs: Representation, crowdsourcing and baselines. *arXiv preprint arXiv:2010.02815*.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. *arXiv preprint arXiv:2004.14530*.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia, July. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Arndt Riester. 2019. Constructing qud trees. In *Questions in Discourse*, pages 164–193. Brill.

Craige Roberts. 1996. Information structure in discourse: Toward a unified theory of formal pragmatics. *The Ohio State University Working Papers in Linguistics*, 49:91–136.

Luise Schricker and Tatjana Scheffler. 2019. Ranking of potential questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 143–148.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, pages 109–147.

Christiane Von Stutterheim and Wolfgang Klein. 1989. Referential movement in descriptive and narrative discourse. In *North-Holland Linguistic Series: Linguistic Variations*, volume 54, pages 39–76. Elsevier.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. Ted-q: Ted talks and the questions they evoke. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1118–1127.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December.

A Supplementary material for "Similarity or deeper understanding? Analyzing the TED-Q dataset of evoked questions"

The following lists of examples are not exhaustive, but exemplify some of the patterns we observed, as referenced in the paper.

False positives by BERT, coded correctly by both annotators.

(1) It was first suggested by Lyman Spitzer, the father of the space telescope, in 1962, and he took his inspiration from an eclipse. You've all seen that. That's a solar eclipse.
Question: `Do eclipses have any other effects?`

(2) I feel so fortunate that my first job was working at the Museum of Modern Art on a retrospective of painter Elizabeth Murray. I learned so much from her.
Question: `What did she think of those early works?`

(3) Mastery is about sacrificing for your craft and not for the sake of crafting your career. How many inventors and untold entrepreneurs live out this phenomenon? We see it even in the life of the indomitable Arctic explorer Ben Saunders, who tells me that his triumphs are not merely the result of a grand achievement,
Question: `What would be an example of this?`

(4) His favorite novel was "The [Unknown] Masterpiece" by Honoré de Balzac, and he felt the protagonist was the painter himself. Franz Kafka saw incompletion when others would find only works to praise, so much so that he wanted all of his diaries, manuscripts, letters and even sketches burned upon his death. His friend refused to honor the request, and because of that, we now have all the works we now do by Kafka.
Question: `Where is that quote from`

False positives by BERT also coded incorrectly by both annotators.

(5) Even in the developed world, it takes a period of three weeks to often years for a patient to get a comfortable socket, if ever. Prosthetists still use conventional processes like molding and casting to create single-material prosthetic sockets. Such sockets often leave intolerable amounts of pressure on the limbs of the patient, leaving them with pressure sores and blisters.
Question: `How to solve this problem?`
Question: `Is there anything to be done to speed this process along?`

(6) In most of our minds, the vagabond is a creature from the past. The word "hobo" conjures up an old black and white image of a weathered old man covered in coal, legs dangling out of a boxcar, but these photographs are in color, and they portray a community swirling across the country, fiercely alive and creatively free, seeing sides of America that no one else gets to see. Like their predecessors, today's nomads travel the steel and asphalt arteries of the United States.
Question: `Why is the vagabond a creature from the past?`

(7) It's a very heterogenous city in a way that Baltimore or San Francisco is not. You still have the lobe of people involved with government, newspapers, politics, columnists. TEDxRio is down in the lower right, right next to bloggers and writers.
Question: `What exactly is TEDxRio?`

False negatives by BERT, coded correctly by both annotators.

(8) The reason, I would come to find out, was that their prosthetic sockets were painful because they did not fit well. The prosthetic socket is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle. Even in the developed world, it takes a period of three weeks to often years for a patient to get a comfortable socket.
Question: `How did they cope with the situation in the mean time?`

(9) We call that our flower petal starshade. If we make the edges of those petals exactly right, if we control their shape, we can control diffraction, and now we have a great shadow. It's about 10 billion times dimmer than it was before, and we can see the planets beam out just like that.
Question: `What exactly is being talked about?`

(10) They're masters because they realize that there isn't one. Now it occurred to me, as I thought about this, why the archery coach told me at the end of that practice, out of earshot of his archers, that he and his colleagues never feel they can do enough for their team, never feel there are enough visualization techniques and posture drills to help them overcome those constant near wins. It didn't sound like a complaint, exactly, but just a way to let me know, a kind of tender admission, to remind me that he knew he was giving himself over to a voracious, unfinished path that always required more.
Question: `What more would they like to be doing?`

(11) Many of you might be wondering why anyone would choose a life like this, under the thumb of discriminatory laws, eating out of trash cans, sleeping under bridges, picking up seasonal jobs here and there. The answer to such a question is as varied as the people that take to the road, but travelers often respond with a single word: freedom. Until we live in a society where every human is assured dignity in their labor so that they can work to live well, not only work to survive, there will always be an element of those who seek the open road as a means of escape, of liberation and, of course, of rebellion.
Question: `Homelessness is rebellion from what?`

False negatives by BERT also coded incorrectly by both annotators.

(12) So if the returns are the same or better and the planet benefits, wouldn't this be the norm? Are investors, particularly institutional investors, engaged? Well, some are, and a few are really at the vanguard.
Question: `Is this about stock exchange?`

(13) But this is the thing: What gets us to convert success into mastery? This is a question I've long asked myself. I think it comes when we start to value the gift of a near win.
Question: `anything else?`

(14) Masters are not experts because they take a subject to its conceptual end. They're masters because they realize that there isn't one. Now it occurred to me, as I thought about this, why the archery coach told me at the end of that practice, out of earshot of his archers, that he and his colleagues never feel they can do enough for their team, never feel there are enough visualization techniques and posture drills to help them overcome those constant near wins.
Question: `How does that apply in practical terms here?`

(15) No one loses their inner demons by taking to the road. Addiction is real, the elements are real, freight trains maim and kill, and anyone who has lived on the streets can attest to the exhaustive list of laws that criminalize homeless existence. Who here knows that in many cities across the United States it is now illegal to sit on the sidewalk, to wrap oneself in a blanket,
Question: `Where do they go?`

False negatives by BERTSCORE, correct by LEMMAOVERLAP

(16) The world is changing in some really profound ways, and I worry that investors aren't paying enough attention to some of the biggest drivers of change,
Question: `What is the change?`

(17) The world is changing in some really profound ways, and I worry that investors aren't paying enough attention to some of the biggest drivers of change, especially when it comes to sustainability. And by sustainability, I mean the really juicy things, like environmental and social issues and corporate governance. I think it's reckless to ignore these things, because doing so

can jeopardize future long-term returns.
Question: `How can it jeopordize the future`

False positives by LEMMAOVERLAP; correct by BERTSCORE

(18) That's why it's hard. The light from the star is diffracting. It's scattering inside the telescope, creating that very bright image that washes out the planet.
Question: `What planet is in the red circle?`

(19) It was first suggested by Lyman Spitzer, the father of the space telescope, in 1962, and he took his inspiration from an eclipse. You've all seen that. That's a solar eclipse.
Question: `What does this have to do with Spitzer's suggestion?`

(20) I feel so fortunate that my first job was working at the Museum of Modern Art on a retrospective of painter Elizabeth Murray.
Question: `What did the painter think of her early works?`