




Evaluation of Semantic Answer Similarity Metrics

June 25,
NLPD 2022

Overview

- 
- Challenges in Question Answering (QA)
 - Transformer-based SAS model metrics
 - Statistical overview of the datasets
 - Fine-tuning with Sentence BERT (SBERT)
 - Error categories of model predictions
 - Contributions and future work

QA model evaluation challenge

- Who makes more money: NFL or Premier League?
 - Ground-truth answer:
National Football League
 - Predicted Answer:
the NFL

QA model evaluation challenge

Question: Who makes more money: NFL or Premier League?

Ground-truth answer: National Football League

Predicted Answer: the NFL

EM: 0.00

F₁: 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

Semantic similarity

- Descriptions of the same entity, place, time, action, etc
- Different words or sentences having the same meaning in a given **context**
(Zeng, 2007)

What is Semantic Answer Similarity (SAS) model?

Question: Who makes more money: NFL or Premier League?

Ground-truth answer: National Football League

Predicted Answer: the NFL

EM: 0.00

F₁: 0.00

Top-1-Accuracy: 0.00

SAS: 0.9008

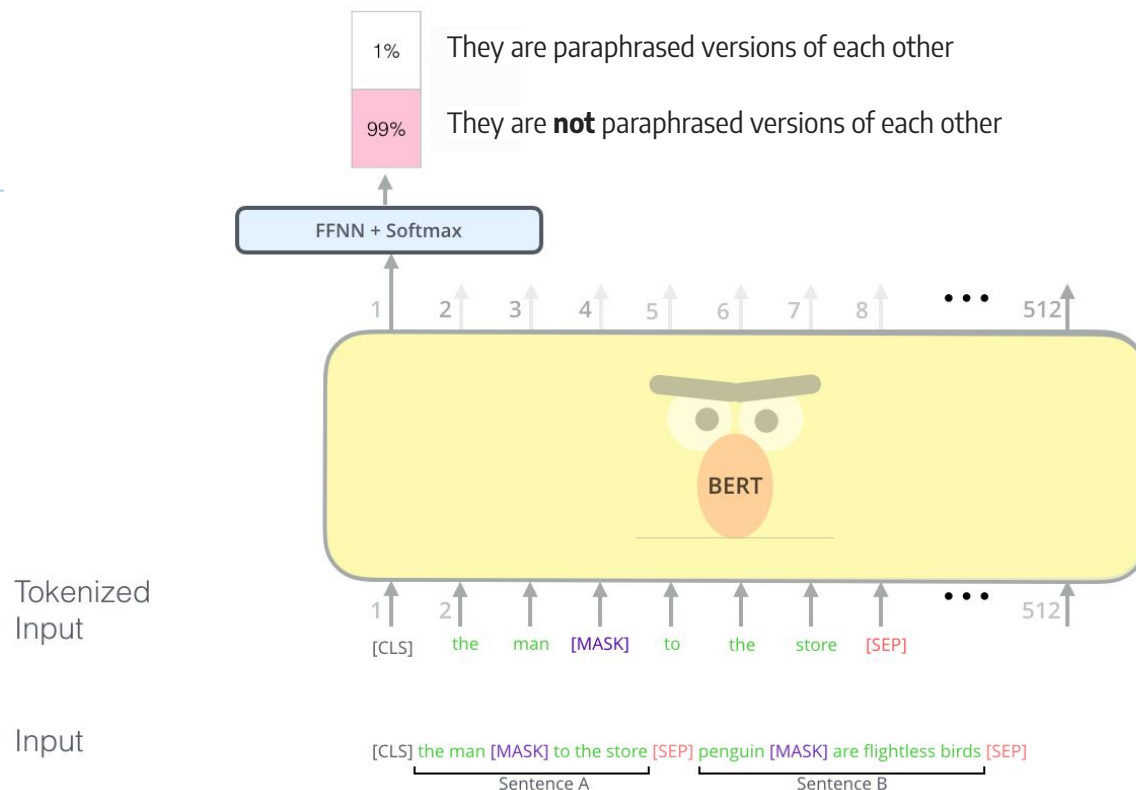
Human Judgment: 2 (definitely correct prediction)

f_{BERT}: 0.4317

f'_{BERT}: 0.4446

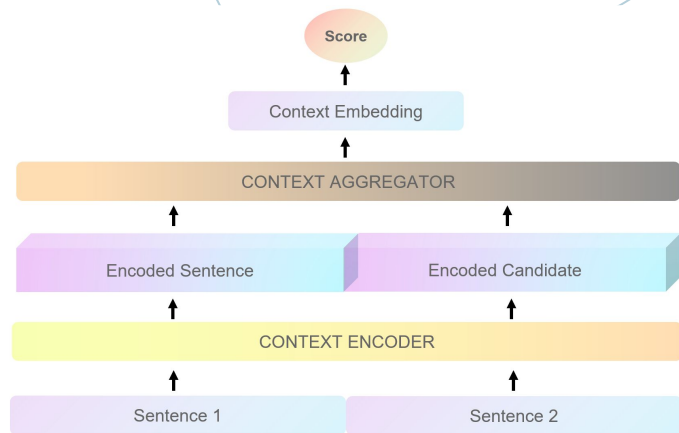
Bi-Encoder: 0.5019

Transformers, aka, BERT, the magic sauce



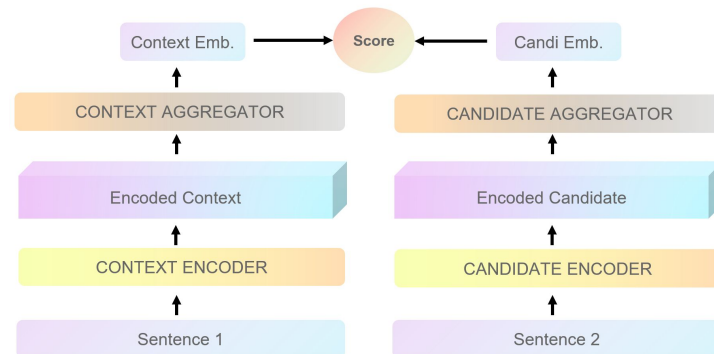
Models under

Bi-Encoder

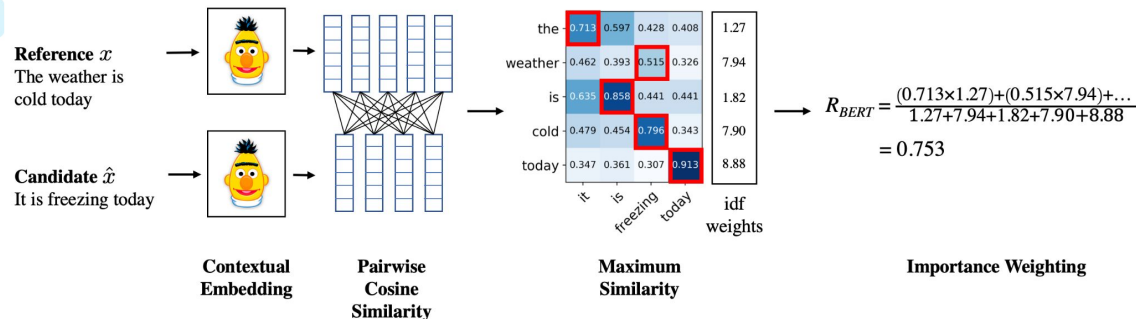


BERTScore

https://github.com/Tiiiger/bert_score



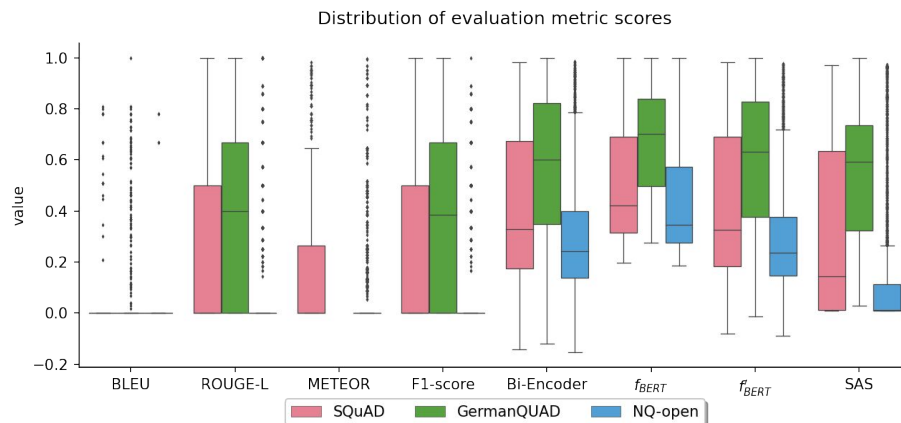
Cross-Encoder



Datasets

	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
$F_1 = 0$	565	124	3030
$F_1 \neq 0$	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.



Answer categories

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
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
Co-reference	Requires resolution of a relationship	Who is Marconi in "We built this city"?	The father of the radio	Italian inventor Guglielmo Marconi
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
Wrong gold label	The ground-truth label is incorrect	How do you call a person who cannot speak?	Sign language	Mute

Model Errors

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

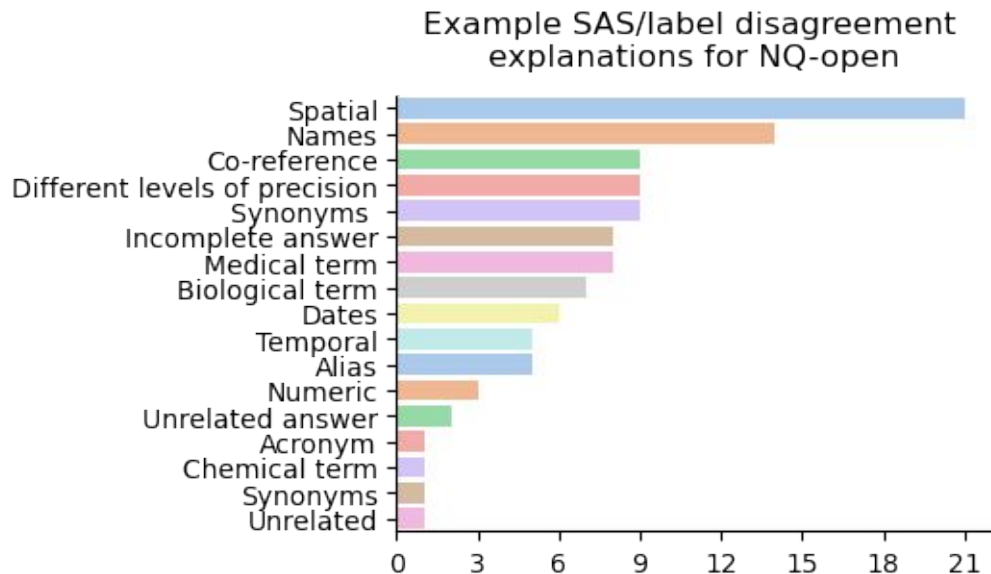


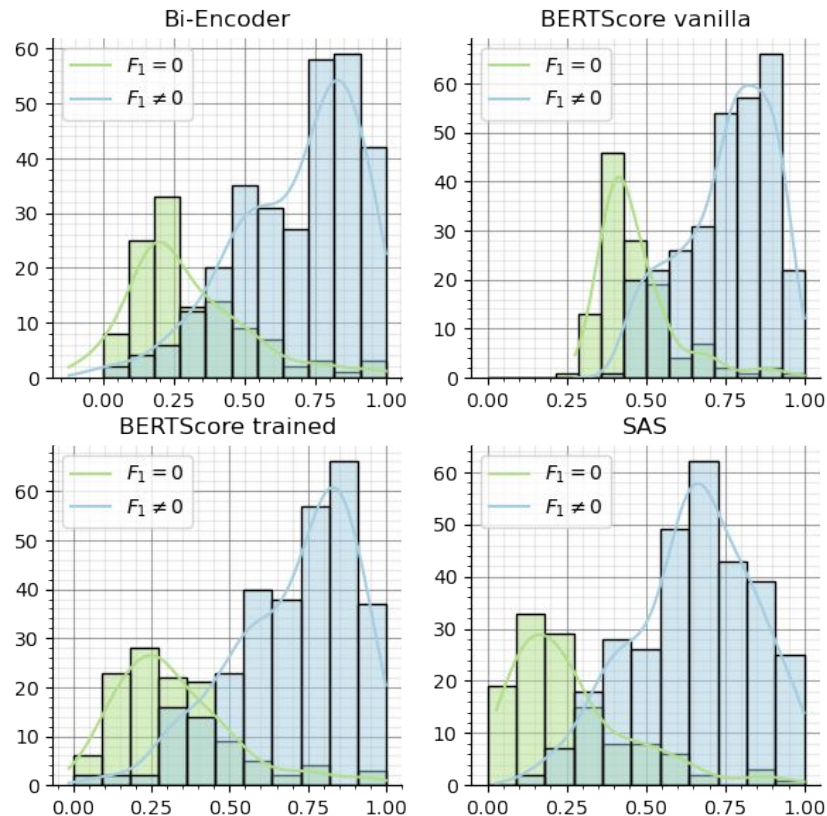
Figure 5: Subset of NQ-open test set, where SAS score < 0.01 and human label is 2, manually annotated for explanation of discrepancies. Original questions and Google search has been used to assess the correctness of the gold labels.

GermanQuAD

Metrics	GermanQuAD					
	$F_1 = 0$			$F_1 \neq 0$		
	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.

Distribution of metric scores for GermanQuAD ($F_1 = 0$ vs. $F_1 \neq 0$)



Embedding Layer Extraction

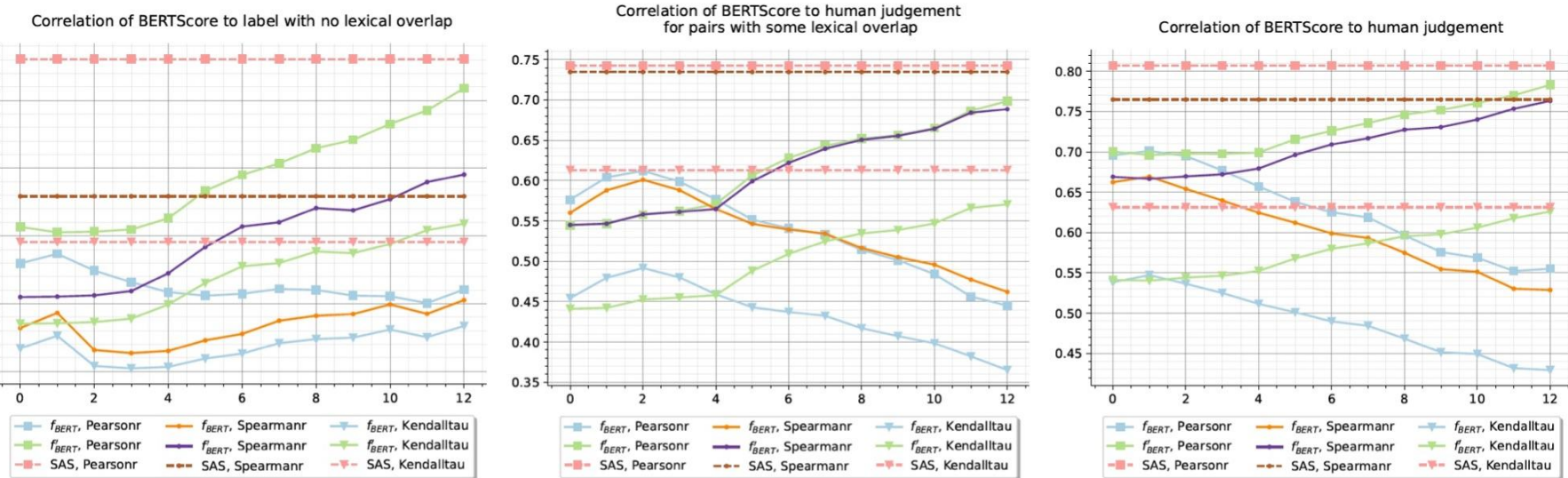
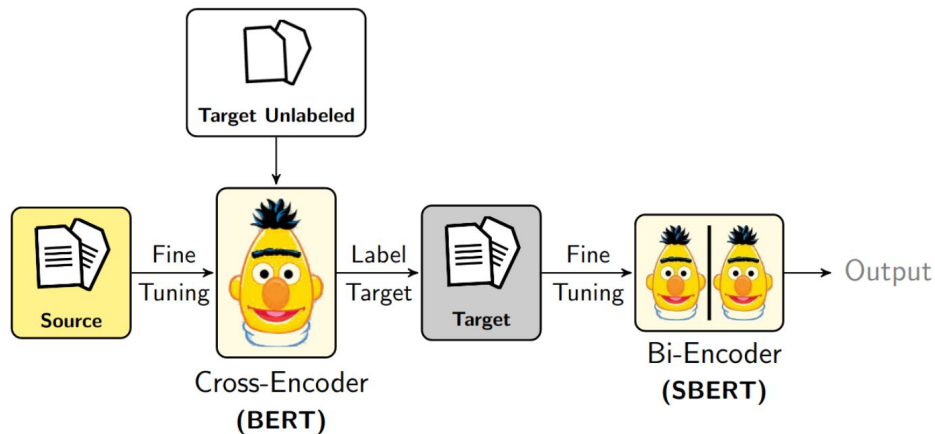


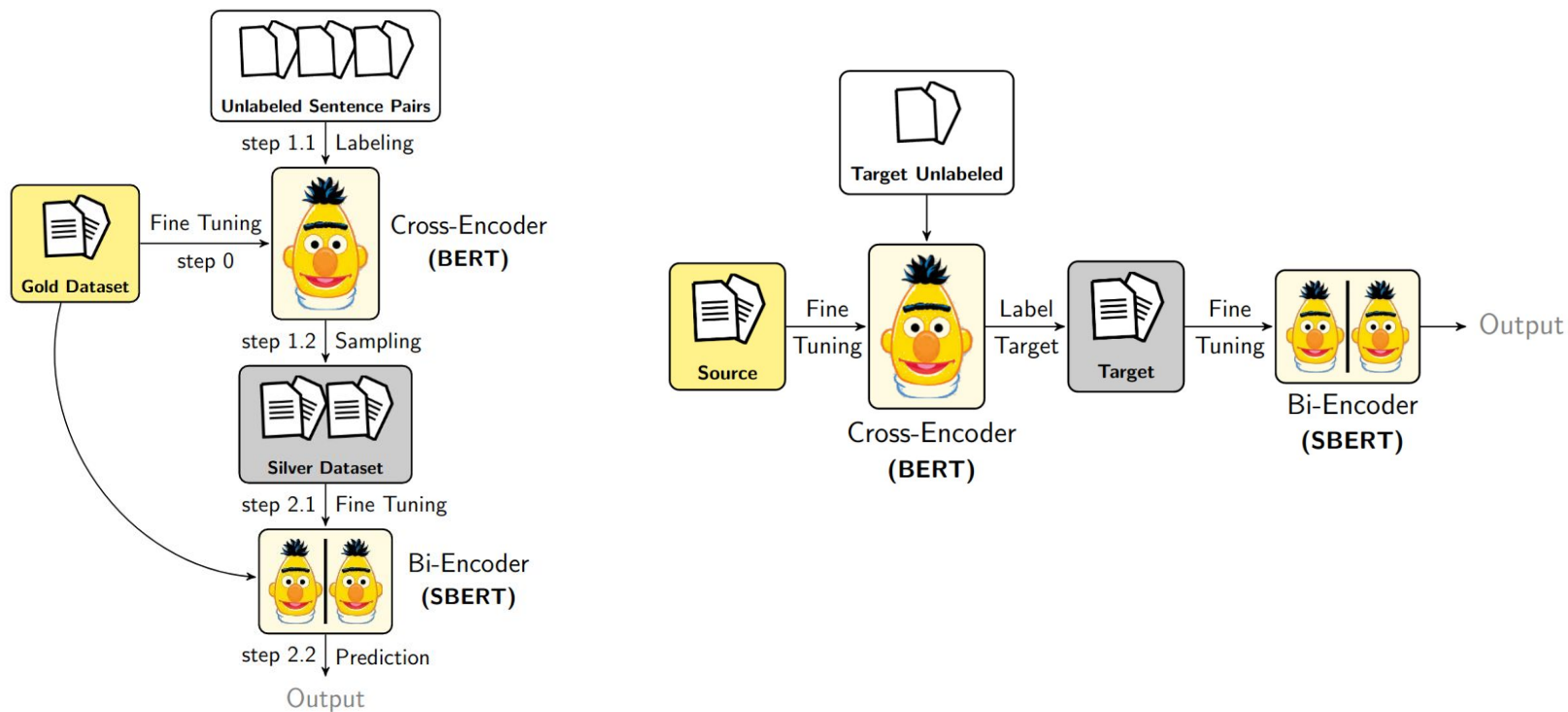
Figure 4: Pearson, Spearman's, and Kendall's rank correlations for different embedding extractions for when there is no lexical overlap ($F_1 = 0$), when there is some overlap ($F_1 \neq 0$) and aggregated for the SQuAD subset. f_{BERT} is BERTScore vanilla and f'_{BERT} is BERTScore trained.

Fine-tuning with Augmented SBERT

- New name pairs dataset based on Wagner, 2017
- US public figures on Wikipedia and DBpedia
- Sum of name pairs (random + aliases): ~40.000



Fine-tuning with Augmented SBERT



Fine-tuning with Augmented SBERT: Examples

Name1	Name2	Similarity Score	Random / Alias
Murray Feingold	Elvin Charles Stakman	0.011	Random (CE labelled)
Tona Rozum	Tona Rozum	0.968	Random (CE labelled)
Elvin Charles Stakman	Stakman Elvin C.	1	Alias (always 1)
Holly Ryder	Lisa Marie Abato	1	Alias (always 1)

Metric Comparison

Metrics	SQuad				NQ-open			
	$F_1 = 0$		$F_1 \neq 0$		$F_1 = 0$		$F_1 \neq 0$	
	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.

Contributions

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

Future work

- 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*.

About Us



Farida Mustafazade

Quantitative researcher

GAM Systematic, Cambridge, UK

farida.mustafazade.15@ucl.ac.uk



Peter F. Ebbinghaus

Team Lead SEO

Teufel Audio, Berlin, Germany

peter.ebbinghaus@posteo.de