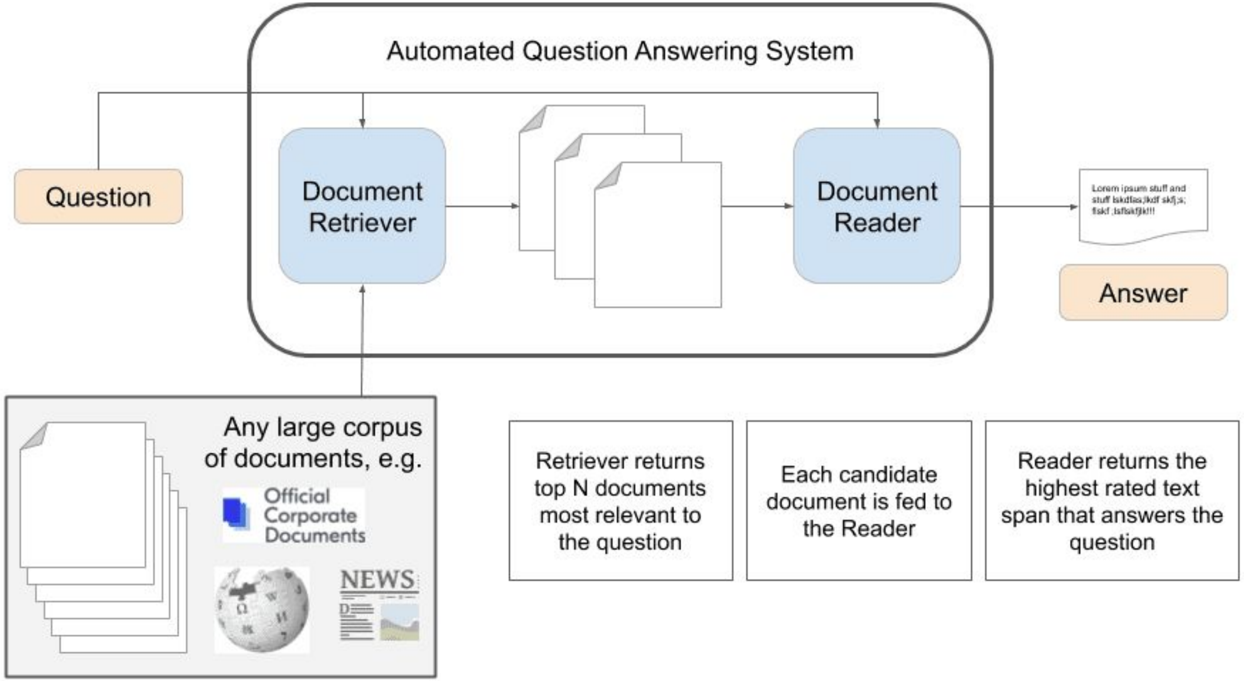
Evaluation of Semantic Answer Similarity Metrics

Overview

- Question Answering (QA)
- Semantic Answer Similarity (SAS)
- SAS model metrics
- Similarity score distributions
- Does it matter how deep the model is?
- Where do the models get confused?
- Contributions and future work

QA Pipeline


Automated Question Answering System
Question
Document Retriever
Document Reader
Answer
Any large corpus of documents, e.g.
Official Corporate Documents
NEWS
Retriever returns top N documents most relevant to the question
Each candidate document is fed to the Reader
Reader returns the highest rated text span that answers the question

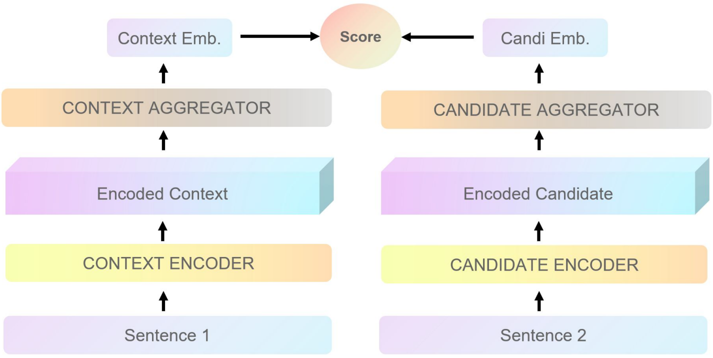
What is Semantic Answer Similarity (SAS)?

Question: How many plant species are estimated to be in the Amazon region?
Context: The region is home to about 2.5 million insect species, tens of thousands of plants, and some 2,000 birds and mammals. To date, at least 40,000 plant species [. . .]
Ground-Truth Answer: "40,000"
Predicted Answer: "tens of thousands"
Exact Match: 0.00
F1-Score: 0.00
Top-1-Accuracy: 0.00
SAS: 0.55
Human Judgment: 0.50

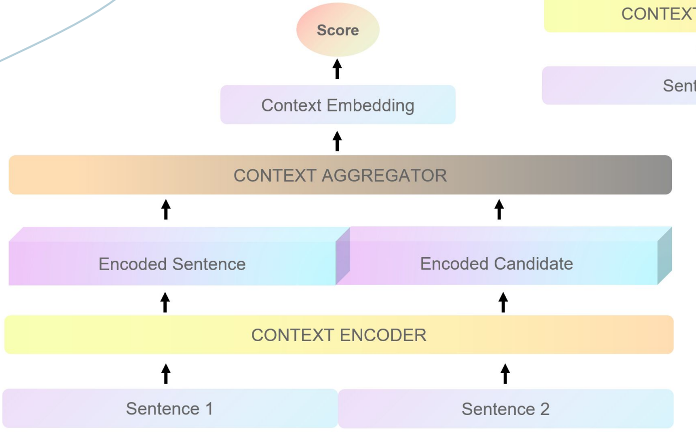
Question: Who killed Natalie and Ann in Sharp Objects?
Ground-truth answer: Amma
Predicted Answer: Luke
EM: 0.00
$\mathbf{F_1}$: 0.00
Top-1-Accuracy: 0.00
SAS: 0.0096
Human Judgment: 0
$\mathbf{f_{BERT}}$: 0.226
$\mathbf{f'_{BERT}}$: 0.145
Bi-Encoder: 0.208
$\mathbf{\tilde{f}_{BERT}}$: 0.00
Bi-Encoder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

Models under 🕵️

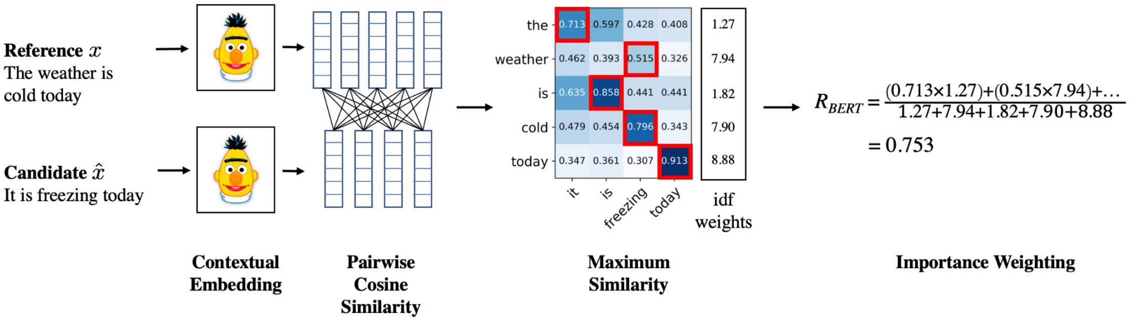
- Bi-Encoder



- Cross-encoder



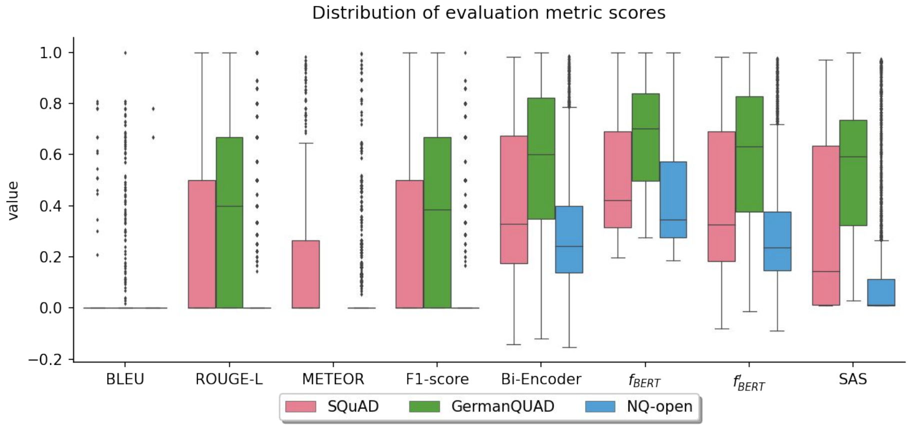
- BERTScore



Distribution of 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: 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.



Metric Comparison

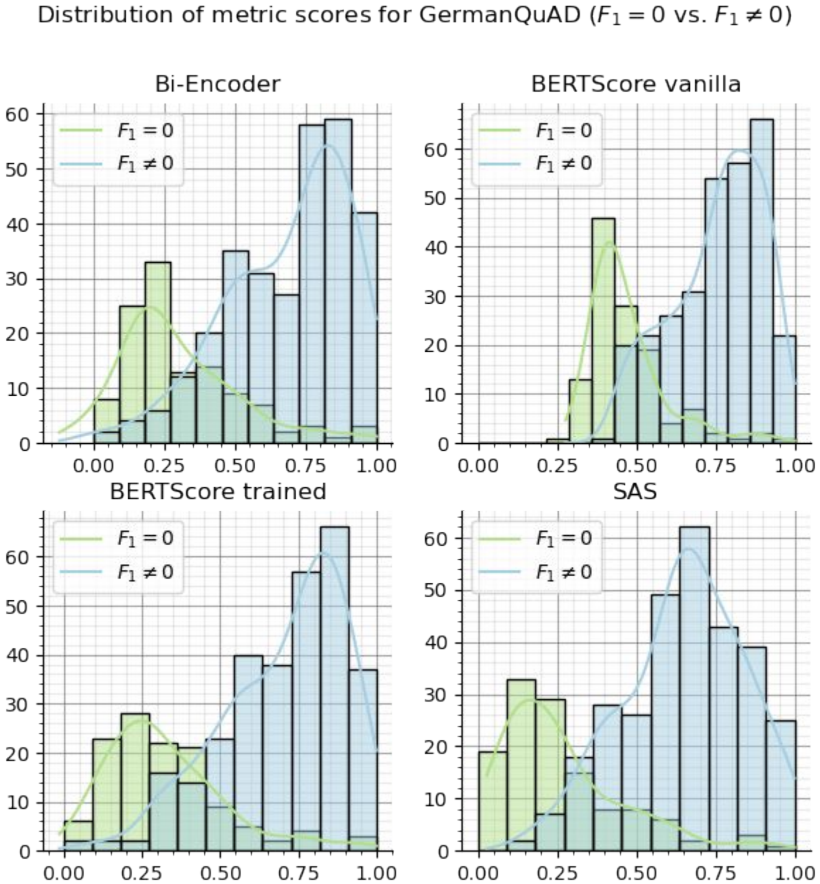
Metrics	SQuad						NQ-open					
	$F_1 = 0$			$F_1 \neq 0$			$F_1 = 0$			$F_1 \neq 0$		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
BLEU	0.000	0.000	0.000	0.182	0.168	0.159	0.000	0.000	0.000	0.052	0.054	0.051
ROUGE-L	0.100	0.043	0.041	0.556	0.537	0.455	0.220	0.163	0.159	0.450	0.458	0.377
METEOR	0.398	0.207	0.200	0.450	0.464	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 4: 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.

GermanQuAD

	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.



Embedding Layer Extraction

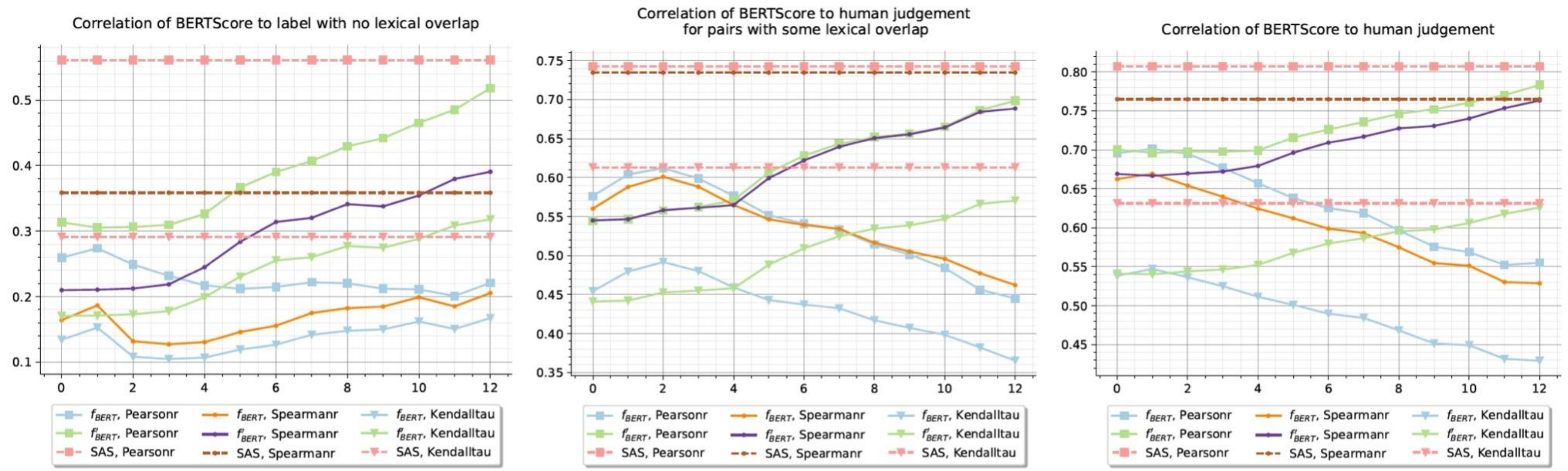


Figure 4: Pearson, Spearman's, and Kendall's rank correlations for different embedding extractions for when there is no lexical overlap ($F_1 = 0$), when there is some overlap ($F_1 \neq 0$) and aggregated for the SQuAD subset. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

Model Errors

- Names
  - Geographical names
  - Co-references
  - Aliases
- Translation
- Numeric types
  - Numbers & Dates
- Synonyms
- Scientific terminology

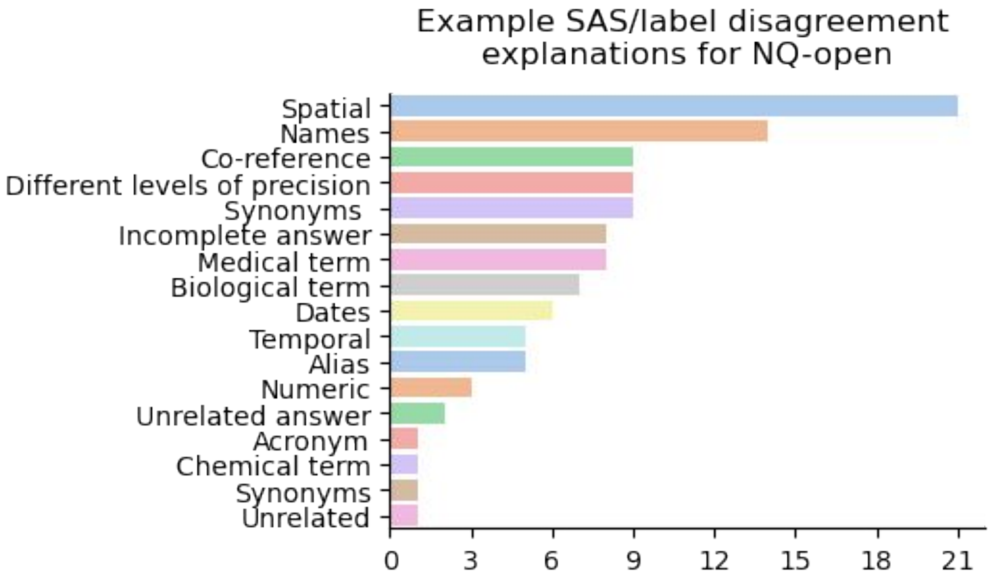
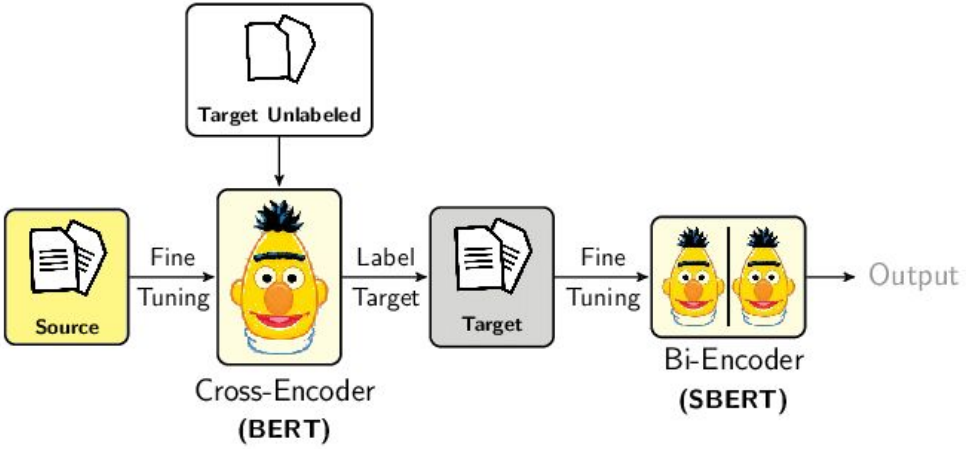


Figure 5: Subset of NQ-open test set, where SAS score < 0.01 and human label is 2, manually annotated for explanation of discrepancies. Original questions and Google search has been used to assess the correctness of the gold labels.

Contributions

- Improved (by finding and correcting for errors) a paper to be published on EMNLP 2021 (workshop paper)
- A new names dataset
- Relabelled NQ-open dataset
- An improved single configuration of BERTScore model



Future work

- Annotate prediction and ground-truth pairs (instead of two ground-truth answers)
- Improve the performance on non-common names in English contexts and spatial names
- Use BERTScore as a training objective to generate soft predictions, allowing the network to remain differentiable end-to-end

References

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

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Marc-Antoine Rondeau and Timothy J Hazen. 2018. Systematic error analysis of the stanford question answering dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 12–20.

Teven Le Scao and Alexander M. Rush. 2021. How many data points is a prompt worth? *arXiv:2103.08493*.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *arXiv arXiv:1003.1141*.

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

Team SASsy

Peter Ebbinghaus
Senior Manager SEO
Teufel Audio, Berlin, Germany

Farida Mustafazade
Quantitative researcher
GAM Systematic, Cambridge, UK