HOLMS: Alternative Summary Evaluation with Large Language Models

Yassine Mrabet
mrabety@mail.nih.gov

Dina Demner-Fushman
ddemner@mail.nih.gov

U.S. National Library of Medicine
National Institutes of Health
8600 Rockville Pike, Bethesda, MD, 20894

Abstract

Efficient document summarization requires evaluation measures that can not only rank a set of systems based on an average score, but also highlight which individual summary is better than another. However, despite the very active research on summarization approaches, few works have proposed new evaluation measures in the recent years. The standard measures relied upon for the development of summarization systems are most often ROUGE and BLEU which, despite being efficient in overall system ranking, remain lexical in nature and have a limited potential when it comes to training neural networks. In this paper, we present a new hybrid evaluation measure for summarization, called HOLMS, that combines both language models pre-trained on large corpora and lexical similarity measures. Through several experiments, we show that HOLMS outperforms ROUGE and BLEU substantially in its correlation with human judgments on several extractive summarization datasets for both linguistic quality and pyramid scores.

1 Introduction

.

Generating human readable summaries of textual documents is posed to remain a key technology in our information era. Whether summarizing news, scientific articles, encyclopedias, or social media posts, the demand for a faster consumption of the most relevant information is expected to grow hand in hand with the amount of information available online.

A current bottleneck to a wider adoption of automatic summarizers is the lack of efficient solutions addressing both the relevance of the generated summaries and their linguistic quality. One component of this current limitation is the lack of efficient evaluation measures that address both aspects.

The two most cited and widely adopted measures in various summarization datasets and summarization challenges are ROUGE (Lin and Hovy, 2003; Lin, 2004) and BLEU (Papineni et al., 2002). While both measures have been shown to be highly efficient baselines in ranking summarization systems based on their average score (Dang and Vanderwende, 2007; Dang and Owczarzak, 2008; Dang and Owczarzak, 2009; Owczarzakg, 2010; Owczarzak and Dang, 2011), fewer studies have examined their relevance for individual summary ranking and their adequacy with linguistic quality and fluency.

With the new levels of performance achieved by neural language models in a variety of natural language processing tasks, several insights point towards the high potential of their contextual language encoding for language representation. Most of the state-of-the-art models such as T5 (Raffel et al., 2019), BERT (Devlin et al., 2019), GPT (Radford et al., 2018; Radford et al., 2019) and the Universal Sentence Encoder (Cer et al., 2018), are built from very large corpora, which reduces substantially the potential bias from using them to evaluate summaries in a restricted document set or benchmark.

On the other hand, the shallow lexical features of the original texts play a key role in extractive summarization and in pointer-generator approaches. It can also be argued that lexical similarities with gold summaries are implicitly capturing an important portion of the relevant semantics, especially at high similarity values.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5679–5688
Barcelona, Spain (Online), December 8-13, 2020

In this paper, we propose a new evaluation measure for summarization, called HOLMS, relying on both contextual neural representations and lexical similarities. To capture the salient indicators from each side more efficiently, HOLMS relies on a multi-dimensional Gaussian function combining a sequential similarity measure based on neural embeddings and ROUGE. We motivate and present each component in more details in section 3.

We evaluate HOLMS on five different summarization datasets by computing its correlation with human judgements for content relevance and linguistic quality, and show that HOLMS has higher Pearson, Spearman and Kendall correlations than state-of-the-art measures on both aspects.

In the next section, we discuss related works on summarization evaluation. We present HOLMS in details in section 3 and our experiments in section 4. Finally we discuss the results in section 5 before concluding.

2 Related Work

Summary evaluation was studied from several perspectives, including the similarity of the candidate summaries with reference human summaries, their intrinsic linguistic quality and coherence, and their relevance w.r.t. the original document (Cabrera-Diego and Torres-Moreno, 2018; Lloret et al., 2018). In this paper, we focus on extrinsic evaluations when human generated summaries are available as they offer both more specific parameters for the task, and available benchmarks with human-generated scores for automatically generated summaries.

	Lexical	Semantic	Gold sum.	Full text	Reference	Citations
ROUGE	✓	—	✓	—	(Lin and Hovy, 2003)	1,531
BLEU	✓	—	✓	—	(Papineni et al., 2002)	10,628
ROUGE-WE	✓	✓	✓	—	(Ng and Abrecht, 2015)	35
ROUGE-2.0	✓	✓	✓	—	(Ganesan, 2018)	15
AutoSummENG	✓	—	✓	—	(Giannakopoulos and Karkaletsis, 2011)	26
HOLMS	✓	✓	✓	—	—	—

Table 1: Related summarization evaluation metrics

Table 1, we present several summarization evaluation measures and their main characteristics. In terms of usage, ROUGE and BLEU stand out as the most cited measures, albeit a big portion of BLEU citations are likely from language translation papers, and not from research works on summarization.

ROUGE stands for Recall-Oriented Understudy of Gisting Evaluation (Lin, 2004). It allows evaluating system-generated summaries by comparing them with (ideal) summaries created by humans. It relies on computing the ratio of overlapping units between the two summaries and has several variants according to the unit type: e.g., unigrams, n-grams or skip-grams.

BLEU was designed for the evaluation of language translation systems (Papineni et al., 2002). It is a precision metric that counts a unigram as correct if it occurs in a reference translation and tackles redundancies by clipping the maximum number that a candidate word can be matched to its maximum number of occurrences in the gold/reference translation.

Other metrics have been proposed, mostly in the first decade of the twenty first century, in the context of the DUC and TAC challenges, with a few recent exceptions. For instance, ROUGE 2.0 was proposed in 2018 as an extension of ROUGE that relies on Wordnet synonyms (Pedersen et al., 2004) and main topics in addition to the original tokens. AutoSummENG is a lexical method relying on computing similarities between the n-gram graphs of candidate summaries and gold summaries. Edges represent proximity between two n-grams and are weighted by the number of co-occurrences of their vertices (connected n-grams) in a specified window of text.

One of the closest approaches to our work considered applying the ROUGE metric using word embeddings (Ng and Abrecht, 2015). Instead of computing word matching in a binary fashion as in ROUGE, the authors consider a word or n-gram similarity to be either 0 if the candidate word is out of vocabulary, or equal to the dot-product of their embeddings otherwise. While this approach made use of contextual embeddings, it still required the words to be present in both the gold summary and the

candidate summary, which makes it subject to the same limitation of ROUGE, i.e., the lack of semantic generalization that would allow matching synonyms and paraphrases. Results-wise, their approach was tested only on one dataset, and under-performed substantially ROUGE-SU4 and other ROUGE variants in terms of Pearson's correlation.

In contrast, HOLMS does not restrict the coverage of neural representations with lexical constraints, but explores new associative ways to get the best of both worlds. The embeddings component of HOLMS also relies on a sequential similarity function that grants increasingly more weight to (matching) sub-sequences with higher embeddings similarity. As far as we know, HOLMS is the first summarization evaluation measure that (a) takes into account embeddings similarity with a sequential perspective, and (b) uses a Gaussian function to combine lexical and embeddings-based similarities.

3 HOLMS

HOLMS stands for Hybrid Lexical and MOdel-based evaluation of Summaries. It relies on both deep contextual embeddings and lexical similarities.

The Embeddings-based Similarity (**ES**) used in HOLMS is computed in a sequential method. The intuition behind *ES* is that embeddings similarity between a word in a system summary and a word in a reference summary should be more pronounced if the contiguous words also have high embeddings similarity. Such perspective can also be extended to consecutive n-grams to capture conceptual similarity between sequences.

Concretely, to compute *ES*, a first step is to transform the input texts from both the gold and system summaries into two sets of consecutive n-grams: $G = \{g_1, \ldots, g_l\}$ for the gold summary, and $A = \{a_1, \ldots, a_m\}$ for a system summary. An embedding vector $V_L(x)$ is then generated for each n-gram x from a given language model L.

A second step is to compute the best distance value for a given gold n-gram g_i, $Dist_A(g_i)$, with a filtering method where each system n-gram $a \in A$ can only be used once as the best match for any given gold n-gram $g \in G$.

$$Dist_A(g_i) = Min_{\{a_j \in A_i\}} \ euc(V_L(g_i), V_L(a_j)) \tag{1}$$

With euc being the euclidean distance, and A_k the dynamic set computed by removing the previously matched n-grams in A.

The last step is inspired from *TextFlow* (Mrabet et al., 2017), and consists in using the (consecutive) distance values as coordinates on a curve and computing the area under curve as the overall distance. This is best shown with an example as in figure 1. The two example sentences, S, and G are from a news article[1], and were slightly modified for ease of presentation:

- G: "The man who ate the $120K banana art on the wall said that he was not sorry and that he was performing art by eating it."
- S: "An artist claimed he was performing art by eating a banana used as a center piece in the art work of a colleague."

Drawing the curve connecting the distance values as coordinates allows adding more weight to highly matching sub-sequences and less weight to weakly matched sub-sequences. The final value of *ES* is then computed as the complementary of the area under curve normalized by $|G| \times Max_{euc}$ to obtain values in the $[0, 1]$ range.

The Lexical Similarity used in HOLMS is the ROUGE-1 recall. There is no restriction on the lexical similarity measure, or ensemble of measures, that could be used for the lexical component

[1] https://cutt.ly/UrSccG7

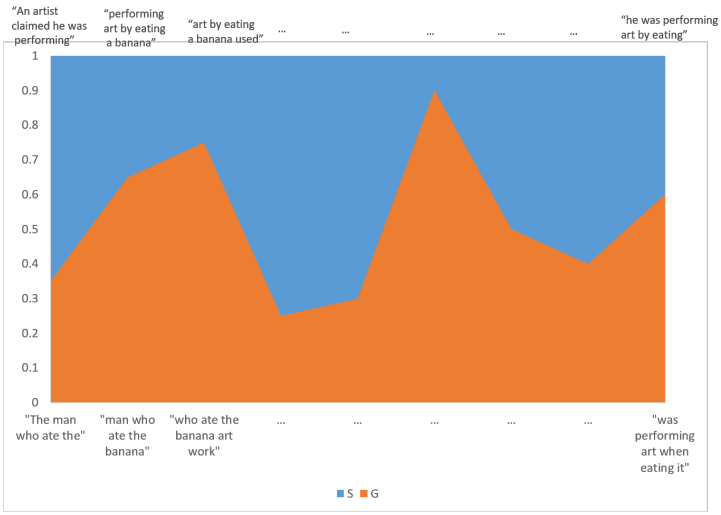

Figure 1: Illustration of the area under curve representing the HOLMS value.

of HOLMS. We picked ROUGE-1 for our first proof-of-concept as it has shown a good potential in ranking systems based on their average performance and is also widely used in the community. For the remainder of the paper, we will refer to ROUGE-1, ROUGE-2 ,and ROUGE-SU4 as R_1, R_2, and R_{SU4}.

The Combination of both aspects (lexical and neural) needs to take into account the implicit and relative links existing between shallow lexical similarities and embeddings-based similarity. To this effect, we build HOLMS using a bound three-dimensional Gaussian function that highlights further the summary pairs on which both measures agree, by promoting or downgrading exponentially strong agreements and strong disagreements. The function has its peak at 1 for perfect agreement on the quality of a summary and a low at 0 for total disagreement between the two measures (cf. figure 2).

$$HOLMS_{\{ES,R_1\}}(x,y) = \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right) \tag{2}$$

with x_0 and y_0 the coordinates of ES and R_1 peaks and σ_x^2 and σ_y^2 the respective spreads.

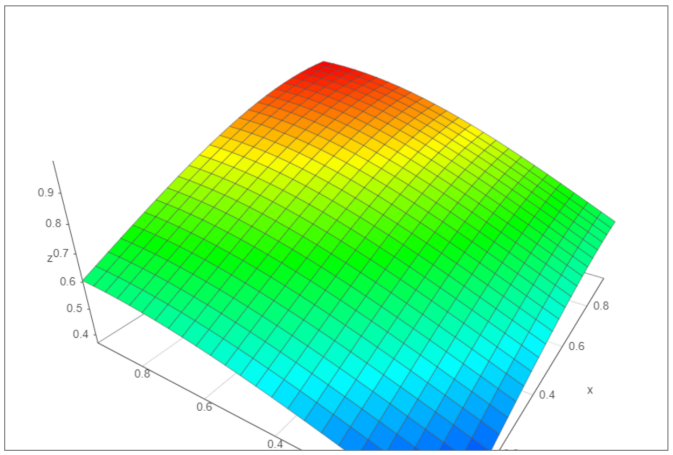

Figure 2: HOLMS: structure and value range (3D Gaussian peaks and spreads are both set to 1).

In practice, the values on the X and Y axes (representing the values of the evaluation measures) will never exceed the x and y coordinates of the peak, hence the bound nature of the function.

4 Evaluation

We used 4-grams as our units to compute the ES and HOLMS values. The peaks and spreads are set to 1. To compare HOLMS with a hybrid baseline, we also compute the correlation results for $linear$, a simple equal-weighted linear combination of ES and R_1.

We tested several variations of ES using three different sources for neural embeddings:

- The BERT large uncased model based on transformers and trained on large corpora such as Google books and Wikipedia using word masking and consecutive sentences classification.
- The universal sentence encoder (Cer et al., 2018) based on auto-encoders. We used the transformer-based version in our experiments instead of the slightly less accurate deep averaging version.
- Glove embeddings (Pennington et al., 2014) based on both word co-occurrence and local context windows.

In our preliminary experiment on the TAC 2011 dataset (Owczarzak and Dang, 2011), we observed that both the BERT embeddings and Glove embeddings underperformed substantially the embeddings from the universal sentence encoder (use) by more than 51%. We experimented with both the CLS class embeddings and the sum of the tokens embeddings from BERT, and with different vector dimensions for Glove. Subsequently, we used only the universal sentence encoder for our extended experiments on the 5 datasets. In future works, we plan to test more recent language models such as T5 (Raffel et al., 2019) and SciBERT (Beltagy et al., 2019).

We compute the correlations of HOLMS with human judgments of automatic summaries on 5 datasets from the summarization benchmarks introduced in the DUC and TAC challenges from 2007 to 2011(Dang and Vanderwende, 2007; Dang and Owczarzak, 2008; Dang and Owczarzak, 2009; Owczarzakg, 2010; Owczarzak and Dang, 2011). We use the three standard correlation measures: *Pearson correlation (P)*, *Spearman rank coefficient (S)*, and *Kendall's rank coefficient (K)*.

Each dataset consists of a set of source documents, a set of human-generated summaries for each document (called models in the challenges), and manual annotations of the relevance score for each candidate summary generated by the participating systems. The score for each candidate summary is computed as the average similarity between the system summary and the set of reference summaries. For content relevance, human assessors first selected the important content units (SCUs) then used the pyramid method (Passonneau et al., 2005) to score automatically the system-generated summaries based on the SCUs. In addition to content relevance scoring, the assessors also annotated the system-summaries with linguistic quality scores ranging from 1 to 5 for the DUC 2007, TAC 2008, and TAC 2011 datasets. Several scores ranging from 1 to 5 were first assigned to assess the system summaries according to five questions or aspects: grammaticality, non-redundancy, referential clarity, focus, structure, and coherence. The final linguistic quality score is then obtained by averaging the answers/scores given to all sub-questions. Further details about the data collection and annotation methods are described in the challenges overview papers and websites (Dang and Vanderwende, 2007; Dang and Owczarzak, 2008; Dang and Owczarzak, 2009; Owczarzakg, 2010; Owczarzak and Dang, 2011).

In the DUC and TAC editions, the ROUGE measure achieved state-of-the-art performance on its correlation with human judgments for both content relevance and linguistic quality. However, most evaluations of the TAC and DUC tracks relied only on average system scores and did not focus on ranking individual summaries.

Tables 2 and 3 present the correlations of BLEU, ROUGE, and HOLMS, with the linguistic quality

scores computed manually by the NIST assessors[2].

Table 2 presents the Pearson, Spearman, and Kendall correlations on the linguistic quality of individual summaries. Table 3 presents the correlations on the average linguistic quality scores of each participating system.

Tables 4 and 5 present the correlations with Pyramid scores computed manually by NIST assessors to measure the content relevance of the system summaries. Table 4 presents the Pearson's (P), Spearman's (S), and Kendall's (K) correlations on individual summaries. Table 5 presents the correlations on the average scores of each participating system.

Dataset	Corr.	BLEU	R_1	R_2	R_{SU4}	ES	Lin.	HOLMS
DUC 2007	P	-.059 (.017)	.216	.168	.216	.139	.186	**0.238**
	S	-.066 (.008)	.121	**.137**	.121	.094	.110	0.109
	K	-.058 (.008)	.090	**.101**	.090	.070	.080	0.076
TAC 2008	P	-.010 (.604)	.132	.128	.126	.117	.132	**.136**
	S	-.011 (.578)	.147	.144	.145	.145	.156	**.159**
	K	-.010 (.576)	.110	.108	.108	.111	.119	**.121**
TAC 2011	P	.003 (.896)	.361	.263	.294	.342	.365	**.376**
	S	.001 (.967)	.242	.260	.235	**.286**	.272	.276
	K	.001 (.965)	.183	.196	.178	**.215**	.206	.208
Average	P	-.022	<u>.236</u>	.186	.212	.199	.227	**.25**
	S	-.026	.170	<u>.180</u>	.167	.175	.179	**.181**
	K	-.022	.127	**.135**	.125	<u>.132</u>	**.135**	**.135**

Table 2: Correlations with Individual Summaries' Linguistic Quality

Dataset	Corr.	BLEU	R_1	R_2	R_{SU4}	ES	Lin.	HOLMS
DUC07	P	.569	.352 (.022)	.326	.352	.328 (.034)	.344 (.026)	**.753**
	S	.474	.427 (.005)	.301	.427	.353 (.022)	.390 (.011)	**.643**
	K	.336	.286 (.009)	.192	.286	.245 (.026)	.259 (.018)	**.451**
TAC08	P	.132 (.324)	**.434**	.430	.417	.383 (.003)	.416 (.001)	.411 (.001)
	S	.161 (.227)	.379 (.003)	**.433**	.404	.324 (.013)	.351 (.007)	.362 (.005)
	K	.105 (.248)	.268 (.003)	**.309**	.287	.225 (.013)	.243 (.007)	.254 (.005)
TAC11	P	.308 (.030)	.733	.705	.739	.729	.739	**.741**
	S	.210 (.144)	.341 (.015)	**.361**	.358	.331 (.019)	.342 (.015)	.325 (.021)
	K	.150 (.128)	.242 (.014)	**.263**	.263	.226 (.022)	.230 (.019)	.219 (.026)
Average	P	.336	<u>.506</u>	.487	.503	.480	.500	**.634**
	S	.281	.382	.365	<u>.396</u>	.336	.361	**.443**
	K	.197	.265	.255	<u>.279</u>	.232	.243	**0.310**

Table 3: Correlations with Average Systems' Linguistic Quality

[2](P: Pearson, S: Spearman, K: Kendall). All statistical p-values are strictly below 0.001 unless otherwise specified between brackets. Best results are highlighted in bold. Second best are underlined.

Dataset	Corr.	BLEU	R_1	R_2	R_{SU4}	ES	Lin.	HOLMS
DUC 2007	P	-.064	.365	.316	.365	.357	**.379**	0.376
	S	-.058	.340	.322	.340	.348	.358	**.42** (.095)
	K	-.051	.250	.235	.250	.257	.265	**.59** (.001)
TAC 2008	P	.000 (.982)	**.546**	.465	.495	.468	.541	.538
	S	.000 (.995)	.537	.486	.507	.480	**.547**	.539
	K	.000 (.999)	.376	.336	.353	.332	**.381**	.376
TAC 2009	P	-.008 (.679)	.665	.645	.665	.650	**.685**	.682
	S	-.002 (.916)	.630	.609	.630	.604	**.643**	.639
	K	-.002 (.925)	.460	.442	.460	.436	**.471**	.467
TAC 2010	P	.115	.690	.653	.690	.708	**.726**	.722
	S	.116	.715	.662	.715	.704	**.739**	.735
	K	.094	.525	.480	.525	.516	**.547**	.543
TAC 2011	P	.154	.630	.568	.606	.642	**.660**	**.660**
	S	.162	.596	.552	.566	.609	**.630**	.627
	K	.129	.425	.391	.401	.434	**.452**	.450
Average	P	.039	.579	.529	.564	.565	**.598**	.596
	S	.043	.564	.526	.552	.549	.583	**.592**
	K	.059	.407	.377	.398	.395	.423	**.485**

Table 4: Correlations with Individual Summaries' Pyramid Scores

Dataset	Corr.	BLEU	R_1	R_2	R_{SU4}	ES	Lin.	HOLMS
DUC 2007	P	.598	.615	.690	.615	.640	.632	**.769**
	S	.604	.762	.699	.762	.698	.738	**.876**
	K	.474	.573	.511	.573	.499	.547	**.698**
TAC 2008	P	.242 (.067)	.882	.907	.887	**.911**	.908	.905
	S	.109 (.417)	.865	**.908**	.884	.905	.889	.889
	K	.097 (.283)	.706	**.757**	.729	.724	.722	.724
TAC 2009	P	.427	.940	.911	.940	.944	**.955**	.954
	S	.350 (.009)	.894	**.952**	.894	.922	.923	.922
	K	.240 (.010)	.730	**.824**	.730	.787	.775	.772
TAC 2010	P	.369 (.015)	.928	.977	.928	.978	.968	**.985**
	S	.505	.950	.917	.950	.958	**.969**	**.969**
	K	.364	.824	.781	.824	.829	**.872**	**.872**
TAC 2011	P	.614	.954	.954	.975	.969	.972	**.976**
	S	.566	**.909**	.889	.888	.858	.897	.902
	K	.400	**.742**	.736	.724	.675	.728	.736
Average	P	.450	.864	.888	.869	.888	.887	**.917**
	S	.426	.876	.873	.876	.868	.883	**.912**
	K	.315	.715	.722	.716	.703	.729	**.760**

Table 5: Correlations with Average Systems' Pyramid Score

5 Discussion

Common observations. BLEU underperformed substantially all other measures in our experiments, except for the correlation on the average systems linguistic score in DUC 2007, which is likely due to shorter summaries. This could be caused mostly by the design of BLEU that was aimed at sentence-level machine translation. To capture only the relative performance of each summary and each system, we normalized the BLEU scores by the maximum score obtained by a system/summary for a given document. This improved the results (presented in section 4) compared with the correlation of the raw BLEU scores but was not enough to close the gap with the other evaluation measures. Another common observation is the relatively low level of correlation of all measures on linguistic quality when compared to content relevance correlations (pyramid scores). This can be explained in part by a higher level of subjectivity for linguistic quality inducing a higher assessor bias.

Evaluation of the linguistic quality of individual summaries. HOLMS outperformed the ROUGE variants, its embeddings-based component *ES* and the linear baseline in terms of Pearson correlation with a relative improvement ranging from 3% to 10.1%. The fact that HOLMS led to an improvement over both of its components (ES and R_1) and over the linear baseline suggests that:

- The embeddings-based sequential similarity ES and ROUGE-1 brought different but complementary perspectives on the linguistic quality of a given summary.
- HOLMS was better suited to take advantage of those different perspectives than the linear equal-weighted baseline.

In terms of Spearman's and Kendall's correlation factors, the picture was slightly more nuanced, with ES performing better on TAC 2011, HOLMS performing better on TAC 2008, and ROUGE-2 performing better on DUC2007. The macro-average is less prone to dataset-specific bias and showed that HOLMS performed better than all other measures.

Evaluation of the average linguistic quality of summarization systems. When analyzing the correlation values for average system scores, the benefits brought by HOLMS are even more substantial with a relative improvement of +25.2% on average (cf. table 3). The ablation study led to the interesting observation of the ROUGE variants outperforming the embeddings-based similarity and the linear combination. This suggests that: (1) lexical similarity measures can have a more relevant coarse-grained picture on system-level linguistic quality, and that neural language models are not necessarily better suited to rank extractive systems based on linguistic quality when gold summaries are available, and (2) neural language models still have a distinct perspective as shown by the better results obtained by HOLMS, and can make relevant hybrid methods more efficient at system ranking than lexical measures alone.

Evaluation of content relevance for individual summaries. The hybrid methods outperformed the ROUGE variants and the embeddings based similarity (ES) on content relevance with HOLMS performing better than the linear baseline on average in terms of Spearman and Kendall correlation factors, while maintaining comparable Pearson values (cf.table 4). In terms of components, the picture was less mixed, with ROUGE-1 performing slightly better on average on the three correlation measures. This shows that when gold summaries are available, performing better than lexical evaluation measures on content relevance is not as straightforward as computing embeddings similarities or taking into account sub-sequence similarities with measures like ES. Going to the embeddings space seemed to result in a relatively small loss of information when compared to ROUGE-1 for individual summary ranking. *One potential explanation is that at the scale of one (small) document, the precision of a restricted terminology is greater than the precision of large (contextual) language models*. This finding, together with the higher performance of HOLMS on the evaluation of content relevance for individual summaries, supports the theoretical effectiveness of pointer-generator approaches that combine abstractive and extractive functions.

Evaluation of content relevance for summarization systems. The improvement provided by HOLMS on the evaluation of average system performance for content relevance was noticeable over all baselines and datasets, with an average increase in correlation of 3.2%, 3.2%, and 4.2% respectively for Pearson's, Spearman's and Kendall's correlation factors. The lexical and neural components had comparable correlation results, with the linear baseline consistently under-performing HOLMS. This validates further the observation that the methods used to combine the lexical and neural language model spaces for summary evaluation can play a key role in improving systems evaluation when designed appropriately.

In a more general note, the proximity of the correlation results obtained by ES and the ROUGE variants raise an interesting question on the codification of language (or meaning) by neural embeddings and the extent to which their underlying representations provide an actual semantic generalization or rather a symbolic compression that remains tuned for data patterns that are both complex and shallow. What we can observe from our experiments is that the Gaussian combination through HOLMS outperforms both lexical and neural measures. This shows that embeddings do provide distinct and complementary information to the discrete/lexical information considered by ROUGE, but that the differences might not be as wide as could be expected. Moving forward, we are likely going to need either substantially different language representation spaces, or the integration of a different source of semantics such as knowledge bases and their associated neural graph models.

Limitations. In this paper, we did not address abstractive summarization due to the lack of sufficiently large abstractive summarization datasets with human judgments of content quality and relevance. We expect the neural-embedding component of HOLMS to have a higher impact in generative summarization approaches, and acquiring or building such datasets is one of our short-term objectives.

6 Conclusions

We presented a new summarization evaluation measure, called HOLMS , based on a sequential n-gram embeddings similarity and ROUGE. In our experiments on 5 summarization evaluation benchmarks, HOLMS performed consistently better on average than its individual components, BLEU, and a linear combination baseline. HOLMS can also be used as a framework, as many more variations can be tested, including the use of consecutive skip-grams as the input instead of n-grams, and the combination of sentence level similarity and n/skip-gram similarity. Moving forward, we think that summarization systems should be evaluated on a combination of discrete and contextual or semantic evaluation measures. Such extension fits naturally in HOLMS through the insertion of additional dimensions in its Gaussian function. Our results suggest that such combination is likely to bring a higher level of correlation with human assessments of both linguistic quality and content relevance.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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 3606–3611.

Luis Adrián Cabrera-Diego and Juan-Manuel Torres-Moreno. 2018. Summtriver: A new trivergent model to evaluate summaries automatically without human references. *Data & Knowledge Engineering*, 113:184–197.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *Proceedings of the Text Analysis Conference (TAC)*.

Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the tac 2009 summarization track. *Proceedings of the Text Analysis Conference (TAC)*.

Hoa Trang Dang and Lucy Vanderwende. 2007. Overview of duc 2007 tasks and evaluation results. In *Document Understanding Conference (DUC)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

George Giannakopoulos and Vangelis Karkaletsis. 2011. Autosummeng and memog in evaluating guided summaries. In *TAC*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post2Conference Workshop of ACL*.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52(1):101–148.

Yassine Mrabet, Halil Kilicoglu, and Dina Demner-Fushman. 2017. Textflow: A text similarity measure based on continuous sequences. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 763–772.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. *Proceedings of the Text Analysis Conference (TAC)*.

Karolina Owczarzakg. 2010. Overview of the tac 2010 summarization track. *Proceedings of the Text Analysis Conference (TAC)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rebecca J Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in duc 2005. In *Proceedings of the document understanding conference (DUC 05), Vancouver, BC, Canada*.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.