

# SUM-QE: a BERT-based Summary Quality Estimation Model

Stratos Xenouelas<sup>1</sup>  
Marianna Apidianaki<sup>2</sup>

Prodromos Malakasiotis<sup>1</sup>  
Ion Androutsopoulos<sup>1</sup>



**A**

Oh look! A super box that makes a summary out of many documents. I won't have to read tons of articles any more!!!

Nice! But how do you know this summary is good?

Ah... you are right! What am I gonna do? I know what qualities a summary must have.

Can you show me?

There they are!

- Q<sub>1</sub>: Grammaticality
- Q<sub>2</sub>: Non Redundancy
- Q<sub>3</sub>: Referential Clarity
- Q<sub>4</sub>: Focus
- Q<sub>5</sub>: Structure & Coherence

**D**

And how well do you correlate with humans?

Pretty well, actually!! Come take a closer look at the table on the right.

I am not totally convinced. It seems you have difficulties with Q<sub>2</sub>...

Indeed, but this is half the truth. Come with me below.

	DUC-05			DUC-06			DUC-07			
	$\rho$	$\tau$	$r$	$\rho$	$\tau$	$r$	$\rho$	$\tau$	$r$	
Q1 Grammaticality	BEST-ROUGE	0.213	0.128	0.033	-0.049	-0.044	0.331	0.387	0.283	0.506
	GPT-2	0.678	0.511	0.637	0.391	0.280	0.593	0.780	0.586	0.675
	BERT-FR-LM	0.437	0.319	0.025	0.524	0.354	0.667	0.598	0.453	0.566
	BiGRU-ATT-S-1	0.119	0.079	0.116	0.263	0.182	0.459	0.119	0.085	0.494
	BiGRU-ATT-M-1	0.190	0.144	0.091	0.619	0.462	0.757	0.332	0.235	0.662
Q2 Non redundancy	BERT-FT-S-1	0.156	0.160	0.040	0.613	0.466	0.771	0.315	0.215	0.584
	BERT-FT-M-1	0.681	0.543	0.817	0.907	0.760	0.929	0.845	0.672	0.930
	BERT-FT-M-5	0.675	0.543	0.805	0.889	0.749	0.902	0.851	0.684	0.896
	BERT-FR-NS	0.185	0.130	-0.138	0.462	0.315	0.494	0.478	0.340	0.565
	BERT-FR-LM	0.437	0.319	0.025	0.524	0.354	0.667	0.598	0.453	0.566
Q3 Referential clarity	BIGRU-ATT-S-1	0.119	0.079	0.116	0.263	0.182	0.459	0.119	0.085	0.494
	BIGRU-ATT-M-1	0.190	0.144	0.091	0.619	0.462	0.757	0.332	0.235	0.662
	BIGRU-ATT-M-5	0.156	0.160	0.040	0.613	0.466	0.771	0.315	0.215	0.584
	BERT-FT-S-1	0.330	0.232	0.499	0.677	0.517	0.679	0.756	0.576	0.689
	BERT-FT-M-1	0.333	0.232	0.494	0.791	0.615	0.789	0.761	0.596	0.799
Q4 Focus	BERT-FT-M-5	0.712	0.564	0.802	0.883	0.732	0.925	0.840	0.680	0.902
	BEST-ROUGE	0.381	0.284	0.166	0.411	0.329	0.372	0.449	0.347	0.407
	BIGRU-ATT-S-1	0.150	0.110	0.153	0.355	0.242	0.644	0.433	0.321	0.533
	BIGRU-ATT-M-1	0.199	0.118	0.194	0.366	0.259	0.653	0.533	0.372	0.553
	BIGRU-ATT-M-5	0.154	0.097	0.160	0.493	0.371	0.691	0.645	0.462	0.657
Q5 Structure & Coherence	BERT-FT-S-1	0.645	0.471	0.578	0.814	0.636	0.853	0.873	0.704	0.902
	BERT-FT-M-1	0.664	0.491	0.642	0.776	0.608	0.842	0.893	0.745	0.905
	BERT-FT-M-5	0.791	0.621	0.739	0.875	0.710	0.911	0.818	0.636	0.867
	BEST-ROUGE	0.391	0.300	0.039	0.080	0.056	0.023	0.370	0.292	0.293
	BERT-FR-NS	0.200	0.153	-0.140	0.171	0.120	0.285	0.418	0.280	0.015

**E**

	DUC-05	DUC-06	DUC-07
Q1	3.77 (± 0.42)	3.58 (± 0.60)	3.54 (± 0.78)
Q2	4.41 (± 0.20)	4.23 (± 0.26)	3.71 (± 0.31)
Q3	2.99 (± 0.50)	3.11 (± 0.52)	3.20 (± 0.66)
Q4	3.15 (± 0.41)	3.60 (± 0.39)	3.30 (± 0.47)
Q5	2.18 (± 0.46)	2.39 (± 0.51)	2.42 (± 0.59)

Observe the table on my left. Q<sub>2</sub> has the highest manual scores and with the lowest standard deviation! Can you understand why this might be a problem?

Hmm... The differences between the systems are small and you struggle to put them in the correct order?

I see! It is quite clear now. Thank you Bert!

That's right! This is better illustrated in the diagram here.

**B**

Haaave you met Bert? He is super genius and can deal with many tasks! What do you think Bert?

Just give me data and the rest is up to me! See my proposal below.

(a)  $S_{Q_i} = R_i(h_i)$   
(b)  $S_{Q_i} = R(h)[i]$   
(c)  $S_{Q_i} = R_i(h)$   
 $i = 1, \dots, 5$   
 $R(h) = W^R h + b^R$

(a) Single Task (S-1) (b) Multi-Task-1 (M-1) (c) Multi-Task-5 (M-5)

$S_{Q1} S_{Q2} S_{Q3} S_{Q4} S_{Q5}$   $S_{Q1} S_{Q2} S_{Q3} S_{Q4} S_{Q5}$   $S_{Q1} S_{Q2} S_{Q3} S_{Q4} S_{Q5}$

$R_1 R_2 R_3 R_4 R_5$   $R$   $R_1 R_2 R_3 R_4 R_5$

$E_1 E_2 E_3 E_4 E_5$   $E$   $E$

Summary Summary Summary

SUM-QE Baseline

$\mathcal{E} = \text{BERT}$   $h = \text{CLS}$   $\mathcal{E} = \text{BiGRU-ATT}$   $h = \sum_i a_i h_i$

**C**

I can see you are using multi-task learning. Why is that?

It will help me learn richer representations and make better predictions, especially when the qualities are highly correlated.

See in the heatmaps around us how qualities are correlated.

DUC-05

DUC-06

DUC-07

**F**

I'm glad I could help! What are your plans now?

Well with your help I managed to predict linguistic quality.

I was wondering if I could learn how to predict content related aspects without human references.

I would also like to experiment with more datasets from different domains.

Oh! Oh! And I want to see if I can estimate the quality of other types of texts, coming, for instance, from NLG or sentence compression.

Nice! Just give me data when you're ready and let me see what I can do.

I think my work here is done. Take care!

Before you go... Do you have any friends I could meet?

Bye Bert! Thanks again for your help!

Don't forget to read the paper!

You can also find the code at <https://github.com/nlpaueb/SumQE>

Well yes! I could introduce you to Albert and Roberta. Say the word and it's done!

See you around!