Evaluating the Examiner: The Perils of Pearson Correlation for Validating Text Similarity Metrics

Gisela Vallejo[1] **Timothy Baldwin**[1,2] **Lea Frermann**[1]
[1]The University of Melbourne [2]MBZUAI
gvallejo@student.unimelb.edu.au,
{tbaldwin,lea.frermann}@unimelb.edu.au

Abstract

In recent years, researchers have developed question-answering based approaches to automatically evaluate system summaries, reporting improved validity compared to word overlap-based metrics like ROUGE, in terms of correlation with human ratings of criteria including fluency and hallucination. In this paper, we take a closer look at one particular metric, QuestEval, and ask whether: (1) it can serve as a more general metric for long document similarity assessment; and (2) a single correlation score between metric scores and human ratings, as the currently standard approach, is sufficient for metric validation. We find that correlation scores can be misleading, and that score distributions and outliers should be taken into account. With these caveats in mind, QuestEval can be a promising candidate for long document similarity assessment.

1 Introduction

Methods which can provide accurate estimates of document content similarity are critical to tasks such as news analysis and fact-checking (Shaar et al., 2020). Researchers have proposed a broad range of metrics to estimate document similarity (Sai et al., 2020), from n-gram overlap metrics such as BLEU (Papineni et al., 2002) and Meteor (Lavie and Agarwal, 2007) for machine translation, and ROUGE (Lin, 2004) for automatic summarisation, to embedding-based metrics such as BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019). However, these metrics have been shown to rely heavily on superficial features, correlate poorly with human annotations, and perform poorly over longer document pairs (Hanna and Bojar, 2021; Balasubramanian et al., 2020; Kryscinski et al., 2019; Koto et al., 2022).

A more radical recent proposal has been to use question-answering (QA) based models (Wang et al., 2020; Scialom et al., 2021), to automatically generate question–answer pairs from a source document, and estimate similarity by the proportion of questions that can be successfully answered on the basis of the target document. While such approaches were designed to evaluate automatic summarisation in a reference-free manner, i.e., compare a full (long) document with its (short) summary, they can in principle be applied to arbitrary document pairs. In this paper we ask whether the QuestEval method (Scialom et al., 2021) scales to varying-length document pairs, and in particular, can be used to calculate the similarity between same length documents reliably. In other words, we are comparing two evaluation settings: long–short document pairs vs. documents of the same length. In Table 1, we present the different document length scenarios in terms of average length.

Data	Avg. Len. Doc 1	Avg. Len. Doc 2
ABC News	86	86
SemEval	535	535
SummEval	63	359

Table 1: Average document length (words) in each dataset. In the case of SummEval, Doc 1 denotes a summary while Doc 2 the source text.

Consistent with other work on the evaluation of similarity metrics (including the original QuestEval paper), we explore this question by measuring the Pearson correlation between the estimated similarity scores and a gold standard. Pearson correlation is notoriously susceptible to outliers (Sai et al., 2020; Mathur et al., 2020), so in addition to the raw correlation values, we perform detailed analysis of the distribution of the gold and predicted similarity scores (via inspection of scatter plots). We find that reported correlations can be inflated by a small number of outliers, caused by a skewed distribution in the gold standard, and are thus hot fully reflective of the quality of QuestEval.

Our contributions are as follows: (1) we eval-

uate QuestEval on three different datasets, and demonstrate that it is robust to increasing document lengths; (2) we showcase the perils of presenting Pearson correlation coefficients for metric evaluation in isolation, without examining the raw data distribution; and (3) we suggest visualization strategies which expose possible data biases to the interpretation of raw correlation values.

2 Background

2.1 Evaluating text similarity evaluation

Most common automatic metrics for evaluating summarisation like BLEU and ROUGE, and BERTScore measure lexical overlap. In the case of BLEU and ROUGE, this is based on n-gram overlap, interpolated over different values of n, and with an additional brevity penalty in the case of BLEU. BERTScore, on the other hand, abstracts away from the tokens in calculating similarity based on contextualized embeddings of each token in the respective documents.

While these metrics are computationally inexpensive, they do not penalize critical content divergences (e.g. due to "hallucination" under summarisation: Wang et al. (2020)) or repetitions, and are poor at capturing meaning-critical differences in polarity. Such shortcomings were a large part of the motivation behind QA-based metrics such as QuestEval, which were shown by the authors to be more adept at evaluating factual consistency. We note that subsequent work of Koto et al. (2022) showed that with appropriate model and layer selection, BERTScore is actually superior in evaluating all aspects of summary quality, including factuality. Additionally, unlike the metrics above, QuestEval does not require a reference summary, as it is exclusively based on the consistency between document and generated summary (although varieties of the metric *can* leverage human annotations).

2.2 QuestEval

QuestEval is QA-based pipeline that generates question–answer pairs from a source document, and measures similarity by the proportion of those questions which can be successfully answered based on the target document. While in the context of summarisation evaluation, this is based on the source document and summary, respectively (to test how faithfully the summary captures the content of the source document), this can be applied to document similarity by performing the calculation in both directions and averaged. That is, for a document pair (d_i, d_j), separate scores can be calculated taking each of d_i and d_j as the source document, and the remaining document as the target document.

QuestEval consists of a question generation (QG) and a question answering (QA) model. In question generation, QuestEval selects nouns and named entities as gold-standard answers, and generates questions for them. The model generates questions for each of the nouns and name entities and discards the ones that the QA module is not able to answer correctly. The QuestEval metric comprises two evaluations, which measure whether the summary contains *only* true information (precision), and conversely whether it contains *all* important information (recall). Both the QG and QA components are a fine-tuned version of T5 (Raffel et al., 2020) using SQuAD-v2 (Rajpurkar et al., 2018). Even though SQuAD – where answers are generated based on Wikipedia paragraphs – is not comparable to typical summarization datasets which consist of news articles, the original QuestEval paper showed that the method is robust to the domain shift between component pre-training data and final application. This paper further asks whether QuestEval extends to document similarity assessment more generally, between arbitrary document pairs.

It is worth mentioning that the typical input limit of 512 tokens of pre-trained language models does not affect QuestEval, because the model generates and answers questions based on pre-identified nouns in their *local* context of five sentences. Thus, there is no limit on the document length that QuestEval can be applied to.

3 Experimental Setup

Here, we describe the datasets and evaluation methods we use to test QuestEval's applicability to long documents, as well as reliability across datasets and reference annotations.

3.1 Data

We experiment with three datasets: (1) SummEval, made up of article–summary pairs (long–short); (2) ABC News, consisting of article–article pairs (long–long); and (3) SemEval, also made up of article–article pairs (long–long). In each case, a given document pair is associated with one or more ground-standard labels.

Condition	Measure	Data	r	ρ
Long–Short	Coherence	SummEval	0.22	0.21
Long–Short	Consistency	SummEval	0.41	0.33
Long–Short	Fluency	SummEval	0.30	0.20
Long–Short	Relevance	SummEval	0.35	0.31
Long–Long	Doc Sim	ABC News	0.33	0.10
Long–Long	Doc Sim	SemEval	0.77	0.74

Table 2: Pearson (r) and Spearman (ρ) correlation coefficients for QuestEval scores under different data conditions.

Condition	Measure	Data	r	ρ
Long–Short	Coherence	SummEval	0.22	0.20
Long–Short	Consistency	SummEval	0.37	0.30
Long–Short	Fluency	SummEval	0.25	0.18
Long–Short	Relevance	SummEval	0.33	0.30
Long–Long	Doc Sim	ABC News	<u>0.11</u>	<u>0.06</u>
Long–Long	Doc Sim	SemEval	0.77	0.72

Table 3: Pearson (r) and Spearman (ρ) correlation coefficients, after removing outliers. We underline the most drastic drops.

SummEval (Fabbri et al., 2021) consists of 1600 generated summaries from 16 different models generated for a random sample of 100 articles from the CNN/DailyMail dataset (Hermann et al., 2015), and was used in the original QuestEval publication (Scialom et al., 2021). The average length of each generated summary and source document is 63 and 359 words respectively. Each summary was rated by three experts and five non-experts (crowdworkers) regarding coherence, consistency, fluency, and relevance. In our experiments, we only use the expert ratings for all four dimensions. Note that coherence and fluency are intrinsically intra-document properties, independent of the source document. As such, QuestEval is a slightly odd choice of method, given that it compares the source document with the summary. In line with the original QuestEval paper, however, we include these results based on the hypothesis that there should be some influence on the ability to correctly answer questions if the summary lacks coherence or fluency.

ABC News (Lee et al., 2005) consists of 1225 document-pairs, created by exhaustively pairing 50 news articles taken from the Australian Broadcasting Corporation (ABC) news service. The average article length is 86 words. Each article pair was rated by 8-10 annotators for similarity on a five-point scale from 1 (highly unrelated) to 5 (highly related). In our experiments, we compare QuestEval scores against the average annotated similarity per article pair.

SemEval (Chen et al., 2022) was published as part of SemEval-2022 Task 8: Multilingual news article similarity. The full dataset contains 10K pairs of documents from 10 languages, including both monolingual (two documents in the same language, e.g., English) and cross-lingual (documents in different languages, e.g., English vs. Arabic) pairs. Here we only use the 1348 pairs of the training split where both documents are English.[1] The average article length is 535 words. Document pairs were labeled by trained annotators for a variety of axes of similarity (tone, style, narrative, temporal and geographical range, and entities) as well as overall similarity. Annotations were collected on a four-point scale from 1 (very dissimilar) to 4 (very similar).[2] In our experiments, we include only the *overall* similarity score, which we correlate with QuestEval similarity.

3.2 Validating QuestEval Scores

We obtained QuestEval scores for all three datasets using QuestEval version 0.1.1[3] and calculated the Pearson and Spearman correlation coefficients of the respective gold labels with our QuestEval scores. We report the results in Table 2. It is widely known that correlation scores are susceptible to outliers (Sai et al., 2020; Mathur et al., 2020), rendering the findings less robust. To assess the robustness of observed correlations, we additionally inspect the full distributions of gold ratings and QuestEval scores in Figure 1 in the form of kernel density estimation (KDE) plots, onto which we superimpose the regression line of best fit based on Pearson correlation. We also include the raw scatter plots in Appendix C for comparison.

4 Results

In analysing the results, we investigate: (1) whether QuestEval is document-length agnostic, i.e., scales from the original scenario of article–summary

[1] Noting that the script for reproducing the dataset occasionally failed, so that we evaluate on 74% of the data described in Chen et al. (2022).

[2] The original annotations were collected on the reverse scale (4: most dissimilar), but we flip the scores for consistency with the other results.

[3] The authors provide this link with the source code to reproduce the scores reported in the paper: https://github.com/recitalAI/QuestEval/releases/tag/v0.1.1

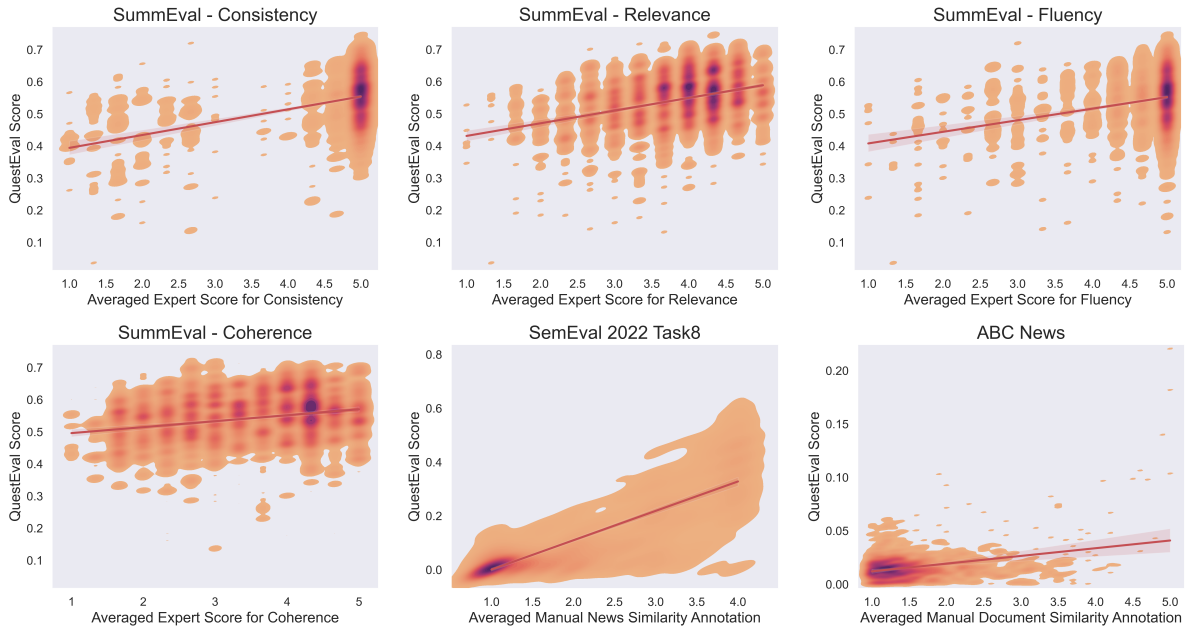

Figure 1: Visualised correlation (heat map of raw data + correlation line) for QuestEval with several human annotated metrics for SummEval, ABC News, and SemEval.

(long–short) similarity to estimating article–article (long–long) similarity in terms of raw Pearson Correlation scores; (2) whether QuestEval correlates with ratings of document similarity, departing from the dimensions of coherence, consistency, fluency, and relevance as originally assessed; and (3) how robust the observed Pearson and Spearman correlations are across all data conditions and ground-truth labels.

QuestEval as a measure of long document similarity The correlation coefficients reported in Table 2 address questions (1) and (2). The top block in the table shows our reproduction of the original QuestEval evaluation setup (Scialom et al., 2021).[4] Our numbers are comparable to the original reported scores, and confirm that QuestEval best captures consistency (i.e,. content similarity) and to a lesser extent accounts for the other three axes of summary quality. The bottom block of Table 2 shows the correlation of QuestEval with the respective manual document similarity scores in the ABC News and SemEval datasets. Both are either close or exceed the best evaluation score obtained for summary evaluation, suggesting that the metric indeed can be employed to estimate long document similarity. However, given the coefficient's high sensitivity to outliers — and consequently the distribution of reference and QuestEval scores — we next assess the robustness of the reported score.

Robustness of QuestEval validation Validating automatic evaluation metrics in terms of their correlation to human labels seems intuitive, however, correlation scores like Pearson are susceptible to outliers. This is particularly pertinent in cases where rank (or label) distributions are skewed, as is often the case when collecting human similarity ratings. Consider the data densities implied for the human quality/similarity ratings in Figure 1, i.e., densities along the x-axis. For most metrics (with the exception of relevance and coherence in SummEval), human labels are concentrated at one end of the spectrum, suggesting that instances labelled with unusual ratings are outliers and to some degree atypical. We can thus achieve high Pearson correlation scores under these highly atypical data conditions.

Conversely, if the outliers were removed, the correlation would drop substantially. Following Mathur et al. (2020), we removed outliers in all datasets based on QuestEval scores x by means of the Median Absolute Deviation (MAD) as shown below:

$$\text{cutoff} < \frac{|x - \text{median}(x)|}{\text{MAD}(x)}$$

[4] Compared to QUESTEVAL$_{W_{\text{uniform}}}$ our coherence, consistency, and relevance scores are 1–2 points lower and fluency scores are 1.3 points higher than those reported in the paper. We also include Spearman, which is not reported in the orginal paper.

Data	Cutoff	# of Outliers
ABC News	5.5	20
SemEval	10	39
SummEval	3.5	16

Table 4: Selected cutoff parameter for each dataset for outliers removal as well as total number of removed outliers.

We selected a different cutoff for each of the datasets, taking as reference box plots, and depict cutoffs and the total amount of outliers in Table 4. Raw scatter plots of the data including removed outliers are illustrated in Figure 2. We report the obtained results in Table 3 and show how the correlations drop for all datasets. The effect is particularly pertinent in the case of ABC News, with a drop of about 22 absolute points in Pearson correlation. Here, the removal of a small number of outliers (similarity > 4.0) would reduce correlation close to zero. On the other hand, for the SemEval 2022 documents, we observe a relatively wide spread of human labels, and correspondingly small impact of removing outliers, and can conclude that the high correlation with QuestEval scores (Table 2) is reliable.

We observe a similar trend for the best-correlated SummEval score of Consistency, for which 89.4% of the data points were labeled with a score > 4.0. SummEval Relevance and Coherence scores are more evenly spread, leading to lower, albeit much more robust, estimates of Pearson correlation. Beyond that, we are aware that Pearson correlation is sensitive to outliers and Spearman correlation is less robust when the distribution happens to have clusters. None of these metrics are perfect and therefore it is crucial to understand the data, plot the distributions in scatter plots and conclude how informative are correlation coefficients.

5 Analysis and Discussion

From our results we can observe that summarisation evaluation metrics and more specifically, QuestEval have utility for tasks beyond summarisation, especially where there is no access to gold human annotations. In our case, we showed that QuestEval scores do correlate with the overall news article similarity scores of SemEval. However, this is not the case for every metric, as we were also able to show with dimensions like document similarity, consistency, and fluency. Moreover, we showed that in isolation Pearson correlation coefficients with human ratings are not a reliable signal for the quality of an evaluation metric, due to their sensitiveness to outliers. We recommend to visualise score distributions in tandem with calculating the correlation to ensure that it is not affected by a minority of outliers. This is consistent with the observations of Mathur et al. (2020) in their analysis of WMT task results. We observed that QuestEval scores are distributed in the range of 0–1 for almost all datasets/measurements except for ABC News, motivating us to look more closely at this dataset. In the Appendix we present some examples with high document similarities but low QuestEval scores. While we are aware that QuestEval values are lower than expected for those examples, the similarity rating is also arguable. For both cases, almost none of the entities overlap in the depicted documents; this could be the reason why QuestEval scores are low. We also propose to take into consideration several correlation coefficients as we show in Table 2. In addition to that, it is also important to understand the data by plotting it to look for useful patterns.

6 Conclusion

In this paper we investigated whether automatic QA-based metrics for summarisation evaluation can be adopted to compare long documents. We also conducted a more detailed evaluation of the robustness of Pearson correlation for similarity metric evaluation, and found that correlation-based metrics need to be validated by plotting and understanding labels and score distributions. In future work, we plan to extend our work to different languages.

References

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a name? are BERT named entity representations just as good for any other name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214, Online. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on*

Semantic Evaluation (SemEval-2022), pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2022. FFCI: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73:1553–1607.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Michael D Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 27.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

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 (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for NLG systems. *CoRR*, abs/2008.12009.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong

Kong, China. Association for Computational Linguistics.

A Limitations

We are aware that our analysis may be biased because we focus only on English data. Additionally, due to time constrains we were not able to comprehensively clean the SemEval data, so there may be remnant noise.

B ABC News Examples

See Table 5 for examples where the gold-standard similarity is high but QuestEval score is exceedingly low compared to a sample of documents that are indeed very similar and get high scores from annotations as well as from QuestEval.

C Scatterplots

Figure 2 is a complement to the kernel density plots of Figure 1, and presents the raw scatter plots for the different datasets and removed outliers.

Averaged Similarity: 3.7 – QuestEval Score: 0.0004	
The Bush administration has drawn up plans to escalate the war of words against Iraq, with new campaigns to step up pressure on Baghdad and rally world opinion behind the US drive to oust President Saddam Hussein. This week, the State Department will begin mobilising Iraqis from across North America, Europe and the Arab world, training them to appear on talk shows, write opinion articles and give speeches on reasons to end President Saddam's rule.	The Iraqi capital is agog after the violent death of one of the world's most notorious terrorists, but the least of the Palestinian diplomat's worries was the disposal of Abu Nidal's body, which lay on a slab in an undisclosed Baghdad morgue. Abu Nidal's Fatah Revolutionary Council is held responsible for the death or injury of almost 1000 people in 20 countries across Europe and the Middle East in the three decades since he fell out with Yasser Arafat over what Abu Nidal saw as Arafat's willingness to accommodate Israel in the Palestinian struggle.
Averaged Similarity: 3.9 – QuestEval Score: 0.0003	
U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials.	The Iraqi capital is agog after the violent death of one of the world's most notorious terrorists, but the least of the Palestinian diplomat's worries was the disposal of Abu Nidal's body, which lay on a slab in an undisclosed Baghdad morgue. Abu Nidal's Fatah Revolutionary Council is held responsible for the death or injury of almost 1000 people in 20 countries across Europe and the Middle East in the three decades since he fell out with Yasser Arafat over what Abu Nidal saw as Arafat's willingness to accommodate Israel in the Palestinian struggle.
Averaged Similarity: 5.0 – QuestEval Score: 0.182	
An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing.	Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry.

Table 5: Examples from ABC News with high gold-standard similarity but very low QuestEval scores compared to an document pair having high scores in both annotations and QuestEval score.

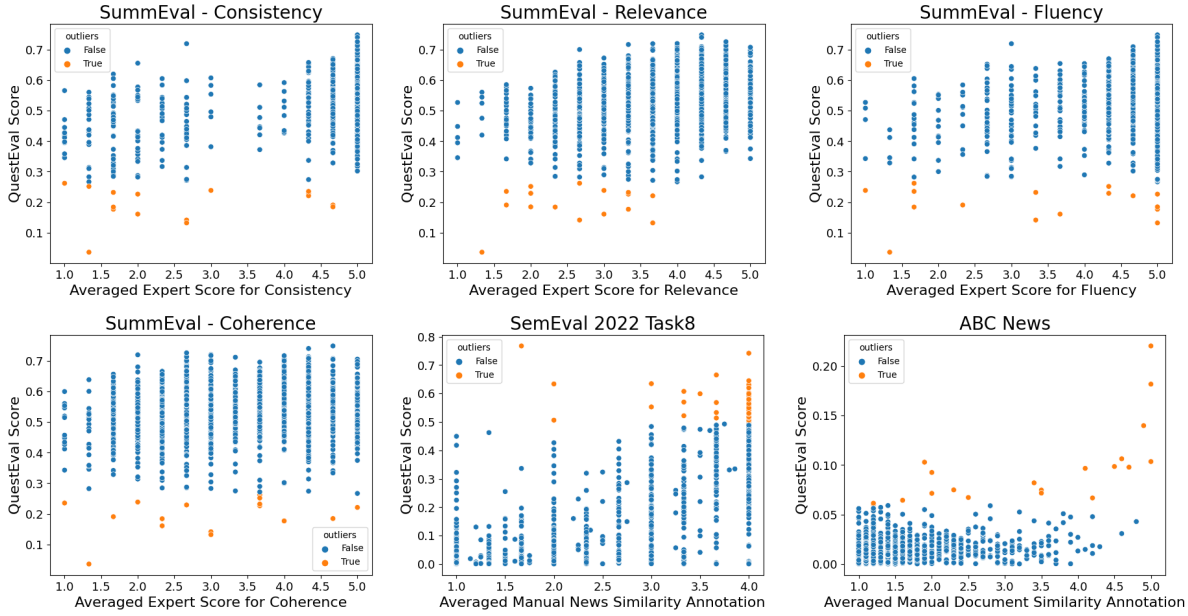

Figure 2: Raw scatter plots of QuestEval vs. gold-standard scores for SemEval, ABC News and SummEval. Data points in orange represent the removed outliers.