Evaluation of Semantic Answer Similarity Metrics

Farida Mustafazade **Peter Ebbinghaus**

Abstract

We propose a cross-encoder augmented BERTScore model for semantic answer similarity trained on our new dataset of non-common names in English contexts. There are several issues with the existing general machine translation (MT) or natural language generation (NLG) evaluation metrics, and question answering (QA) systems are indifferent in that sense. 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 lexical overlap is important not only to compare models fairly but also 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 superior results in the case of no lexical overlap.

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 concerns with the current models, but they also influence the development of a new generation of models. In addition, the consensus is to have a simple metric as opposed to a highly configurable and parameterizable one so that the development and the hyperparameter tuning do not add more layers of complexity to already complex QA pipelines. SAS, a cross-encoder-based metric for the estimation of semantic answer similarity (Risch et al., 2021), provides one such metric to compare answers based on semantic similarity.

The central objective of this research project is to analyse pairs of answers like in Figure 1 where (Risch et al., 2021)'s Semantic Answer Similarity (SAS) cross-encoder metric differs from human judgement and find patterns resulting in such errors. The main hypotheses that we will aim to test thoroughly through experiments are twofold. First of all, lexical-based metrics are not well suited in the automated evaluation of QA models. 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) as well as three lexical n-gram-based automated metrics. Then in Section 6, we thoroughly analyse the evaluation datasets described in the previous section and conduct an in-depth qualitative analysis of the errors. Since the human labels are discrete, while the model outputs come from a continuous distribution, our error analyses methodology involves using filtering on some thresholds to discretise the scores. Finally, in Section 4, we mitigate some of those issues by training the models further on external datasets for some categories and systematically generated datasets for others. In Section 7, we summarise our contributions and discuss ways in which as part of future work, the model could be improved and used in real-life applications.

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

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

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

definition of semantic and contextual synonyms. Min et al. (2021) noted that open-domain QA is inherently ambiguous because of the uncertainties in the language itself. They observed that automatic evaluation based on exact match (EM) fails to capture semantic similarity, which is observed in 60% of the ground-truth and prediction pairs in NQ-Open dataset. They, further, reported that 13%-17% of the predictions that fail automated evaluations are either definitely correct or partially correct. As a result, human evaluation improves the accuracy of the open-domain QA systems they tested on by 17%-54% compared to the EM used by the automated evaluation system.

Two out of four semantic textual similarity (STS) metrics that we analyse and the model that we eventually train depend on BERTScore, which is introduced in (Zhang et al., 2019). This metric is not one-size-fits-all. On top of choosing a suitable contextual embedding and model, there is an additional 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 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.

Risch et al. (2021) expand on this idea and further address the issues with existing general MT, NLG, which entails as well generative QA and extractive QA evaluation metrics. These include, but are not limited to, reliance on string-based methods, such as EM, F1-score, and top-n-accuracy. There is a more generic problem of evaluating STS. 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, one a subset of a well-known English SQuAD dataset (Rajpurkar et al., 2018), one German GermanQuAD dataset (Möller et al., 2021), plus a subset of NQ-open (Min et al., 2021).

Looking into error categories revealed problematic data types, where entities, particularly names, turned out to be the leading category. This is also what Si et al. (2021) try to solve using knowledge-base (KB) mined aliases as additional ground-truth answers. In contrast to Si et al. (2021), we generate a standalone names dataset from another dataset, described in greater detail in Section 3. The authors try to accomplish higher EM scores, which is defined as the maximum exact match on any of the correct answers in the expanded answer set.

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 used 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).

3 Data

We perform our analysis on the three manually annotated by three human raters subsets[1] of larger datasets provided by Risch et al. (2021). Unless specified otherwise, we will refer to the subsets by the 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** (Möller et al., 2021) 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., 2019), GermanQuAD was created with the goal of preventing strong lexical overlap between questions and answers. For this reason, questions were constantly rephrased via synonyms and altered syntax as well as complex questions were encouraged. SQuaD and GermanQuAD contain a pair of answers and a hand-labelled notation 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. (2019)'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 having a non-vague interpretation. The human annotators will attach a label `2` to all predictions that answer the question correctly are "definitely correct", `1` - "possibly correct", and `0` - "definitely incorrect". 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 two answers can be correct without semantic similarity even though only one correct answer is expected. In comparison to annotation for SQuAD and GermanQuAD, we conclude that 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.

In an attempt to further improve the NQ-open subset we build on (Risch et al., 2021)'s filter to only include question-answer pairs with one given ground-truth by manually re-labelling incorrect labels as well as filtering out vague questions. We focus on the most obvious cases described in Section 3 as well as we provide an improvement across all metrics in Section 6.

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.

	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.

3.2 Augmented datasets

In Section 6, we find that for NQ-open, in the majority of cases, the underlying QA model but also BERTScore are both failing in its prediction of names and their similarity to ground-truth answers respectively. To resolve this issue, we provide a new dataset that consists of 13,593 name pairs and employ the Augmented SBERT approach (Thakur et al., 2021) whereby we use the cross-encoder model to label a new dataset consisting of name pairs (only USA at the moment) to then train a bi-encoder model on the resulting dataset.

[1] https://semantic-answer-similarity.s3.amazonaws.com/data.zip

We deploy the already presented cross-encoder/stsb-roberta-large to label our new name pairs dataset to fine-tune the aforementioned T-Systems-onsite/cross-en-de-roberta-sentence-transformer which we then use in our Bi-Encoder as well as BERTScore trained semantic answer similarity metrics. These are commonly referred to as silver labels. The underlying dataset is created from the open source "Politicians on Wikipedia and DBpedia" dataset[2] which includes the names of 1,167,261 persons that have a page on Wikipedia and DBpedia. Out of these we only use those based in the U.S. as the English cross-encoder model has difficulties with labelling names that are not as common in the U.S. Besides, the questions in NQ-open are on predominantly U.S. related topics. Then we 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, resulting in a dataset where 75 per cent of the values have a score of 0.012 or smaller. Only 23 pairs receive a score higher than 0.5. To add to pairs which only by chance ended up describing the same individual, as in the case of *mark davis* and *tona rozum*, we make use of another benefit of the dataset: it includes different ways of writing a person's name, such as *gary a labranche* and *labranche gary* but also 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. In this context 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 a person but were mistakenly included in the dataset as, for example, the names of various members of Tampa Bay Rays minor league who have one page for all members.

We find that only adding the first variation of names, improves the overall performance of the bi-encoder model which is trained on the new name pairs dataset. Since most persons 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 to the aforementioned random pairs. All name variation pairs receive a manually annotated score of 1 because they are synonymous and refer to the same person.

[2] https://github.com/Kandy16/people-networks/tree/dbpedia-data/dbpedia-data/final_datasets

4 Models

The focus of our research lies on different semantic similarity metrics and their underlying models. As a human baseline, the original paper reports correlations between the labels by the first and the second annotator for subsets of SQuAD and GermanQuAD and omits these for the NQ-open subset since they are not publicly available. Maximum correlations achieved are just under 0.7 Spearman rank correlation for SQuAD and 0.64 for GermanQuAD, while maximum Kendall rank correlation achieved is just under 0.5 and 0.6 for SQuAD and GermanQuAD respectively.

The baseline semantic similarity models we have considered are bi-encoder, BERTScore vanilla, and BERTScore trained, whereas the focus will be on cross-encoder (SAS) performance. Table 5 outlines the exact configurations used for each model. For NQ-open specifically, METEOR and ROUGE-L remain essential baselines, too, because the lexical-overlap based metrics performs similar to the semantic similarity metrics on NQ-open.

The **bi-encoder** approach model is based on the sentence Transformer structure (Reimers and Gurevych, 2019). An advantage of the bi-encoder model architecture is that it calculates the embeddings of the two input texts separately. Thus, the embeddings for the ground-truth answer can be pre-computed early and compared later with the prediction answers embeddings. The model we use can be applied to English and German texts because it was trained in both languages. As described in (Risch et al., 2021), it is based on xlm-roberta-base, and it was further trained on an unreleased multi-lingual paraphrase dataset which resulted in the model paraphrase-xlm-r-multilingual-v1 which then in turn was fine-tuned on an English-language STS benchmark dataset (Cer et al., 2017) and a machine-translated German version of the same benchmark[3], resulting in the model T-Systems-onsite/cross-en-de-roberta-sentence-transformer.

While the bi-encoder approach calculates separate embeddings for a pair of answers, the **cross-encoder architecture** used for SAS (Risch et al., 2021) concatenates the answers with a special separator token. Pre-computation is not possible with the cross-encoder approach because it takes both input texts into account at the same time to calculate

[3] https://github.com/t-systems-on-site-services-gmbh/german-STSbenchmark

embedding, as opposed to calculating embedding separately.

Opposed to the bi-encoder (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. 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 as the bi-encoder model (Cer et al., 2017). For German, on the other hand, 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 that is used in experiments.

BERTScore (Zhang et al., 2019) uses Transformer-based language models to generate contextual embeddings, then match the tokens of the ground-truth answer and prediction (or the second answer in SQuAD and GermanQuAD), followed by creating a score from the maximum cosine similarity of the matched tokens. The implementation from Zhang et al. (2019)[4] is used for our evaluation, with minor changes to accommodate for missing key-value pairs for the T-Systems-onsite/cross-en-de-roberta-sentence-transformer model type and 12 respectively for BERTScore trained, as we are using the last layer for the trained version and only the second layer representations for vanilla type BERTScore. We are following Risch et al. (2021)'s approach of comparing two different BERTScore implementations with each other: **BERTScore vanilla** is based on the standard pre-trained BERT language model bert-base-uncased for English (SQuAD and NQ-open) as well as 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 approach, called T-Systems-onsite/cross-en-de-roberta-sentence-transformer.

In Section 6 we observe that BERTScore trained outperforms SAS for answer-prediction pairs without lexical overlap - which in addition amount to the largest group in NQ-open. Therefore as well as because cross-encoders are in general slower than bi-encoder approaches, we analyse the two approaches more thoroughly as well and find that bi-encoder and BERTScore trained don't perform well on names in particular. We therefore use our new name pairs dataset to train T-Systems-onsite/cross-en-de-roberta-sentence-transformer on it with the same hyperparameters as were used to train paraphrase-xlm-r-multilingual-v1 on the English-language STS benchmark dataset, which resulted in T-Systems-onsite/cross-en-de-roberta-sentence-transformer (see below).

[4] https://github.com/Tiiiger/bert_score

5 Experiments

To evaluate the shortcomings of lexical-based metrics in the context of question answering, we compare scores on evaluation datasets from BLEU, ROUGE-L, METEOR, F_1-score and the semantic answer similarity metrics, i.e. Bi-Encoder, BERTScore vanilla, BERTScore trained, and Cross-Encoder (SAS). To address the second hypothesis, we delve deeply into every single dataset and find differences between different types of answers, e.g. names and numbers. As can be observed from Figure 2, lexical-based metrics show considerably lower results than any of the semantic similarity approaches as described in Risch et al. (2021). In line with what the authors found, BLEU indeed 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. Regarding semantic answer similarity metrics, the highest correlations are achieved in the case of BERTScore based trained models, followed closely by bi- and cross-encoder models. We found some inconsistencies regarding the performance of the cross-encoder based SAS metric. 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.

Score distribution for SAS and BERT Trained shows that SAS scores are heavily tilted towards 0 as per Figure 6. We also analyse BERTScore trained thoroughly, however since the labels are not a continuous variable, we will rely heavily on the other two rank correlations, namely Spearman and Kendall's rank correlations, similar to (Chen et al.,

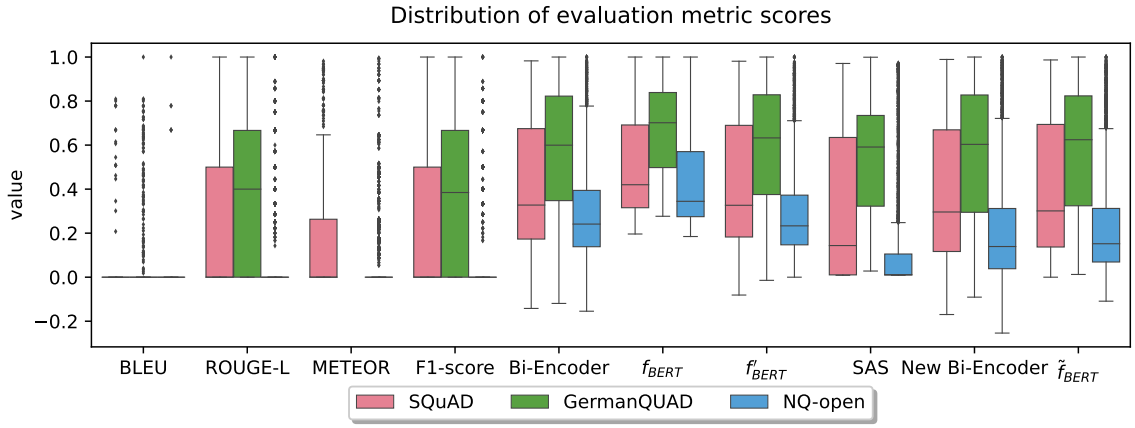

Figure 2: Comparison of all (similarity) scores for the pairs in evaluation datasets. METEOR computations for GermanQuAD are omitted since it is not available for German.

2019). Furthermore, using a combination of metric values, a considerable number of mislabelled pair cases, mainly in NQ-open, have been discovered.

	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 2: 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.

6 Error Analysis

This section is entirely dedicated to highlighting the major categories of problematic samples. We also include towards the end of subsection 6.3 details of the updated NQ-open dataset.

6.1 SQuAD

In Figure 3, we analyse SQuAD subset dataset of answers and we observe a similar phenomenon as in the original paper 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. 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 extreme when label is indicative of semantic similarity and SAS is giving scores below 0.25, totalling to only 9 cases overall, there are three **spatial translations**, in other words, it struggles with non-common names in English contexts, which we find even more evidence in the case of NQ-open. There is an encoding-related example with 12 and 10 special characters each which to our team seems to be a mislabelled example. We haven't identified any other consistent error categories for SQuAD.

6.2 GermanQuAD

Evaluating GermanQuAD more closely proves that SAS correlates the strongest to human annotation as Figure 5 shows.

For GermanQuAD, SAS fails to identify semantic similarity in cases where the answers are **synonyms or translations** which also include technical terms that rely on Latin (e.g. *vis viva* and *living forces* (translated) (SAS score: 0.5), *Anorthotiko Komma Ergazomenou Laou* and *Progressive Party of the Working People* (transl.) (0.04), *Nährgebiet* and *Akkumulationsgebiet* (0.45), *Zehrgebiet* and *Ablationsgebiet* (0.43)). This is likely the case because SAS does not use a multilingual model. Since multilingual models have not been implemented for cross-encoders yet, this remains an area for future research. The general hypothesis is supported by significantly higher BERTScore trained scores for the same pairs (0.43-0.5) apart from

	SQuad						NQ-open					
	$F_1 = 0$			$F_1 \neq 0$			$F_1 = 0$			$F_1 \neq 0$		
Metrics	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
BLEU	0.000	0.000	0.000	0.182	0.168	0.159	0.000	0.000	0.000	0.052	0.054	0.051
ROUGE-L	0.100	0.043	0.041	0.556	0.537	0.455	0.220	0.163	0.159	0.450	0.458	0.377
METEOR	0.398	0.207	0.200	0.450	0.464	0.378	0.233	0.152	0.148	0.188	0.179	0.139
F1-score	0.000	0.000	0.000	0.594	0.579	0.497	0.000	0.000	0.000	0.394	0.407	0.337
Bi-Encoder	0.487	0.372	0.303	0.684	0.684	0.566	0.294	0.212	0.170	0.454	0.446	0.351
f_{BERT}	0.249	0.132	0.108	0.612	0.601	0.492	0.156	0.169	0.135	0.165	0.142	0.112
f'_{BERT}	0.516	0.391	0.318	0.698	0.688	0.571	0.319	0.225	0.181	0.452	0.449	0.354
SAS	**0.561**	0.359	0.291	**0.743**	**0.735**	**0.613**	**0.422**	0.196	0.158	**0.662**	**0.647**	**0.512**
New Bi-Encoder	0.501	0.391	0.318	0.694	0.690	0.572	0.338	0.252	0.203	0.501	0.501	0.392
$\tilde{f}_{BERT}$	0.519	**0.399**	**0.324**	0.707	0.698	0.581	0.351	**0.257**	**0.208**	0.498	0.507	0.398

Table 3: Pearson, 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.

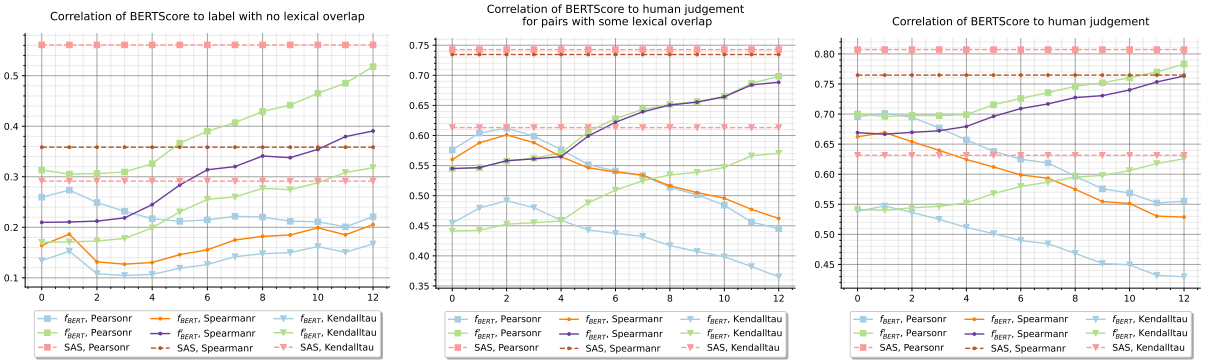

Figure 3: Pearson, Spearman's, and Kendall's rank correlations for different embedding extractions for when there is no lexical overlap ($F_1 = 0$), when there is some overlap ($F_1 \neq 0$) and aggregated for the SQuAD subset. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

Anorthotiko Komma Ergazomenou Laou (0.06) which is noticeable because the multilingual model used in BERTScore trained was fine-tuned on English and German but the training set of the underlying XLM-Roberta model included more than ten times as many Greek tokens than Latin ones (Conneau et al., 2020). Difficult as well are text-based **calculations and numbers** (transl.: *46th day before Easter Sunday* and *Wednesday after the 7th Sunday before Easter* (0.41), *24576 kB/s* and (transl.) *Execute three transmissions per microframe (125 µs) with up to 1024 bytes* (0.26)).

Apart from that, **aliases or descriptions of relations** which point to the same person or object: (*Thayendanegea* and *Joseph Brant* (0.028) are the same person but SAS fails to recognise it, BERTScore vanilla and BERTScore trained both find some similarity (0.36, 0.22); *Goring House* and *Buckinghams Haus* (0.29) refer to the same object but one is the official name, the other one a description of the same, again BERTScore vanilla and BERTScore trained identify more similarity (0.44, 0.37); *Aschraf Marwan* and *The husband of Nasser's daughter Mona* (translated) (0.17) where the first is a name and the latter a description of the same person, BERTScore vanilla discovers more similarity (0.38), BERTScore trained slightly less than SAS (0.15).

Overall, error analysis for GermanQuAD is limited to a few cases because it is the smallest dataset of the three. For this reason as well as because all metrics perform worst on it, we focus on our error analysis on NQ-open.

6.3 NQ-open

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). Thus, in the majority of cases, the underlying QA model predicted the

wrong answer. Apart from that, all four semantic similarity metrics perform considerably worse on NQ-open than on SQuAD and GermanQuAD - for answer-prediction pairs in particular that have no lexical overlap ($F_1 = 0$) which amount to 95 per cent of all answer-prediction pairs with the label `0` indicating incorrect predictions having no lexical overlap with the ground-truth answer. This is expected for wrong answers. All four metrics perform only slightly better than METEOR or ROUGE-L, thus adding no value via their semantic approach in the majority of all cases in NQ-open.

We also observe that for answer-prediction pairs which include numbers, e.g. an amount, a date or a year, SAS, as well as BERTScore trained, differ in many cases significantly from the label indication. By our definition of semantic similarity, 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: we observe for SAS that the pair of *11* and *eleven* has a score of 0.09 whereas the pair of *eleven* and *11* has a score of 0.89. Conducted experiments to improve the performance on numbers can be found in Appendix B.

Apart from numbers we find that for BERTScore trained as well as bi-encoder overall performance related to names is subpar: for *Alexandria Ocasio-Cortez* and *Kevin McCarthy* they indicate a similarity of 0.16 and 0.11 respectively whereas SAS correctly identifies no semantic similarity (0.01). A potential reason for identified similarity might consist in both names referring to politicians but as Figure 1 shows, the issue can be observed for more common names, too. We use our new name pairs dataset to fine-tune to bi-encoder model. Results are mentioned in the section below after addressing more general issues with NQ-open first.

After correcting for encoding errors and fixing the labels (one-way as of now) 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, as shown in Table 4. Figure 4 depicts the major error categories for when SAS scores range below 0.25 while human annotations indicate a label of 2, while Table 9 defines those categories along with providing a sample question, ground-truth answer and prediction triple. 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.

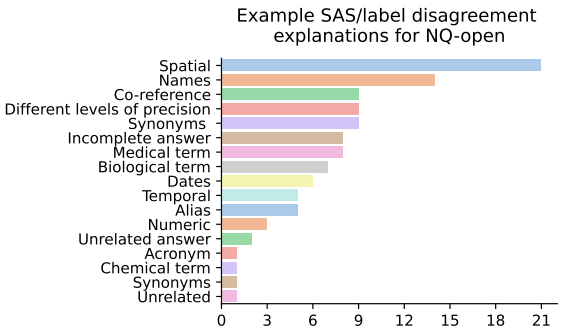

Figure 4: Subset of NQ-open test set, where SAS score < 0.01 and human label is 2, manually annotated for explanation of discrepancies. Original questions and Google search has been used to assess the correctness of the gold labels.

	NQ-open					
	$F_1 = 0$			$F_1 \neq 0$		
Metrics	r	ρ	τ	r	ρ	τ
ROUGE-L	37.7	23.9	24.5	10.7	10.9	11.4
METEOR	22.7	9.9	10.1	-3.7	3.4	3.6
Bi-Encoder	21.1	13.2	13.5	16.3	15.7	16.0
f_{BERT}	21.2	10.7	10.4	12.1	12.0	12.5
f'_{BERT}	20.4	12.4	12.7	15.7	15.1	15.3
SAS	25.6	17.3	17.7	18.7	17.2	17.4

Table 4: Improvements in correlation figures for NQ-open subset after re-labelling. All numbers reported in percentages.

We observe an improvement of 14 (Spearman) to 15 (Kendall) per cent on the newly created NQ-open subset for BERTScore trained and for bi-encoder we see an uplift of 19 per cent for both Spearman and Kendall when applying our fine-tuned sentence-transformer model (see Table 3). This refers only to pairs with no lexical overlap. For both bi-encoder and BERTScore trained the improvement for $F_1 \neq 0$ is smaller still in double-digits.

Figure 1 shows a representative example of where both BERTScore trained and bi-encoder have significantly improved towards identifying no semantic similarity where is none.

7 Conclusion

We have found a few 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**.

Currently, the comparison to annotator labels is performed on ground-truth answers taken from subsets of SQuAD and GermanQuAD datasets, and only for NQ-open we have a prediction and answer pair. There are two main reasons why we focus more on NQ-open. Firstly, 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. Secondly, effectively all semantic similarity metrics failed to have a high correlation to human labels, more so when there was no lexical overlap. NQ-open subset had quite a few issues associated with the labels as well as some other minor issues. Removing duplicates, re-labelling wrong labels one-way and re-calculating all metrics led to significant improvements across the board for semantic similarity metrics An element of future research would be improving the performance of all metrics on non-common names in English contexts and spatial names. The idea of using a KB, such as Freebase or Wikipedia, as explored in Si et al. (2021), could be used to find equivalent answer to named geographical entities as well. In addition, we found that bi-encoders are not only outperforming cross-encoders on answer-prediction pairs without lexical overlap but that they are also faster than cross-encoders which makes them more applicable in real-world scenarios. It is also relatively easy to enhance bi-encoder models as demonstrated by improved results after training on the names dataset. This could be essential for companies as well because models most probably won't understand the relationships between different employees and stakeholders mentioned in internal documents. Future research is needed for answer-prediction pairs without lexical overlap, the main use case of semantic answer similarity. As another dimension of future research and yet one more reason to have a preference towards BERTScore is the ability to use BERTScore as a training objective to generate soft predictions, allowing the network to remain differentiable end-to-end. Both SAS and BERTScore trained should be considered as metrics to evaluate the performance of a QA model and should be thoroughly compared in the context of the given dataset.

Acknowledgements

We would like to thank Ardhendu Singh, Julian Risch, Malte Pietsch and XCS224U course facilitators, Ankit Chadha in particular, as well as Christopher Potts for their constant support.

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.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv:1911.02116].*

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

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. 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.

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

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

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. 2019. 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.

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]*.

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 Model configurations

B 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 6.

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 28% of NQ-Open data. We observe a similar distribution of scores in Figure 6.

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.

	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 5: 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` archictectures respectively. Most of the models use absolute position embedding.

	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 6: 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 7: Kendall's performance on SQuAD dataset, with and without numbers.

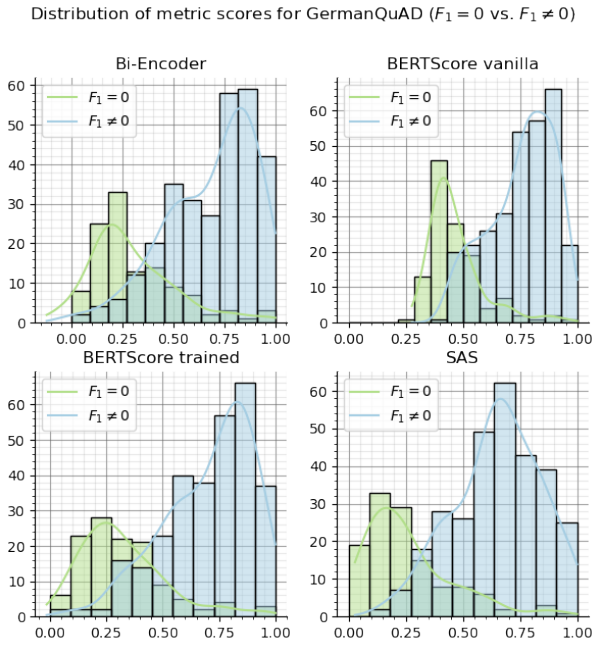

Figure 5: Distribution of scores across labels for answer-pairs in GermanQuAD.

C Distribution of scores

D Hyperparameter tuning

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.

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

Table 8: Experimental setup for hyperparameter tuning of cross-encoder augmented BERTScore.

E Error categories

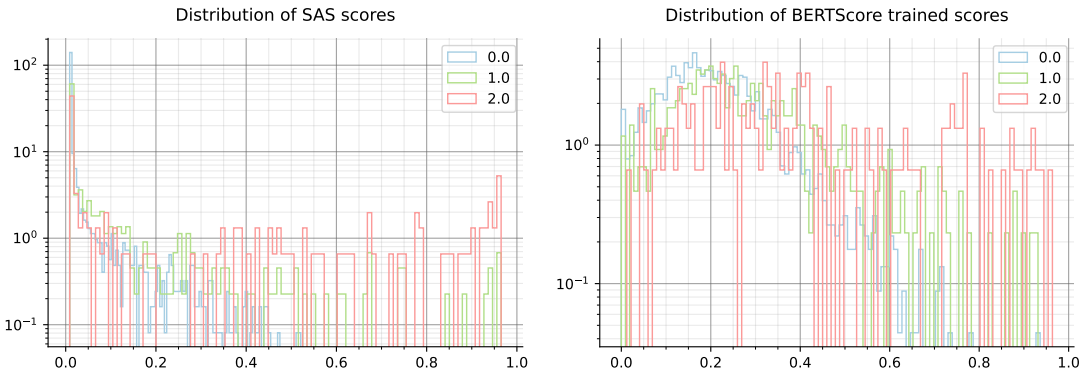

Figure 6: Distribution of SAS and BERT Trained scores for NQ-Open when $F_1 = 0$.

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 in Buckinghamshire	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 9: Category definitions and examples from annotated NQ-open dataset.