Evaluation of Semantic Answer Similarity Metrics

[1]**Farida Mustafazade**
[1]GAM Systematic

`farida.mustafazade.15@ucl.ac.uk`

[2]**Peter Ebbinghaus**
[2]Teufel Audio

`peter.ebbinghaus@posteo.de`

Abstract

There are several issues with the existing general machine translation or natural language generation evaluation metrics, and question-answering (QA) systems are indifferent in that context. To build robust QA systems, we need the ability to have equivalently robust evaluation systems to verify whether model predictions to questions are similar to ground-truth annotations. The ability to compare similarity based on semantics as opposed to pure string overlap is important to compare models fairly and to indicate more realistic acceptance criteria in real-life applications. We build upon the first to our knowledge paper that uses transformer-based model metrics to assess semantic answer similarity and achieve higher correlations to human judgement in the case of no lexical overlap. We propose cross-encoder augmented bi-encoder and BERTScore models for semantic answer similarity, trained on a new dataset consisting of name pairs of US-American public figures. As far as we are concerned, we provide the first dataset of co-referent name string pairs along with their similarities, which can be used for training.

1 Introduction

Having reliable metrics for evaluation of language models in general, and models solving difficult question answering (QA) problems, is crucial in this rapidly developing field. These metrics are not only useful to identify issues with the current models, but they also influence the development of a new generation of models. In addition, it is preferable to have an automatic, simple metric as opposed to expensive, manual annotation or a highly configurable and parameterisable metric so that the development and the hyperparameter tuning do not add more layers of complexity. SAS, a cross-encoder-based metric for the estimation of semantic answer similarity (Risch et al., 2021), provides one such

Question: Who makes more money: NFL or Premier League?
Ground-truth answer: National Football League
Predicted Answer: the NFL
EM: 0.00
F_1: 0.00
Top-1-Accuracy: 0.00
SAS: 0.9008
Human Judgment: 2 (definitely correct prediction)
f_{BERT}: 0.4317
f'_{BERT}: 0.4446
Bi-Encoder: 0.5019

Figure 1: Representative example from NQ-open of a question and all semantic answer similarity measurement results.

metric to compare answers based on semantic similarity.

The central objective of this research project is to analyse pairs of answers similar to the one in Figure 1 and to evaluate evaluation errors across datasets and evaluation metrics.

The main hypotheses that we will aim to test thoroughly through experiments are twofold. Firstly, lexical-based metrics are not well suited for automated QA model evaluation as they lack a notion of context and semantics. Secondly, most metrics, specifically SAS and BERTScore, as described in (Risch et al., 2021), find some data types more difficult to assess for similarity than others.

After familiarising ourselves with the current state of research in the field in Section 2, we describe the datasets provided in (Risch et al., 2021) and the new dataset of names that we purposefully tailor to our model in Section 3. This is followed by Section 4, introducing the four new semantic answer similarity approaches described in (Risch et al., 2021), our fine-tuned model as well as three lexical n-gram-based automated metrics. Then in Section 5, we thoroughly analyse the evaluation datasets described in the previous section and conduct an in-depth qualitative analysis of the errors.

Finally, in Section 6, we summarise our contributions.

2 Related work

We define semantic similarity as different descriptions for something that has the same meaning in a given context, following largely (Zeng, 2007)'s definition of semantic and contextual synonyms. In Min et al. (2021), the human annotators attach a label 2 to all predictions that are "definitely correct", 1 - "possibly correct", and 0 - "definitely incorrect". Automatic evaluation based on exact match (EM) fails to capture semantic similarity for definitely correct answers, where 60% of the predictions are semantically equivalent to the ground-truth answer. Just under a third of the predictions that do not match the ground-truth labels were nonetheless correct. They also mention other reasons for failure to spot equivalence, such as time-dependence of the answers or underlying ambiguity in the questions.

QA evaluation metrics in the context of SQuAD v1.0 (Rajpurkar et al., 2016) dataset are analysed in Bulian et al. (2022). They thoroughly discuss the limitations of EM and F1 score from n-gram based metrics, as well as the importance of context including the relevance of questions to the interpretation of answers. A BERT matching metric (Bert Match) is proposed for answer equivalence prediction, which performs better when the questions are included alongside the two answers, but appending contexts didn't improve results. Additionally, authors demonstrate better suitability of Bert Match in constructing top-k model's predictions. In contrast, we will cover multilingual datasets, as well as more token-level equivalence measures, but limit our focus on similarity of answer pairs without accompanying questions or contexts.

Two out of four semantic textual similarity (STS) metrics that we analyse and the model that we eventually train depend on bi-encoder and BERTScore (Zhang et al., 2019). The bi-encoder approach model is based on the Sentence Transformer structure (Reimers and Gurevych, 2019a), which is a faster adaptation of BERT for the semantic search and clustering type of problems. BERTScore uses BERT to generate contextual embeddings, then match the tokens of the ground-truth answer and prediction, followed by creating a score from the maximum cosine similarity of the matched tokens. This metric is not one-size-fits-all. On top of choosing a suitable contextual embedding and model,

there is an optional feature of importance weighting using inverse document frequency (idf). The idea is to limit the influence of common words. One of the findings is that most automated evaluation metrics demonstrate significantly better results on datasets without adversarial examples, even when these are introduced within the training dataset, while the performance of BERTScore suffers only slightly. (Zhang et al., 2019) uses machine translation (MT) and image captioning tasks in experiments and not QA. (Chen et al., 2019) apply BERT-based evaluation metrics for the first time in the context of QA. Even though they find that METEOR as an n-gram based evaluation metric proved to perform better than the BERT-based approaches, they encourage more research in the area of semantic text analysis for QA. Moreover, (Bulian et al., 2022) uses only BERTScore base as one of the benchmarks, while we explore the larger model, as well as a finetuned variation of it.

Authors in (Risch et al., 2021) expand on this idea and further address the issues with existing general MT, natural language generation (NLG), which entails as well generative QA and extractive QA evaluation metrics. These include reliance on string-based methods, such as EM, F1-score, and top-n-accuracy. The problem is even more substantial for multi-way annotations. Here, multiple ground-truth answers exist in the document for the same question, but only one of them is annotated. The major contribution of the authors is the formulation and analysis of four semantic answer similarity approaches that aim to resolve to a large extent the issues mentioned above. They also release two three-way annotated datasets: a subset of the English SQuAD dataset (Rajpurkar et al., 2018), German GermanQuAD dataset (Möller et al., 2021), and NQ-open (Min et al., 2021).

As depicted in Table 4 and Section 5, the leading problematic data type category is entities, particularly those involving names. Si et al. (2021) analyse Natural Questions (NQ) (Kwiatkowski et al., 2019b), TriviaQA (Joshi et al., 2017) as well as SQuAD and address the issue that current QA benchmarks neglect the possibility of multiple correct answers. They focus on the variations of names, e.g. nicknames, and improve the evaluation of Open-domain QA models based on a higher EM score by augmenting ground-truth answers with aliases from Wikipedia and Freebase. In our work, we focus solely on the evaluations of answer eval-

uation metrics and generate a standalone names dataset from another dataset, described in greater detail in Section 3.

Our main assumption is that better metrics will have a higher correlation with human judgement, but the choice of a correlation metric is important. Pearson correlation is a commonly used metric in evaluating semantic text similarity (STS) for comparing the system output to human evaluation. (Reimers et al., 2016) show that Pearson power-moment correlation can be misleading when it comes to intrinsic evaluation. They further go on to demonstrate that no single evaluation metric is well suited for all STS tasks, hence evaluation metrics should be chosen based on the specific task. In our case, most of the assumptions, such as normality of data and continuity of the variables behind Pearson correlation do not hold. Kendall's rank correlations are meant to be more robust and slightly more efficient in comparison to Spearman as demonstrated in (Croux and Dehon, 2010).

Soon after Transformers took over the field, adversarial tests resulted in significantly lower performance figures, which increased the importance of adversarial attacks (Niven and Kao, 2019). General shortcomings of language models and their benchmarks led to new approaches such as Dynabench (Kiela et al., 2021) and AdvGLUE (Wang et al., 2021). There are other shortcomings of large language models, including environmental and financial costs (Bender et al., 2021).

3 Data

We perform our analysis on three subsets of larger datasets annotated by three human raters and provided by (Risch et al., 2021). Unless specified otherwise, these will be referred to by their associated dataset names.

3.1 Original datasets

SQuAD is an English-language dataset containing multi-way annotated questions with 4.8 answers per question on average. **GermanQuAD** is a three-way annotated German-language question/answer pairs dataset created by the deepset team which also wrote (Risch et al., 2021). Based on the German counterpart of the English Wikipedia articles used in SQuAD, GermanQuAD is the SOTA dataset for German question answering models. To address a shortcoming of SQuAD that was mentioned in (Kwiatkowski et al., 2019a), GermanQuAD was

created with the goal of preventing strong lexical overlap between questions and answers. Hence, more complex questions were encouraged, and questions were rephrased with synonyms and altered syntax. SQuAD and GermanQuAD contain a pair of answers and a hand-labelled annotation of 0 if answers are completely dissimilar, 1 if answers have a somewhat similar meaning, and 2 if the two answers express the same meaning. **NQ-open** is a five-way annotated open-domain adaption of Kwiatkowski et al. (2019a)'s Natural Questions dataset. NQ-open is based on actual Google search engine queries. In case of NQ-open, the labels follow a different methodology as described in (Min et al., 2021). The assumption is that we only leave questions with a non-vague interpretation (see Table 4). Questions like *Who won the last FIFA World Cup?* received the label 1 because they have different correct answers without a precise answer at a point in time later than when the question was retrieved. There is yet another ambiguity with this question, which is whether it is discussing FIFA Women's World Cup or FIFA Men's World Cup. This way, the two answers can be correct without semantic similarity even though only one correct answer is expected.

The annotation of NQ-open indicates truthfulness of the predicted answer, whereas for SQuAD and GermanQuAD the annotation relates to the semantic similarity of both answers which can lead to differences in interpretation as well as evaluation. To keep the methodology consistent and improve NQ-open subset, vague questions with more than one ground-truth labels have been filtered out. We also manually re-label incorrect labels as well as filter out vague questions.

Table 1 describes the size and some lexical features for each of the three datasets. There were 2, 3 and 23 duplicates in each dataset respectively. Dropping these duplicates led to slight changes in the metric scores.

3.2 Augmented dataset

For NQ-open, the largest of the three datasets, names was the most challenging category to predict similarity as per 2. While it includes city and country names as well, we focus on the names of public figures in our work. To resolve this issue, we provide a new dataset that consists of ~40,000 (39,593) name pairs and employ the Augmented SBERT approach (Thakur et al., 2021): we use the

	SQuAD	GermanQuAD	NQ-open
Label 0	56.7	27.3	71.7
Label 1	30.7	51.5	16.6
Label 2	12.7	21.1	11.7
$\mathbf{F_1 = 0}$	565	124	3030
$\mathbf{F_1 \neq 0}$	374	299	529
Size	939	423	3559
Avg answer size	23	68	13

Table 1: Percentage distribution of the labels and statistics on the subsets of datasets used in the analyses. The average answer size column refers to the average of both the first and second answers as well as ground-truth answer and predicted answer (NQ-open only). $F_1 = 0$ indicates no string similarity, $F_1 \neq 0$ indicates some string similarity. Label distribution is given in percentages.

cross-encoder model to label a new dataset consisting of name pairs and then train a bi-encoder model on the resulting dataset. We discuss the deployed models in more detail in Section 4.

The underlying dataset is created from an open dbpedia-data dataset (Wagner, 2017) which includes the names of more than a million public figures that have a page on Wikipedia and DBpedia, including actors, politicians, scientists, sportsmen, and writers. Out of these we only use those with a U.S. nationality as the questions in NQ-open are on predominantly U.S. related topics. We then shuffle the list of 25,462 names and pair them randomly to get the name pairs that are then labelled by the cross-encoder model. It includes different ways of writing a person's name including aliases. For example, *Gary A Labranche* and *Labranche Gary*, or aliases like *Lisa Marie Abato*'s stage name *Holly Ryder* as well as e.g. Chinese ways of writing such as *Rulan Chao Pian* and 卞趙如蘭. We filter out all examples where more than three different ways of writing a person's name exist because in these cases these names don't refer to the same person but were mistakenly included in the dataset. For example, names of various members of Tampa Bay Rays minor league who have one page for all members. Since most public figures in the dataset have a maximum of one variation of their name, we only leave out close to 800 other variations this way, and can add 14,131 additional pairs. These are labelled as 1 because they refer to the same person.

4   Models / Metrics

The baseline semantic similarity models considered are bi-encoder, BERTScore vanilla, and BERTScore trained, whereas the focus will be on cross-encoder (SAS) performance. Table 6 outlines the exact configurations used for each model.

A cross-encoder architecture (Humeau et al., 2020) concatenates two sentences with a special separator token and passes them to a network to apply multi-head attention over all input tokens in one pass. Pre-computation is not possible with the cross-encoder approach because it takes both input texts into account at the same time to calculate embeddings. A well-known language model that makes use of the cross-encoder architecture is BERT (Devlin et al., 2018). The resulting improved performance in terms of more accurate similarity scores for text pairs comes with the cost of higher time complexity, i.e. lower speed, of cross-encoders in comparison to bi-encoders. A bi-encoder calculates the embeddings of the two input texts separately by mapping independently encoded sentences for comparison to a dense vector space which can then be compared using cosine similarity. The separate embeddings result in higher speed but reduced scoring (Chen et al., 2020). In our work, both cross- and bi-encoder architectures are based on Sentence Transformers (Reimers and Gurevych, 2019b).

Risch et al. (2021) used a separate English and German model for the cross-encoder because there is no multi-lingual cross-encoder implementation available yet. We use BERTScore implementation from (Zhang et al., 2019) For BERTScore trained, the last layer representations were used, while for vanilla type BERTScore, only the second layer as per Figure 3. BERTScore vanilla is based on bert-base-uncased for English (SQuAD and NQ-open) and deepset's gelectra-base (Chan et al., 2020) for German (GermanQuAD), whereas BERTScore trained is based on the *multi-lingual* model that is used by the bi-encoder (May, 2020). BERTScore trained outperforms SAS for answer-prediction pairs without lexical overlap, the largest group in NQ-open, but neither of the models perform well on names. New name pairs are used to train the Sentence Transformer, which can be found in (Mustafazade and Ebbinghaus, 2021).

5   Analysis

To evaluate the shortcomings of lexical-based metrics in the context of QA, we compare BLEU, ROUGE-L, METEOR, F_1 and the semantic answer similarity metrics, i.e. Bi-Encoder, BERTScore

vanilla, BERTScore trained, and Cross-Encoder (SAS) scores on evaluation datasets. To address the second hypothesis, we delve deeply into every single dataset and analyse for disagreements with human judgement. As can be observed from Table 2 and Table 3, lexical-based metrics show considerably lower results than any of the semantic similarity approaches.

| | SQuad | | | | NQ-open | | | |
| | $F_1 = 0$ | | $F_1 \neq 0$ | | $F_1 = 0$ | | $F_1 \neq 0$ | |
Metrics	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLEU	0.00	0.00	0.17	0.16	0.00	0.00	0.05	0.05
ROUGE-L	0.04	0.04	0.54	0.46	0.16	0.16	0.46	0.38
METEOR	0.21	0.20	0.46	0.38	0.15	0.15	0.18	0.14
F1-score	0.00	0.00	0.58	0.50	0.00	0.00	0.41	0.34
Bi-Encoder	0.37	0.30	0.68	0.57	0.21	0.17	0.45	0.35
f_{BERT}	0.13	0.11	0.60	0.49	0.17	0.14	0.14	0.11
f'_{BERT}	0.39	0.32	0.69	0.57	0.23	0.18	0.45	0.35
SAS	0.36	0.29	**0.74**	**0.61**	0.20	0.16	**0.65**	**0.51**
New Bi-Encoder	0.39	0.32	0.69	0.57	0.25	0.20	0.50	0.39
$\tilde{f}_{BERT}$	**0.40**	**0.32**	0.70	0.58	**0.26**	**0.21**	0.51	0.40

Table 2: Spearman's, and Kendall's rank correlations of annotator labels and automated metrics on subsets of SQuAD and NQ-open. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained, and $\tilde{f}_{BERT}$ is the new BERTScore trained on names.

BLEU lags behind all other metrics, followed by METEOR. Similarly, we found that ROUGE-L and F1 achieve close results. In the absence of lexical overlap, METEOR gives superior results than the other n-gram-based metrics in the case of SQUAD, but ROUGE-L is closer to human judgement for the rest. The highest correlations are achieved in the case of BERTScore based trained models, followed closely by bi- and cross-encoder models. The superior performance of SAS doesn't hold up for the correlation metrics other than Pearson. We observed that SAS score underperformed when $F_1 = 0$ compared to all other semantic answer similarity metrics and overperformed when there is some lexical similarity.

NQ-open is not only by far the largest of the three datasets but also the most skewed one. We observe that the vast majority of answer-prediction pairs have a label 0 (see Table 1). In the majority of cases, the underlying QA model predicted the wrong answer.

All four semantic similarity metrics perform considerably worse on NQ-open than on SQuAD and GermanQuAD. In particular, answer-prediction pairs that have no lexical overlap ($F_1 = 0$) amount to 95 per cent of all pairs with the label 0 indicating incorrect predictions. Additionally, they perform only marginally better than METEOR or ROUGE-L.

| | **GermanQuAD** | | | | | |
| | $F_1 = 0$ | | | $F_1 \neq 0$ | | |
Metrics	r	ρ	τ	r	ρ	τ
BLEU	0.000	0.000	0.000	0.153	0.095	0.089
ROUGE-L	0.172	0.106	0.100	0.579	0.554	0.460
F_1-score	0.000	0.000	0.000	0.560	0.534	0.443
Bi-Encoder	0.392	0.337	0.273	0.596	0.595	0.491
f_{BERT}	0.149	0.008	0.006	0.599	0.554	0.457
f'_{BERT}	0.410	0.349	0.284	0.606	0.592	0.489
SAS	**0.488**	**0.432**	**0.349**	**0.713**	**0.690**	**0.574**

Table 3: Pearson, Spearman's, and Kendall's rank correlations of annotator labels and automated metrics on subsets of GermanQuAD. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

In **SQuAD**, there are only 16 cases where SAS completely diverges from human labels. In all seven cases where SAS score is above 0.5 and label is 0, we notice that the two answers have either **a common substring** or could be used often in the same context. In the other 9 extreme cases when the label is indicative of semantic similarity and SAS is giving scores below 0.25, there are three **spatial translations**. There is an encoding-related example with 12 and 10 special characters each which seems to be a mislabelled example.

Overall, error analysis for GermanQuAD is limited to a few cases because it is the smallest dataset of the three and all language model based metrics perform comparably well - SAS in particular. Regardless, SAS fails to identify semantic similarity in cases where the answers are **synonyms or translations** which also include technical terms that rely on Latin. This is likely the case because SAS does not use a multilingual model. Text-based **calculations and numbers** are also problematic. SAS also fails to recognise **aliases or descriptions of relations** that point to the same person or object.

We also observe that similarity scores for answer-prediction pairs which include numbers, e.g. an amount, a date or a year, SAS, as well as BERTScore trained, diverge from labels. The only semantically similar entities to answers expected to contain a numeric value should be the exact value, not a unit more or less. Also, the position within the pairs seems to matter for digits and their string representation. For SAS that the pair of *11* and *eleven* has a score of 0.09 whereas the pair of *eleven* and
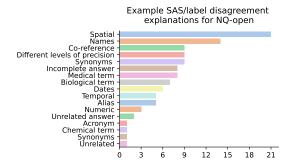
Figure 2: Subset of NQ-open test set, where SAS score < 0.01 and human label is 2, manually annotated for an explanation of discrepancies. Original questions and Google search has been used to assess the correctness of the gold labels.

11 has a score of 0.89.

Figure 2 depicts the major error categories for when SAS scores range below 0.25 while human annotations indicate a label of 2. We observe that entities related to names, which includes spatial names as well as co-references and synonyms, form the largest group of scoring errors. After correcting for encoding errors and fixing the labels manually in the NQ-open subset, totalling 70 samples, the correlations have already improved by about a per cent for SAS. Correcting wrong labels in extreme cases where SAS score is below 0.25 and the label is 2 or when SAS is above 0.5 and label is 0 improves results almost across the board for all models, but more so for SAS.

After removal of duplicates, sample with imprecise questions, wrong gold label or multiple correct answers, we are left with 3559 ground-truth answer/prediction pairs compared to 3658 we started with.

An example for the better performance on names when applying our new bi-encoder and SBERT trained models can be seen in Figure 4, where both models perform well in comparison to SAS and human judgement.

6 Conclusion

Existing evaluation metrics for QA models have various limitations. N-gram based metrics suffer from asymmetry, strictness, failure to capture multi-hop dependencies and penalise semantically-critical ordering, failure to account for relevant context or question, to name a few. We have found patterns in the mistakes that SAS was making. These include **spatial awareness**, **names**, **numbers**, **dates**, **context awareness**, **translations**, **acronyms**, **scientific terminology**, **historical events**, **conversions**, **encodings**.

The comparison to annotator labels is performed on answer pairs taken from subsets of SQuAD and GermanQuAD datasets, and for NQ-open we have a prediction and ground-truth answer pair. For cases with lexical overlap, ROUGE-L achieves comparative results to pre-trained semantic similarity evaluation models at a fraction of computation costs that the other models require. This holds for all GermanQuAD, SQuAD and NQ-open alike, discussed in more detail in Appendix E. Dataset size was one of the reasons why we focused more heavily on the NQ-open dataset. In addition, focusing on the other two would mean less strong evidence on how the metric will perform when applied to model predictions behind a real-world application. Furthermore, all semantic similarity metrics failed to have a high correlation to human labels when there was no token-level overlap, which is arguably the most important use-case for a semantic answer similarity metric as opposed to, say, ROUGE-L. NQ-open happened to have the largest number of samples that satisfied this requirement. Removing duplicates and re-labelling led to significant improvements across the board. We have generated a names dataset, which was then used to fine-tune the bi-encoder and BERTScore model. The latter achieves and beats SOTA rank correlation figures when there is no lexical overlap for datasets with English as the core language. Bi-encoders outperformed cross-encoders on answer-prediction pairs without lexical overlap both in terms of correlation to human judgement and speed, which makes them more applicable in real-world scenarios.

An element of future research would be further improving the performance on names of public figures as well as spatial names like cities and countries. Knowledge-bases, such as Freebase or Wikipedia, as explored in (Si et al., 2021), could be used to find an equivalent answer to named geographical entities. Numbers and dates which is the problematic data type in multi-lingual, as well as monolingual contexts, would be another dimension.

Acknowledgements

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Jiecao Chen, Liu Yang, Karthik Raman, Michael Bendersky, Jung-Jung Yeh, Yun Zhou, Marc Najork, Danyang Cai, and Ehsan Emadzadeh. 2020. DiPair: Fast and accurate distillation for trillion-scale text matching and pair modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2925–2937, Online. Association for Computational Linguistics.

Christophe Croux and Catherine Dehon. 2010. Influence functions of the spearman and kendall correlation measures. *Stat Methods Appl (2010) 19:497–515*.

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

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019a. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019b. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Philip May. 2020. T-systems-onsite/cross-en-de-roberta-sentence-transformer. *Hugging Face*.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, and et al. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 86–111. PMLR.

Farida Mustafazade and Peter F. Ebbinghaus. 2021. Evaluation of semantic answer similarity metrics. https://github.com/e184633/semantic-answer-similarity.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv:2104.12741*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv:1606.05250*.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. *arXiv preprint arXiv:2109.05289*.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv:2010.08240v2 [cs.CL]*.

Claudia Wagner. 2017. Politicians on wikipedia and dbpedia.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xian-Mo Zeng. 2007. Semantic relationships between contextual synonyms. *US-China education review*, 4:33–37.

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

A Error categories

B Implementation details

The original bi-encoder applied in (Risch et al., 2021) uses the multi-lingual T-Systems-onsite/cross-en-de-roberta-sentence-transformer (May, 2020) that is based on xlm-roberta-base which was further trained on an unreleased multi-lingual paraphrase dataset resulting in the model paraphrase-xlm-r-multilingual-v1. The latter was fine-tuned on an English-language STS benchmark dataset (Cer et al., 2017) and a machine-translated German STS benchmark. Similar to the bi-encoder approach, the English SAS cross-encoder model relies on cross-encoder/stsb-roberta-large which was trained on the same English STS benchmark. For German, a new cross-encoder model had to be trained, as there were no German cross-encoder models available. It is based on deepset's gbert-large (Chan et al., 2020) and trained on the same machine-translated German STS benchmark as the bi-encoder model, resulting in gbert-large-sts.

B.1 Hyperparameter tuning

We did an automatic hyperparameter search Table 5 for 5 trials with Optuna (Akiba et al., 2019). Note that cross-validation is an approximation of Bayesian optimization, so it is not necessary to use it with Optuna. The following set of hyperparameters was found to be the best: 'batch': 64, 'epochs': 2, 'warm': 0.45.

To be able to use BERTScore (Zhang et al., 2019), we made minor changes to accommodate for missing key-value pairs for the (May, 2020) model type. In Figure 3, we analyse SQuAD subset dataset of answers and we observe a similar phenomenon as in (Risch et al., 2021) when there is no lexical overlap between the answer pairs: the higher in layers we go in case of BERTScore trained, the higher the correlation values with human labels are. Quite the opposite is observed in the case of BERTScore vanilla, where it is either not as sensitive to embedding representations in case of no lexical overlap or correlations decrease with higher embedding layers. We did an automatic hyperparameter search for 5 trials with Optuna (Akiba et al., 2019). Note that cross-validation is an approximation of Bayesian optimization, so it is not necessary to use it with Optuna. We found the following best hyperparameters: 'Batch': 64, 'Epochs': 2, 'warm': 0.45.

Category	Definition	Question	Gold label	Prediction
Acronym	An abbreviation formed from the initial letters of other words and pronounced as a word	what channel does the haves and have nots come on on directv	OWN	Oprah Winfrey Network
Alias	Indicate an additional name that a person sometimes uses	who is the man in black the dark tower	Randall Flagg	Walter Padick
Co-reference	Requires resolution of a relationship between two distinct words referring to the same entity	who is marconi in we built this city	the father of the radio	Italian inventor Guglielmo Marconi
Different levels of precision	When both answers are correct, but one is more precise	when does the sympathetic nervous system be activated	constantly	fight-or-flight response
Imprecise question	There can be more than one correct answers	b-25 bomber accidentally flew into the empire state building	Old John Feather Merchant	1945
Medical term	Language used to describe components and processes of the human body	what is the scientific name for the shoulder bone	shoulder blade	scapula
Multiple correct answers	There is no single definite answer	city belonging to mid west of united states	Des Moines	kansas city
Spatial	Requires an understanding of the concept of space, location, or proximity	where was the tv series pie in the sky filmed	Marlow Buckinghamshire	in bray studios
Synonyms	Gold label and prediction are synonymous	what is the purpose of a chip in a debit card	control access to a resource	security
Biological term	Of or relating to biology or life and living processes	where is the ground tissue located in plants	in regions of new growth	cortex
Wrong gold label	The ground-truth label is incorrect	how do you call a person who cannot speak	sign language	mute
Wrong label	The human judgement is incorrect	who wrote the words to the original pledge of allegiance	Captain George Thatcher Balch	Francis Julius Bellamy
Incomplete answer	The gold label answer contains only a subset of the full answer	what are your rights in the first amendment	religion	freedom of the press

Table 4: Category definitions and examples from annotated NQ-open dataset.

Batch Size	{16, 32, 64, 128, 256}
Epochs	{1, 2, 3, 4}
warm	uniform(0.0, 0.5)

Table 5: Experimental setup for hyperparameter tuning of cross-encoder augmented BERTScore.

C Numeric errors

Presumably, numbers are difficult to evaluate (for all metrics), including for the underlying QA model of the predictions because we observe a high amount of label 0 cases where the prediction needed to be a number, however the labels in NQ-open are not entirely reliable, more so when they are 0. Therefore, we performed two experiments using NQ-open dataset where we remove all numbers from both ground-truth answers and predictions, and in the second experiment we remove numbers only from ground-truth answers. We further investigated whether numbers and digit are bringing the SAS performance down. We derived a new dataset from NQ-open where any row with a number in ground-truth is removed and then evaluated the four metrics. The removal of numbers further deteriorated the SAS performance, as evident in Table 7.

A similar experiment with SQuAD dataset shows a similar behaviour that SAS performed poorly compared to BERT-trained and Bi-Encoder metrics, but we did not observe a significant drop in performance when rows with numbers in ground-truth are removed from SQuAD since numbers are found only in 13% of SQuAD data compared to
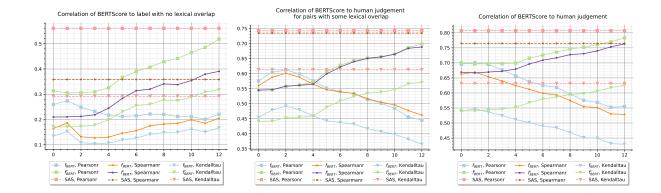
Figure 3: Pearson, Spearman's, and Kendall's rank correlations for different embedding extractions for when there is no lexical overlap ($F_1 = 0$), when there is some overlap ($F_1 \neq 0$) and aggregated for the SQuAD subset. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

	deepset/ gbert-large-sts	cross-encoder/ stsb-roberta-large	T-Systems-onsite/ cross-en-de-roberta -sentence-transformer	bert-base-uncased	deepset/ gelectra-base	Augmented cross-en-de-roberta -sentence-transformer
hidden_size	1,024	1,024	768	768	768	768
intermediate_size	4,096	4,096	3,072	3,072	3,072	3,072
max_position_embeddings	512	514	514	512	512	514
model_type	bert	roberta	xlm-roberta	bert	electra	xlm-roberta
num_attention_heads	16	16	12	12	12	12
num_hidden_layers	24	24	12	12	12	12
vocab_size	31,102	50,265	250,002	30,522	31,102	250,002
transformers_version	4.9.2	-	-	4.6.0.dev0	-	4.12.2

Table 6: Configuration details of each of the models used in evaluations. The architectures for the first two models and our own model follow corresponding sequence classification. T-systems-onsite model as well as our trained model follow `XLMRobertaModel`, and the other two - `BertForMaskedLM` & `ElectraForPreTraining` architectures respectively. Most of the models use absolute position embedding.

28% of NQ-Open data.

	NQ-open			
	$F_1 = 0$			$F_1 \neq 0$
Metrics	w_num	wo_num	w_num	wo_num
f_{BERT}	10.9	13.5	7.1	22.6
Bi-Encoder	13.3	13.1	29.9	25.8
f'_{BERT}	14.4	14.0	29.8	25.9
SAS	11.3	9.7	41.3	35.1

Table 7: Kendall's performance on NQ-open dataset, with and without numbers.

	SQuAD			
	$F_1 = 0$			$F_1 \neq 0$
Metrics	w_num	wo_num	w_num	wo_num
f_{BERT}	8.5	8.3	46.9	49.9
Bi-Encoder	29.2	31.4	56.0	56.8
f'_{BERT}	30.5	32.7	56.3	56.7
SAS	27.6	28.4	60.5	60.8

Table 8: Kendall's performance on SQuAD dataset, with and without numbers.

To investigate further, we created a new numbers dataset consisting of numbers as strings and their respective digit representation (digit/string and string/digit pairs) which were manually labelled as 1. These pairs were complemented by pairs of digits and their consecutive and preceding numbers, labelled as 0. Training the bi-encoder model on this dataset resulted in no change or worse performance, the cross-encoder model on the manually annotated dataset let to non-significant

improvements. Training the bi-encoder model on the dataset with a cross-encoder derived labels led to slightly less poor performance.

D Distribution of scores

Score distribution for SAS and BERTScore trained shows that SAS scores are heavily tilted towards 0 (see Figure 5 and Figure 6).

> **Question**: Who killed Natalie and Ann in Sharp Objects?
> **Ground-truth answer**: Amma
> **Predicted Answer**: Luke
> **EM**: 0.00
> F_1: 0.00
> **Top-1-Accuracy**: 0.00
> **SAS**: 0.0096
> **Human Judgment**: 0
> f_{BERT}: 0.226
> f'_{BERT}: 0.145
> **Bi-Encoder**: 0.208
> f_{BERT}: 0.00
> **Bi-Encoder (new model)**: -0.034

Figure 4: Representative example from NQ-open of a question and all semantic answer similarity measurement results.

E Model Complexity

We have scanned all metrics from Table 2 for time complexity on NQ-open as it is the largest evaluation dataset. Note that we haven't profiled training times as those are not defined for lexical-based metrics, but only measured CPU time for predicting answer pairs in NQ-open. N-gram based metrics are much faster as they don't have any encoding or decoding steps involved, and they take ∼10s to generate similarity scores. The slowest is the cross-encoder as it requires concatenating answers first, followed by encoding, and it takes ∼10 minutes. Concatenation grows on a quadratic scale with the input length due to self-attention mechanism. For the same dataset, bi-encoder takes ∼2 minutes. BERTScore trained takes ∼3 minutes, hence computational costs of BERTScore and bi-encoders are comparable. Additional complexity for all methods mentioned above except for SAS would be marginal when used during training on the validation set. Please note the following system description:

> **System =**'Darwin', **Release=**'20.6.0',
> **Machine=**'x86_64', **Total Memory=**8.00GB,
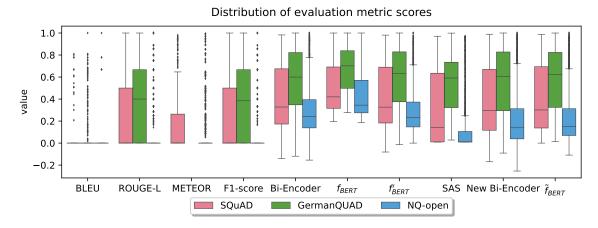> **Total cores=**4, **Frequency=**2700.00Mhz

Figure 5: Comparison of all (similarity) scores for the pairs in evaluation datasets. METEOR computations for GermanQuAD are omitted since it is not available for German.
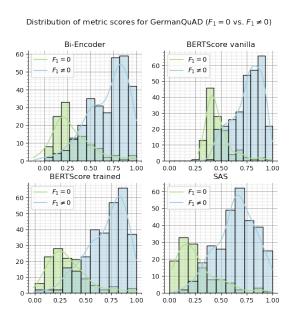

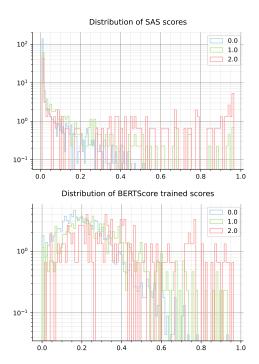Figure 6: Distribution of scores across labels for answer-pairs in GermanQuAD.

Figure 7: Distribution of SAS and BERT Trained scores for NQ-Open when $F_1 = 0$.