Summarization evaluation with ROUGE combinations

Velisarios Miloulis[1], **Prodromos Malakasiotis**[1,2], **Ioannis Pavlopoulos**[1,2], **and Ion Androutsopoulos**[1,2]

[1]Department of Informatics, Athens University of Economics and Business, Greece,
Patission 76, GR-104 34 Athens, Greece,
http://nlp.cs.aueb.gr

[2]Institute for Language and Speech Processing, Research Center Athena, Greece
Artemidos 6 & Epidavrou, GR-151 25 Maroussi, Athens, Greece
http://www.ilsp.gr

Abstract

The development of systems for summarization has been greatly enhanced by automatic evaluation methods. ROUGE is the most widely used method for evaluation, but the metric presents a number of shortcomings. In this paper we study a large number of ROUGE variants, as well as other metrics, and evaluate their performance on the task of summary evaluation. The evaluation takes the form of correlations with human assigned quality scores to system and human generated summaries. We use two different datasets in our experiments, both relevant to the task. We also introduce new approaches to the task, aimed at addressing the drawbacks presented by ROUGE. One method uses the centroid representation of the summaries to produce a score for their quality, while the other combines the available ROUGE metrics. The second method combines all the variants, and produces better results than any other single metric, while also being statistically significant according to pairwise Williams tests.

1 Introduction

Automatic evaluation of summaries is an area of study which was revolutionised after the introduction of ROUGE (Lin, 2004), derived from BLEU. The metric became, and still is, the most used evaluation metric, against which all other metrics are judged. At its basis, it relies on the automatic comparison of system-generated summaries with human-generated reference summaries. The shortcomings of the metric though are widely recognised by the community, and work has been done to evaluate the different ROUGE metrics, in an attempt to recognise the best variant of it . This body of work is significant, because of the large number of ROUGE variants and the fact that it unclear whether it can deal with paraphrasing and other complexities of Natural Language. The identified problems of the metric hinder its usefulness, so it is essential to create a clear picture of its performance in different scenarios. For example, metric variants that were in the past reported to be superior to others (Owczarzak et al., 2012) have been shown to be outperformed when compared to other possible metrics (Graham et al., 2015). To that end, in this study we expand the work of Graham et al. (2015) by evaluating the performance of ROUGE variants and other metrics used for summary evaluation. Specifically, we evaluate 96 ROUGE variants as well as 9 other metrics on the DUC-2004-2004 and BioASQ datasets, both used in summarization tasks. Furthermore, we propose the use of two different methods for evaluation with two permutations each for a total of 4 new methods. These methods aim to either address the issues with ROUGE, or replace ROUGE entirely. We present the results for both datasets and for all 109 metrics, along with statistical significance tests, as correlations with human assessment scores.

2 Related Work

Since its introduction in 2004, ROUGE (Lin and Hovy, 2003; Lin, 2004) has become the standard used method in the task of summarization evaluation. The metric can be expressed in many different variants, which leads to inconsistencies in reported results between published works, as researchers use different variants. The problems arising from the situation have been acknowledged by the community, and considerable work has been to mitigate them.

Owczarzak et al. (2012) tests six different variants, ROUGE-1;2;SU4, with stemming and binary stopword removal, using recall to compute the output score. They use the data from the TAC summarization track from 2008-2011, which uses Pyramid and Responsiveness as base human evaluation metrics. They conclude that ROUGE-2 with stemming and stopwords not removed provides the strongest agreement with human evaluation, even when considering statistical significance.

Rankel et al. (2013), tests a more extensive subset of ROUGE, ROUGE-1;2;3;4;L;W;SU4;BE-HM accuracy, on the task, using the same metrics as human evaluations. They single out a different single metrics as the best for the task than Owczarzak (2012). More importantly, they conclude that a *combination* of ROUGE variants outperform all other systems in AESOP task 2011.

Meanwhile, Ng et al. (2015), recognises the bias towards lexical similarity present in classic ROUGE variants and, proposes the use of word embeddings to capture semantic similarities. The metrics they propose are mutations of ROUGE-1;2;SU4, incorporating the word embeddings in the process of similarity computation for each variant. They show that one of the metrics they propose achieves state-of-the art performance in the AESOP 2011 task of TAC.

Finally, Graham et al. (2015) attempts a review of the whole process of metric evaluation. They evaluate 105 ROUGE variants, against human evaluation by correlation and explore the statistic significance of the difference between metrics, on DUC-2004. We base our study on the work of Graham extensively, and go into further detail about their methodology below.

3 Experiments

3.1 DUC-2004

We used the datasets provided for Tasks 2 and 5 of DUC-2004 (Over and Yen, 2004), and the results of that year's competition. The datasets include human created (model) summaries from newswires chosen by NIST staff, and system (peer) summaries consisting of baseline, manual or system submitted summaries. For each peer summary, a human assessor computed the summary coverage by identifying the peer units that express content of the corresponding human summary (matching PUs). The human assessor also provided an overall coverage estimate (E), defined as the proportion of model units (MUs) in a human summary expressed by a given peer summary. Furthermore, the human assessors were asked to rate the linguistic quality of each peer summary using 7 different criteria.

Graham used these datasets to analyse the current evaluation methodologies applied to summarization metrics. They combined the matching PUs, the overall coverage estimate E and the model units MUs to compute Human Coverage Scores (CS):

$$CS = \frac{\mid MatchingPUs \mid \times E}{\mid MUs \mid} \qquad (1)$$

They also calculated the Mean Linguistic Quality (MLQ) of each summary as the average of the scores of the 7 criteria. MLQ and CS were then combined into a single Human Assessment Score (HAS):

$$HAS = \frac{CS + MLQ}{2} \qquad (2)$$

3.2 BioASQ

BioASQ is a challenge on biomedical semantic indexing question answering. The challenge consists of two tasks, A and B. Task A examines large scale biomedical semantic indexing, while Task B examines biomedical semantic QA (Tsatsaronis et al., 2015).

Task B examines the ability of systems to annotate questions with concepts from relevant ontologies and return either exact or paragraph sized ideal answers. It is organised in two phases, A and

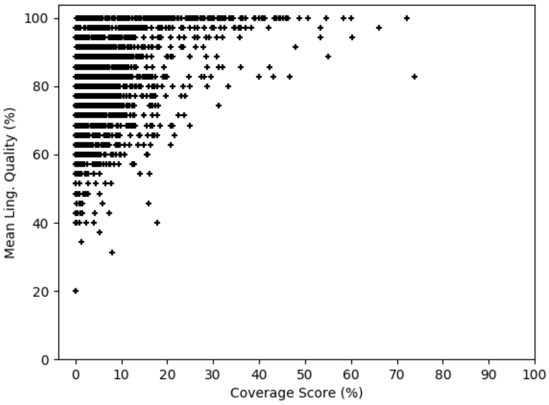

Figure 1: Scatterplot of mean linguistic quality (MLQ) and coverage scores (CS) for human assessments of summaries in DUC-2004.

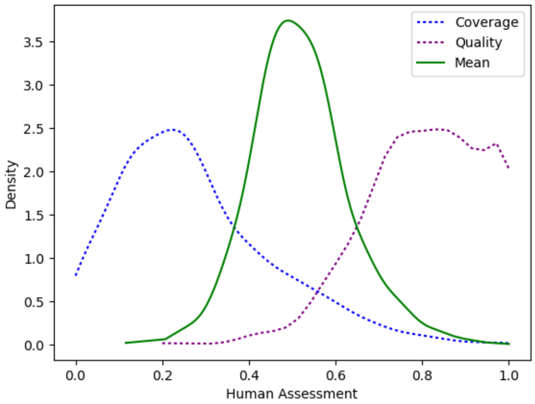

Figure 2: Mean linguistic quality (MLQ) and coverage scores (CS) distributions, with the average of the two distributions.

B. In Phase A the BIOASQ team released questions and the systems had to respond with relevant terminological and ontological concepts, relevant articles, snippets and RDF snippets. In Phase B, the BIOASQ team released questions and gold (correct) relevant concepts, articles, snippets and statements. The participating systems had to respond with ideal answers (i.e., paragraph sized summaries) for all questions, among other, irrelevant to this article types of responses. The answers are called ideal since they are what a human would expect from a biomedical expert. In this article, we focus on the evaluation of Phase B in Task B.

The dataset was created by the BIOASQ team of biomedical experts, and the gold relevant documents were identified using PUBMED. We used the dataset released for the fourth year of the challenge. It consists of 1625 questions and was split into a training, development and test set of 673, 279 and 673 instances respectively (after tokenization and stop word removal)(Tables 1 and 2). For each question in each set a number of system answers is available, and for each system answer its human assessment.

	# questions	max # systems / question	min # systems / question	mean # systems / question
Training set	673	26	2	11.6
Development set	279	20	4	10.1
Test set	257	17	3	10

Table 1: Overview of questions and systems per question for the BIOASQ datasets.

	# input vectors	mean question length	mean gold answer length	mean system answer length
Training set	7842	66.3	317.4	716.8
Development set	2825	64	334	735
Test set	2574	70.2	410.3	769

Table 2: Overview of questions and systems per question for the BIOASQ datasets.

Human assessment was provided by the BIOASQ team of experts, who inspected each system's ideal answer and were asked to evaluate it in terms of information recall (whether the answer contains all the necessary information), information precision (whether it contains no irrelevant information), information repetition (it does not repeat the same information) and readability (it is easily readable and fluent). The criteria were measured in a scale from 1 to 5, with 1 being the worst and 5 the best.

In contrast to the Graham paper, we try to predict the human-assigned score for each of the aforementioned measures by itself instead of combining them all to one.

4 Methods

We will now overview the proposed methods. We attempt to build deep MLP networks to predict the human assessment of the summaries. We employ different strategies when handling the inputs, which is discussed in the following sections, but the process of determining the architecture of the networks is common regardless of their input.

In the BIOASQ dataset, we attempted to infer the human-assigned scores to each of the Recall, Precision, Readability and Repetition measures, on summaries from the fourth year of the competition.

Having four measures to predict, we experimented with both single and multi-task learning . In the single task approach, we define a network which tries to predict a single measures score at a time (S-ROUGE and S-CENT, depending on the inputs). Thus we end up with four networks, one for each measure and each one is optimized to predict the corresponding measure.

In the multi-task approach, for each measure we develop a network which predicts the scores for all four measures at the same time. We chose network configurations dependant on a specific measure out of the four, again ending up with four different networks. We attempt to exploit shared information across all four measures so that the predictive capability for a single measure can be boosted (M-ROUGE and M-CENT, depending on the inputs).

To build the networks, we followed a common approach for both the single and multi-task learning strategies. The process consisted of defining different depth and widths for the architecture, overfitting them on training data and then regularizing the architecture which achieved the lowest Mean Squared Error (MSE) loss during overfitting. During regularization, we added dropout layers and randomly assigned their rate, as well as randomly assigning values for the L2 penalty on the weights. Furthermore, we employed early stopping during training while monitoring the loss on the validation data. We used the Adam optimizer with learning rate $lr = 0.001$((Kingma and Ba, 2014)).

In the case of DUC-2004 the target of the model is HAS, the combinatorial measure which takes into account the coverage expressed by the system summary, as well as the linguistic quality of the system summaries, as annotated by human assessors.

Since the target score in this dataset is a single score, we follow a single task learning approach. The methodology used to define this network is identical to the one described in the single task learning approach for BIOASQ, as are the overfitting and regularization of the networks.

4.1 Deep MLP over ROUGE variations (S-ROUGE & M-ROUGE)

One method we employed was inspired by previous analysis of current evaluation methodologies applied to summarization techniques. Graham et al. (2015) evaluated the performance of different variations of ROUGE metrics on the DUC-2004 dataset. The results suggested that certain variations were much superior to others, and the best performing were not always the ones most commonly used. In this thesis we attempt to continue this analysis by combining all the metric variations, taking advantage of the different insights each of them provides in summarization evaluation.

The ROUGE metrics we used include 8 choices of n-gram counting methods (ROUGE-1;2;3;4;S4;SU4;W;L), binary settings such as word stemming of summaries and inclusion or removal of stop words, as well as the use of recall, precision, or f-score to compute individual summary scores. In total we examined 96 variations of ROUGE.

In addition to the ROUGE variants, we also examine metrics based on similarity between the gold and system generated summaries. Each system summary is converted to a dense vector, representing the centroid of the word embeddings, which is also weighted by term frequency (TF) scores. We call this method CENT, and more specifically CENT-E when the Euclidean and CENT-C when the Cosine similarity are used to compute the similarity between the system summary and the gold (the representation of a gold is computed similarly to a system summary) (Brokos et al., 2016).

We also use 6 different variations of Word Mover's Disctance (WMD). WMD measures the total distance the word embeddings of two summaries have to travel to become identical (Kusner et al., 2015). In the same paper, they introduced a relaxed, faster version of WMD which measures the distance the word embeddings of only one summary have to travel. This led to two different Relaxed WMD (RWMD) variations according to whether the left (RWMD-L) or the right (RWMD-R) summary's distance from the other was measured, in terms of how many words of one are present in the other. They found that the maximum (RWMD-

MAX) of the two distances was the better relaxation. Finally, we differentiate between the use of Euclidean (RWMD-E) and Cosine (RWMD-C) distance as the metric of choice to use in RWMD.

The BLEU scores were computed using the JBLEU library.

In total, we calculated 105 metric variations for each system summary. A list of these metrics was used as input to each of the deep MLP architectures described above.

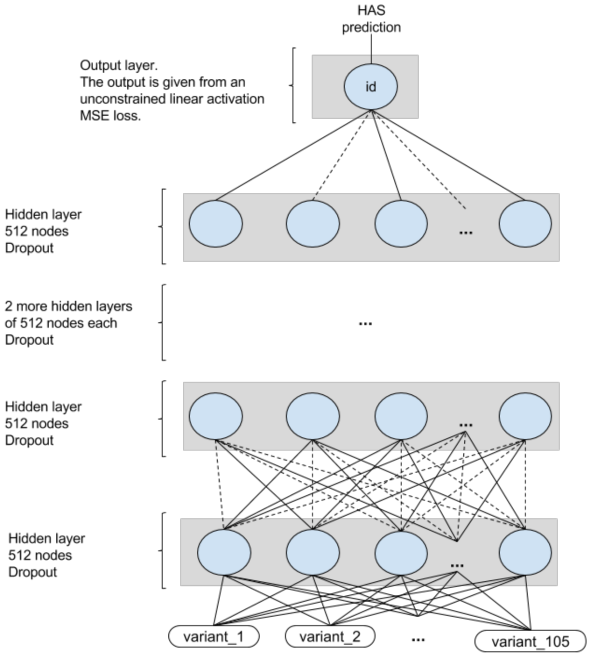

Figure 3: S-ROUGE network used on the DUC-2004 dataset. The network is 5 layers deep with each layer consisting of 512 nodes.

4.2 Deep MLP over difference of centroids (S-CENT & M-CENT)

ROUGE has a number of limitations which is the motivation for the development of the following two methods.

Those limitations have to do both with its implementation and the underlying assumptions. ROUGE has a large number of variants, of which only a small range is commonly used. We include 96 in this study, and depending on the task that number can grow. Furthermore, the n-gram coverage of the summaries, as used by classic ROUGE, favors lexical similarities between the system and human sentences, making it unsuitable for evaluating abstractive summaries, or summaries with a significant amount of paraphrasing. Finally, It is not suitable for evaluating the readability or repetition of the system generated summaries.

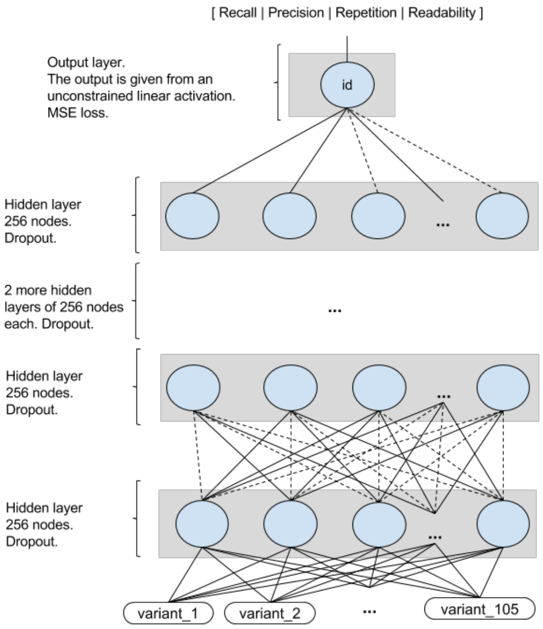

Figure 4: S-ROUGE network used on the BIOASQ dataset. The network is 5 layers deep with each layer consisting of 256 nodes.

In an attempt to combat those limitations we propose the use of the centroid of the summaries' word embeddings, instead of relying on the bag-of-words representation. While previous work has explored the use of dense representations of words or short sequences of words in ROUGE with good results (Ng and Abrecht, 2015), no one has made use of dense representations of whole summaries in the evaluation process to our knowledge.

This method has the advantage of bypassing the need to decide between a large number of ROUGE variants and the uncertainty this choice entails, while also overcoming any bias towards lexical similarity and instead focusing on semantic similarity.

Having a sentence t, the centroid representation of that sentence $\vec{t}$ is the sum of the embeddings of the tokens in the sentence, divided by the number of the tokens in it. In our experiments, we also take into account the IDF scores of the tokens:

$$\vec{t} = \frac{\sum_{j=1}^{|V|} \vec{w}_j \cdot TF(w_j, t) \cdot IDF(w_j)}{\sum_{j=1}^{|V|} TF(w_j, t) \cdot IDF(w_j)} \qquad (3)$$

where $|V|$ is the vocabulary size (approx. 1.7 million words, ignoring stop words), w_j is the j-th vocabulary word, $\vec{w}_j$ its embedding, $TF(w_j, t)$

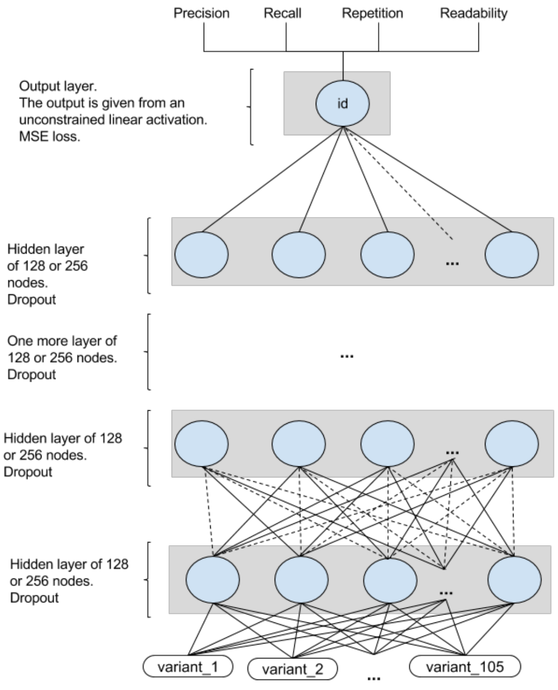
Precision
Recall
Repetition
Readability
Output layer.
The output is given from an unconstrained linear activation.
MSE loss.
id
Hidden layer of 128 or 256 nodes.
Dropout
...
One more layer of 128 or 256 nodes.
Dropout
...
Hidden layer of 128 or 256 nodes.
Dropout
...
Hidden layer of 128 or 256 nodes.
Dropout
...
variant_1
variant_2
...
variant_105

Figure 5: M-ROUGE network used on the BIOASQ dataset. The exact network architecture depends on the metric, since we create one network for each metric. It can be up to 4 layers deep with each layer consisting of 128 or 256 nodes.

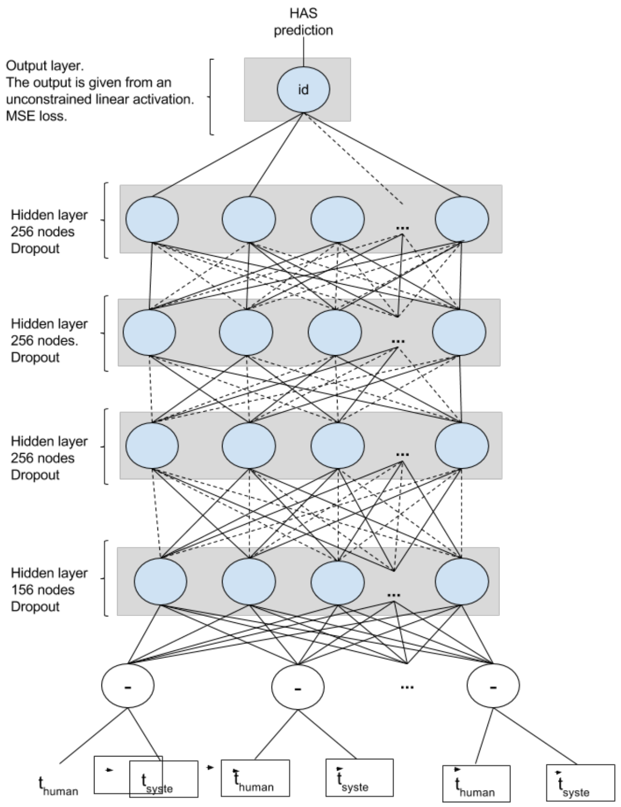

Figure 6: S-CENT network used on the DUC-2004 dataset, with the difference of the centroids of the summaries as input. The network consists of 4 layers of 256 nodes each.

the term frequency of w_j in t, and $IDF(w_j)$ the inverse document frequency of w_j (Manning et al., 2008).

We use 200-dimensional embeddings and precompute the centroid representation of the sentences before using them in our models, resulting in 200-dimensional vectors.

In the case of BIOASQ, the word embeddings used come from applying WORD2VEC to approximately 11 million abstracts obtained from PubMed. The IDF scores are calculated on the vocabulary of those abstracts. In the case of DUC-2004, the word embeddings were obtained from the Wikipedia 2015 + Gigaword 5 GloVe repository and the IDF scores are computed on the vocabulary of the dataset.

Having computed the centroids of each human and system summary, we subtract the system summary centroid representation from the human one and we use the resulting 200-dimensional vector as input to each of the deep MLP architectures described above.

5 Results

5.1 Metric Evaluation by Correlation

We present correlation results between the human evaluation values and the metric values, for both the datasets, for every metric mentioned in this paper. These metrics include the 96 ROUGE variants, the 2 centroid distance based metrics, the 6 variants of WMD, BLEU and the 4 MLP metrics we developed.

The performance evaluation for all the metrics was calculated by employing three different correlation measures, Pearson's r, Spearman's ρ and Kendall's τ. All three metrics have been used to explore the correlation between summaries in the field of NLP, but there are differences between them.

Pearson correlation estimates linear correlations under the assumption of normally distributed data. Spearman correlation estimates the monotonic relationship between two datasets, but it does not assume that they are normally distributed. On the other hand, Kendall's τ is a probabilistic function,

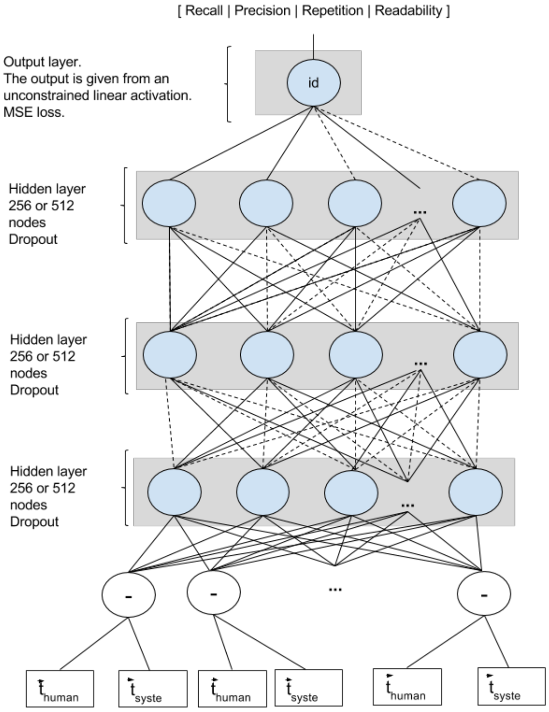

Figure 7: S-CENT network used on the BIOASQ dataset, with the difference of the centroids of the summaries as input. The exact network architecture depends on the metric, since we create one network for each metric. Each network is 3 layers deep, and each layer consists of 256 or 512 nodes.

measuring the discrepancy between the number of concordant and discordant pairs. We decided to present the results sorted by Spearman's ρ, with the two other correlations presented for a complete overview.

5.2 Metric Significance Testing

The superiority of one metric over another should be examined through significance testing. Graham et al, 2015, employed Williams tests to determine the significance of a metric when compared with all the other metrics. We are using the same methodology of applying Williams tests to the difference between the correlations of measures, instead of exploring the significance of each metric on its own. The test can determine whether the metric under examination is statistically more significant than any other, worse performing metric. In the results tables, a • next to the Spearman's ρ value indicates that the particular metric is not significantly outperformed by any other metric that comes after it.

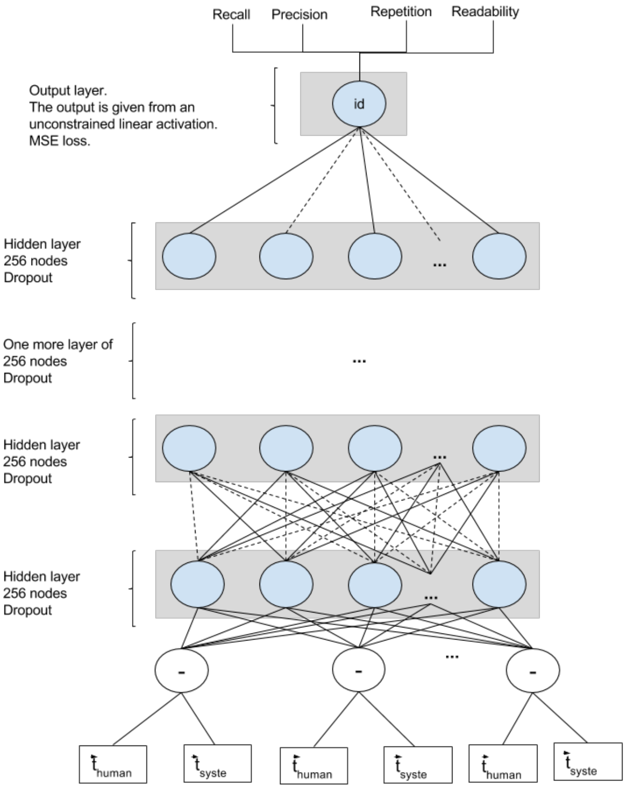

Figure 8: M-CENT network used on the BIOASQ dataset, with the difference of the centroids of the summaries as input. The exact network architecture depends on the metric, since we create one network for each metric. Each network is up to 5 layers deep, and each layer consists of 256 nodes.

5.3 Summarization Metrics Evaluation

Table 3 shows the correlations of each metric with human assessment on the DUC-2004 datasets. The S-ROUGE metric was defined in section 4.1 as the single task learning approach to a deep MLP architecture over an input of all the ROUGE and other variants. This metric achieves the strongest Spearman correlation value, ρ = 0.624, with the human assessment data. The metric also achieves the strongest Pearson's r and Kendall's τ, but it does not achieve statistical significance. The next best metric is ROUGE-2 recall with the words stemmed and stopwords removed, and it is statistically significant in this case. This ROUGE-1 metric is not among the best metrics as examined by Graham et al. (2015), but ROUGE-1 is recommended in Owczarzak et al. (2012).

Table 4 contains the correlations between each metric and the human assigned repetition score in the BIOASQ dataset. We see that the S-ROUGE

Metrics	Stem.	RWS	P/R/F	r	ρ	τ	Metrics	Stem.	RWS	P/R/F	r	ρ	τ
S-ROUGE	-	-	-	0.651	0.624	0.456	R-L	N	N	P	0.587	0.479	0.332
R-1	Y	Y	R	0.647	0.602 ●	0.433	R-L	Y	Y	R	0.589	0.474	0.332
R-1	Y	Y	F1	0.628	0.571	0.409	R-2	N	N	R	0.591	0.474	0.325
R-W	Y	Y	P	0.616	0.552	0.389	R-3	Y	Y	R	0.592	0.473	0.337
R-3	Y	N	R	0.605	0.551	0.395	R-4	N	N	R	0.532	0.467	0.350
R-3	Y	N	F1	0.606	0.549	0.392	R-4	N	N	F1	0.532	0.465	0.348
R-S4	Y	Y	R	0.608	0.548	0.392	R-W	N	N	F1	0.576	0.465	0.321
R-W	Y	N	P	0.605	0.547	0.386	R-L	Y	N	R	0.589	0.464	0.320
R-3	Y	N	P	0.607	0.546	0.389	R-4	N	N	P	0.532	0.463	0.345
R-S4	Y	Y	F1	0.609	0.545	0.388	R-4	N	Y	P	0.449	0.463	0.358
R-2	Y	N	F1	0.628	0.543	0.386	R-4	N	Y	F1	0.448	0.462	0.358
R-L	Y	Y	P	0.616	0.541	0.380	R-4	N	Y	R	0.446	0.462	0.359
R-2	Y	Y	R	0.625	0.539	0.385	RWMD-R-C	-	-	-	0.441	0.455	0.314
R-2	Y	N	P	0.626	0.537	0.380	R-4	Y	Y	P	0.576	0.452	0.338
R-W	Y	N	F1	0.597	0.536	0.378	R-4	Y	Y	F1	0.577	0.451	0.339
R-2	Y	Y	F1	0.624	0.535	0.380	R-4	Y	Y	R	0.577	0.444	0.334
R-1	Y	N	F1	0.622	0.534	0.378	R-3	N	Y	F1	0.496	0.443	0.335
R-S4	Y	Y	P	0.607	0.533	0.378	R-1	N	Y	F1	0.523	0.442	0.306
R-SU4	Y	Y	R	0.606	0.533	0.379	R-3	N	Y	P	0.495	0.442	0.333
R-2	Y	N	R	0.626	0.533	0.378	R-L	N	N	F1	0.576	0.442	0.302
R-SU4	Y	Y	F1	0.606	0.532	0.379	R-2	N	Y	F1	0.540	0.440	0.315
RWMD-L-E	-	-	-	0.544	0.529	0.372	R-3	N	Y	R	0.495	0.437	0.330
R-W	Y	Y	F1	0.606	0.527	0.371	R-S4	N	N	P	0.574	0.436	0.300
R-L	Y	N	P	0.614	0.527	0.372	R-2	N	Y	P	0.525	0.436	0.310
RWMD-L-C	-	-	-	0.532	0.526	0.366	R-SU4	N	N	P	0.574	0.435	0.299
R-S4	Y	N	P	0.612	0.525	0.369	R-SU4	N	Y	P	0.535	0.433	0.300
RWMD-M-C	-	-	-	0.525	0.523	0.365	R-S4	N	Y	P	0.535	0.433	0.300
R-SU4	Y	Y	P	0.605	0.523	0.370	R-S4	N	Y	F1	0.537	0.432	0.299
R-SU4	Y	N	P	0.610	0.520	0.366	R-SU4	N	Y	F1	0.538	0.431	0.299
RWMD-M-E	-	-	-	0.541	0.520	0.362	R-2	N	Y	R	0.543	0.431	0.309
R-S4	Y	N	F1	0.610	0.519	0.366	RWMD-R-E	-	-	-	0.431	0.426	0.294
R-SU4	Y	N	F1	0.609	0.515	0.362	R-SU4	N	Y	R	0.537	0.425	0.295
R-L	Y	Y	F1	0.604	0.515	0.361	R-S4	N	Y	R	0.537	0.425	0.296
R-1	Y	N	P	0.597	0.515	0.364	BLEU	-	-	-	0.417	0.424	0.295
R-1	Y	Y	P	0.576	0.514	0.364	R-W	N	N	R	0.563	0.422	0.287
R-1	Y	N	R	0.605	0.513	0.362	R-S4	N	N	F1	0.573	0.421	0.289
R-2	Y	Y	P	0.614	0.512	0.361	R-SU4	N	N	F1	0.573	0.421	0.289
R-3	N	N	R	0.564	0.508	0.369	R-1	N	N	P	0.522	0.416	0.288
R-3	N	N	P	0.563	0.508	0.368	R-1	N	N	F1	0.528	0.410	0.282
R-3	N	N	F1	0.564	0.508	0.368	R-W	N	Y	P	0.542	0.410	0.278
R-S4	Y	N	R	0.608	0.504	0.356	R-S4	N	N	R	0.569	0.407	0.278
R-L	Y	N	F1	0.603	0.503	0.352	R-SU4	N	N	R	0.569	0.407	0.278
R-SU4	Y	N	R	0.606	0.499	0.352	CENT-C	-	-	-	0.348	0.404	0.276
R-4	Y	N	R	0.588	0.498	0.365	R-L	N	N	R	0.562	0.402	0.273
R-W	Y	N	R	0.590	0.498	0.346	R-L	N	Y	P	0.544	0.401	0.273
R-W	N	N	P	0.584	0.496	0.343	R-1	N	Y	P	0.491	0.400	0.273
R-4	Y	N	F1	0.588	0.496	0.362	R-1	N	N	R	0.486	0.388	0.265
R-4	Y	N	P	0.588	0.495	0.360	R-W	N	Y	F1	0.537	0.384	0.261
R-2	N	N	P	0.587	0.491	0.338	R-L	N	Y	F1	0.538	0.378	0.257
R-1	N	Y	R	0.547	0.487	0.340	R-W	N	Y	R	0.529	0.363	0.245
R-3	Y	Y	P	0.593	0.486	0.344	R-L	N	Y	R	0.531	0.360	0.243
R-2	N	N	F1	0.592	0.486	0.334	CENT-E	-	-	-	0.323	0.347	0.236
R-3	Y	Y	F1	0.593	0.484	0.344	S-CENT	-	-	-	0.293	0.282	0.187
R-W	Y	Y	R	0.594	0.484	0.337							

Table 3: Pearson (r), Spearman (ρ) and Kendall (τ) correlations of BLEU, 96 variants of ROUGE (R-*), centroid distance based metrics (CENT-C, CENT-E), WMD based metrics and methods described in this paper(S-ROUGE, S-CENT) with human assessment in DUC-2004. The ROUGE variants are presented with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F). The results are sorted in descending order by ρ. Correlations marked with ● signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant according to pairwise Williams significance tests.

metric once again achieves the strongest Spearman correlation, $\rho = 0.55$, with a significant margin from the second best metric, which also holds for the other two correlations. The metric is not significantly outperformed by any other metric we examined. The second best correlation is ROUGE-W precision with stemming and stopwords removed, also not among the recommended metrics in the papers of Graham (2015) or Owczarzak (2012). The ROUGE variants most sensitive to repetition appear to be those using precision and F1 to calculate the individual summary scores, and based on the length of common subsequences found in the summaries, which intuitively is expected, especially since the variants that are N-gram based do the worst.

Table 5 shows the correlations between each metric and the human assigned recall score in BIOASQ. The metric that achieves the strongest Spearman correlation with $\rho = 0.529$ is M-ROUGE which was defined in section 4.1 as well. In this case, it refers to the multitask learning approach to the MLP architecture with the same input as described above. The second best approach with $\rho = 0.525$ is S-ROUGE. The metrics retain their relative position when the other two correlation measures are considered. The absolute difference between the ρ values is very small though, and of the two, S-ROUGE is the one not significanlty outperformed by any other metric in the list. In terms of ROUGE variants, the ROUGE-W recall with stopword removal comes third in its correlation strength, with significantly lower correlation value. Worth noting is that the top ROUGE variants in this case are all recall based, and based on the longest common subsequence of words present in the summaries, whether weighted (ROUGE-W) or not (ROUGE-L).

Table 6 shows the correlations between each metric and the human assigned readability score in BIOASQ. Once again, S-ROUGE achieves the strongest correlation, with $\rho = 0.54$. The metric is also not significantly outperformed by any other metric, and the ρ values of the two other correlation measures retain the relative ordering. While the Spearman's ρ for M-ROUGE is not comparable to S-ROUGE, its Pearson's r would put it much nearer the top of the list, but Kendall's τ would maintain its position. The second best metric is ROUGE-W precision, with the stopwords removed, but the difference in correlation values is quite pronounced. As with repetition, the ROUGE variants most sensitive to the readability of sentences are either precision or F1 based, and take into account the length of common subsequences of words.

Table 7 contains the correlations between each metric and the human assigned precision score in BIOASQ. S-ROUGE achieves the strongest correlation, $\rho = 0.463$, and is not significantly outperformed by any other metric. The values of the two other correlation measures would retain its position, while the Pearson's r value for S-CENT and M-CENT would place them in the second and third place respectively. The ROUGE variants that achieve the strongest correlations are precision based, and the second best metric overall is ROUGE-W precision with stemming and stopword removal, with $\rho = 0.425$. It should be expected that the precision based ROUGE variants would do better in a precision evaluation task as was with recall evaluation in Table 5.

Table 8 shows the correlations between each metric and the mean of all four human assigned scores in BIOASQ. In this case, the metric with the strongest correlation is a ROUGE variant, ROUGE-W F1 with stemming and stopword removal, achieving $\rho = 0.519$. Both of the metrics we developed as part of this study, MEAN-ROUGE and MEAN-CENT, fail to achieve a strong Spearman correlation. However, the Pearson correlation of both is very strong, and would place MEAN-CENT at the top of the list, with MEAN-ROUGE third. But even so, their relative placement when considering the Kendall correlation would not change. It is worth noting that while ROUGE-W F1 with stemming and stopword removal comes first when considering the correlation values, its performance when tested against the other metrics that present weaker correlations is not significantly better, as opposed to MEAN-ROUGE and the metrics against which it does better.

6 Discussion

The results of the metrics evaluation certainly raise some questions. It is evident that the method we describe in this study, S-ROUGE, is able to achieve the strongest correlations with human assessment, for both the datasets and for the most tasks. The difference between correlation values for S-ROUGE and the next best metric is quite significant, both in absolute terms and statistically, with the exception of DUC-2004.

Metrics	Stem.	RWS	P/R/F	r	ρ	τ	Metrics	Stem.	RWS	P/R/F	r	ρ	τ
S-ROUGE	-	-	-	0.494	0.550 ●	0.432	R-S4	N	N	F1	0.144	0.111	0.085
R-W	Y	Y	P	0.347	0.439 ●	0.342	R-1	N	N	F1	0.204	0.105	0.080
R-L	Y	Y	P	0.373	0.434 ●	0.339	R-2	N	N	F1	0.131	0.097	0.075
R-W	Y	N	P	0.324	0.425	0.332	R-2	N	Y	F1	0.127	0.083	0.064
R-W	N	Y	P	0.335	0.421	0.329	R-4	Y	N	P	0.166	0.079	0.064
R-L	Y	N	P	0.354	0.420	0.329	R-3	Y	N	F1	0.131	0.076	0.060
R-L	N	Y	P	0.358	0.416	0.325	R-4	Y	Y	P	0.167	0.076	0.064
R-W	N	N	P	0.316	0.414	0.323	R-3	Y	Y	F1	0.132	0.075	0.060
R-L	N	N	P	0.344	0.410 ●	0.320	CENT-E	-	-	-	0.283	0.074	0.055
R-W	Y	Y	F1	0.299	0.347 ●	0.269	M-ROUGE	-	-	-	0.271	0.070	0.052
R-L	Y	Y	F1	0.321	0.337	0.261	R-3	N	N	P	0.157	0.068	0.054
R-W	Y	N	F1	0.281	0.337	0.261	R-4	Y	Y	F1	0.120	0.051	0.043
R-W	N	Y	F1	0.286	0.329	0.255	R-4	Y	N	F1	0.122	0.049	0.040
R-L	Y	N	F1	0.307	0.329	0.255	R-4	N	Y	P	0.137	0.049	0.041
R-W	N	N	F1	0.273	0.325	0.252	R-4	N	N	P	0.141	0.046	0.037
R-L	N	N	F1	0.298	0.319 ●	0.247	R-3	N	Y	P	0.151	0.044	0.036
R-L	N	Y	F1	0.305	0.319	0.247	R-4	N	Y	F1	0.094	0.035	0.029
RWMD-R-C	-	-	-	0.374	0.277	0.213	R-3	N	N	F1	0.109	0.033	0.026
R-1	Y	Y	P	0.319	0.272 ●	0.209	R-4	N	N	F1	0.099	0.025	0.020
RWMD-R-E	-	-	-	0.338	0.258	0.199	R-3	N	Y	F1	0.104	0.021 ●	0.017
R-1	Y	N	P	0.308	0.250	0.193	R-4	N	Y	R	-0.017	-0.007	-0.006
R-SU4	Y	Y	P	0.233	0.237	0.182	R-4	Y	Y	R	-0.006	-0.014 ●	-0.012
R-1	N	Y	P	0.279	0.236	0.182	R-4	Y	N	R	-0.008	-0.029	-0.024
S-CENT	-	-	-	0.43	0.233	0.179	R-4	N	N	R	-0.024	-0.033	-0.028
R-SU4	N	Y	P	0.224	0.222	0.171	R-3	Y	Y	R	-0.022	-0.034	-0.028
R-1	N	N	P	0.280	0.221	0.170	R-SU4	N	Y	R	-0.011	-0.036	-0.030
BLEU	-	-	-	0.191	0.219	0.168	R-3	Y	N	R	-0.017	-0.038	-0.031
RWMD-M-C	-	-	-	0.341	0.219	0.169	R-SU4	Y	Y	R	-0.012	-0.042	-0.035
R-S4	Y	Y	P	0.226	0.216 ●	0.167	R-2	Y	Y	R	-0.027	-0.049	-0.039
R-S4	N	Y	P	0.218	0.205	0.158	R-3	N	Y	R	-0.035	-0.051	-0.042
R-SU4	Y	N	P	0.211	0.202	0.155	R-2	N	Y	R	-0.04	-0.051	-0.041
R-S4	Y	N	P	0.208	0.194	0.149	R-S4	N	Y	R	-0.025	-0.051	-0.041
R-2	Y	N	P	0.212	0.193	0.149	R-3	N	N	R	-0.032	-0.052	-0.043
RWMD-M-E	-	-	-	0.300	0.190	0.146	R-2	Y	N	R	-0.027	-0.053	-0.043
R-SU4	N	N	P	0.205	0.190	0.146	R-2	N	N	R	-0.038	-0.055	-0.044
R-2	Y	Y	P	0.215	0.185	0.144	R-S4	Y	Y	R	-0.028	-0.058	-0.048
R-S4	N	N	P	0.202	0.183 ●	0.140	R-SU4	N	N	R	-0.025	-0.066	-0.054
R-2	N	N	P	0.187	0.157	0.122	R-SU4	Y	N	R	-0.028	-0.070	-0.057
R-SU4	Y	Y	F1	0.168	0.155	0.118	R-S4	N	N	R	-0.032	-0.071	-0.058
R-1	Y	Y	F1	0.241	0.154	0.117	R-S4	Y	N	R	-0.036	-0.077	-0.063
R-SU4	N	Y	F1	0.162	0.147	0.112	RWMD-L-E	-	-	-	0.033	-0.078	-0.063
R-S4	Y	Y	F1	0.160	0.139	0.106	R-W	N	Y	R	-0.014	-0.078	-0.062
M-CENT	-	-	-	0.428	0.138	0.105	RWMD-L-C	-	-	-	0.140	-0.078	-0.063
R-1	N	Y	F1	0.207	0.134	0.102	R-W	Y	Y	R	-0.012	-0.078	-0.062
R-S4	N	Y	F1	0.155	0.133	0.101	R-W	N	N	R	-0.017	-0.084	-0.067
CENT-C	-	-	-	0.322	0.132	0.101	R-W	Y	N	R	-0.016	-0.085	-0.069
R-2	N	Y	P	0.182	0.131	0.103	R-1	Y	Y	R	0.038	-0.092	-0.074
R-1	Y	N	F1	0.232	0.127	0.096	R-1	N	Y	R	0.005	-0.093	-0.075
R-SU4	Y	N	F1	0.151	0.122	0.093	R-L	N	Y	R	-0.031	-0.106 ●	-0.084
R-3	Y	N	P	0.182	0.122	0.097	R-L	Y	Y	R	-0.030	-0.110	-0.088
R-2	Y	N	F1	0.149	0.120	0.092	R-L	N	N	R	-0.041	-0.121	-0.097
R-2	Y	Y	F1	0.148	0.116	0.089	R-1	Y	N	R	0.027	-0.121	-0.098
R-S4	Y	N	F1	0.147	0.116	0.088	R-L	Y	N	R	-0.042	-0.126	-0.101
R-SU4	N	N	F1	0.147	0.116	0.088	R-1	N	N	R	-0.002	-0.129	-0.105
R-3	Y	Y	P	0.184	0.113	0.091							

Table 4: Pearson (r), Spearman (ρ) and Kendall (τ) correlations of BLEU, 96 variants of ROUGE (R-*), centroid distance based metrics (CENT-C, CENT-E), WMD based metrics and methods described in this paper(S-ROUGE, M-ROUGE, S-CENT, M-CENT) with human assessment of the **Repetition** metric in BIOASQ. The ROUGE variants are presented with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F). The results are sorted in descending order by ρ. Correlations marked with ● signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant according to pairwise Williams significance tests.

Metrics	Stem.	RWS	P/R/F	r	ρ	τ	Metrics	Stem.	RWS	P/R/F	r	ρ	τ
M-ROUGE	-	-	-	0.584	0.529	0.417	R-2	N	N	F1	0.266	0.321	0.252
S-ROUGE	-	-	-	0.531	0.525 ●	0.413	R-4	Y	Y	R	0.287	0.320	0.267
R-W	N	Y	R	0.452	0.502	0.396	R-2	N	Y	F1	0.256	0.313	0.248
R-W	Y	Y	R	0.458	0.499 ●	0.394	R-3	Y	N	P	0.199	0.313	0.249
R-W	N	N	R	0.434	0.487	0.383	R-1	Y	N	F1	0.368	0.312	0.242
R-L	N	Y	R	0.466	0.486	0.384	R-1	N	Y	F1	0.345	0.310	0.241
R-W	Y	N	R	0.437	0.484	0.381	R-2	Y	Y	P	0.233	0.309	0.242
R-L	Y	Y	R	0.471	0.483 ●	0.382	R-SU4	N	Y	P	0.22	0.307	0.238
R-L	N	N	R	0.449	0.467	0.368	R-4	Y	N	F1	0.227	0.302	0.246
RWMD-L-E	-	-	-	0.471	0.465	0.369	R-S4	N	Y	P	0.215	0.301	0.234
R-L	Y	N	R	0.451	0.464	0.366	R-4	N	N	R	0.266	0.297	0.250
RWMD-L-C	-	-	-	0.541	0.454	0.360	RWMD-M-E	-	-	-	0.342	0.296	0.230
R-SU4	Y	Y	R	0.394	0.448	0.352	R-3	Y	Y	P	0.199	0.296	0.238
R-SU4	N	Y	R	0.382	0.444	0.348	RWMD-M-C	-	-	-	0.396	0.295	0.228
R-S4	Y	Y	R	0.391	0.439	0.345	R-2	Y	N	P	0.223	0.293	0.228
R-2	Y	Y	R	0.395	0.438	0.347	R-SU4	Y	Y	P	0.220	0.292	0.226
R-2	Y	N	R	0.395	0.434	0.342	R-3	N	N	F1	0.226	0.292	0.237
R-S4	N	Y	R	0.378	0.433	0.340	R-4	Y	Y	F1	0.220	0.289	0.239
R-1	Y	Y	R	0.481	0.431	0.340	R-S4	Y	Y	P	0.216	0.286	0.222
R-SU4	Y	N	R	0.379	0.420	0.329	R-4	Y	N	P	0.176	0.281	0.229
R-S4	Y	N	R	0.379	0.418	0.328	R-3	N	Y	R	0.270	0.281	0.235
R-SU4	N	N	R	0.371	0.417	0.326	R-4	N	Y	R	0.230	0.277	0.237
R-1	Y	N	R	0.470	0.415	0.327	R-1	N	N	F1	0.325	0.273	0.212
R-S4	N	N	R	0.370	0.414 ●	0.324	CENT-E	-	-	-	0.398	0.271	0.210
R-3	Y	N	R	0.343	0.394	0.315	R-4	N	N	F1	0.199	0.271	0.227
R-1	N	Y	R	0.423	0.383	0.300	R-SU4	N	N	P	0.207	0.270	0.209
R-2	N	N	R	0.352	0.383	0.303	R-2	N	Y	P	0.196	0.268	0.212
R-W	N	Y	F1	0.299	0.379	0.295	R-4	Y	Y	P	0.174	0.267	0.221
R-2	N	Y	R	0.337	0.374	0.298	R-3	N	N	P	0.173	0.267	0.217
R-1	N	N	R	0.425	0.373	0.293	R-S4	N	N	P	0.203	0.266	0.206
R-3	Y	Y	R	0.333	0.372	0.301	R-2	N	N	P	0.195	0.260	0.204
CENT-C	-	-	-	0.487	0.371	0.290	R-SU4	Y	N	P	0.206	0.260	0.201
R-W	Y	Y	F1	0.300	0.370	0.288	R-4	N	Y	F1	0.182	0.257	0.219
R-W	N	N	F1	0.285	0.370	0.288	R-S4	Y	N	P	0.202	0.256	0.198
R-2	Y	Y	F1	0.305	0.369	0.288	R-4	N	N	P	0.153	0.256	0.214
R-SU4	N	Y	F1	0.284	0.369	0.287	R-1	Y	Y	P	0.284	0.252	0.194
BLEU	-	-	-	0.229	0.368	0.287	R-3	N	Y	F1	0.208	0.251	0.208
R-2	Y	N	F1	0.302	0.365	0.285	R-4	N	Y	P	0.148	0.245	0.208
R-W	Y	N	F1	0.285	0.363	0.283	RWMD-R-E	-	-	-	0.287	0.244	0.188
R-SU4	Y	Y	F1	0.289	0.363	0.283	R-1	N	Y	P	0.254	0.231	0.179
R-S4	N	Y	F1	0.279	0.359	0.279	RWMD-R-C	-	-	-	0.337	0.231	0.178
R-1	Y	Y	F1	0.398	0.354	0.275	R-3	N	Y	P	0.165	0.230	0.191
R-S4	Y	Y	F1	0.285	0.354	0.276	R-W	N	Y	P	0.190	0.227 ●	0.176
R-L	N	Y	F1	0.311	0.352	0.274	R-W	N	N	P	0.173	0.210	0.163
R-3	Y	N	F1	0.259	0.347	0.275	R-W	Y	Y	P	0.184	0.208	0.161
R-SU4	N	N	F1	0.277	0.342	0.266	R-L	N	Y	P	0.188	0.199	0.154
R-L	Y	Y	F1	0.311	0.341	0.266	R-W	Y	N	P	0.171	0.199	0.154
R-SU4	Y	N	F1	0.280	0.340	0.265	S-CENT	-	-	-	0.418	0.188	0.145
R-4	Y	N	R	0.305	0.337	0.277	R-1	Y	N	P	0.240	0.183	0.141
R-S4	N	N	F1	0.275	0.337	0.262	R-L	Y	Y	P	0.181	0.177	0.136
R-S4	Y	N	F1	0.278	0.336	0.261	R-L	N	N	P	0.165	0.164	0.127
R-L	N	N	F1	0.293	0.330	0.257	R-1	N	N	P	0.214	0.160	0.123
R-3	N	N	R	0.301	0.327	0.267	R-L	Y	N	P	0.161	0.152	0.117
R-3	Y	Y	F1	0.254	0.326	0.262	M-CENT	-	-	-	0.433	0.117	0.091
R-L	Y	N	F1	0.294	0.323	0.253							

Table 5: Pearson (r), Spearman (ρ) and Kendall (τ) correlations of BLEU, 96 variants of ROUGE (R-*), centroid distance based metrics (CENT-C, CENT-E), WMD based metrics and methods described in this paper(S-ROUGE, M-ROUGE, S-CENT, M-CENT) with human assessment of the **Recall** metric in BIOASQ. The ROUGE variants are presented with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F). The results are sorted in descending order by ρ. Correlations marked with ● signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant according to pairwise Williams significance tests.

Metrics	Stem.	RWS	P/R/F	r	ρ	τ	Metrics	Stem.	RWS	P/R/F	r	ρ	τ
S-ROUGE	-	-	-	0.520	0.540 ●	0.427	R-2	N	Y	F1	0.295	0.295	0.234
R-W	N	Y	P	0.420	0.516	0.409	R-3	Y	N	F1	0.290	0.291	0.232
R-W	Y	Y	P	0.421	0.514	0.407	R-4	Y	N	P	0.292	0.29	0.238
R-L	N	Y	P	0.442	0.509	0.403	CENT-C	-	-	-	0.396	0.289	0.223
R-L	Y	Y	P	0.444	0.506	0.401	M-ROUGE	-	-	-	0.417	0.284	0.218
R-W	N	N	P	0.399	0.505	0.401	R-3	Y	Y	F1	0.288	0.281	0.227
R-W	Y	N	P	0.400	0.504	0.400	R-4	Y	Y	P	0.287	0.281	0.234
R-W	Y	Y	F1	0.419	0.503	0.396	R-3	N	N	P	0.290	0.272	0.222
R-W	N	Y	F1	0.414	0.499	0.393	R-4	Y	N	F1	0.271	0.271	0.220
R-L	Y	Y	F1	0.444	0.495	0.390	R-4	Y	Y	F1	0.266	0.265	0.220
R-W	Y	N	F1	0.402	0.495	0.389	R-4	N	N	P	0.268	0.259	0.217
R-W	N	N	F1	0.400	0.494	0.388	R-4	N	Y	P	0.252	0.256	0.219
R-L	N	N	P	0.421	0.492	0.390	R-3	N	N	F1	0.267	0.253	0.205
R-L	N	Y	F1	0.437	0.490	0.386	CENT-E	-	-	-	0.398	0.251	0.194
R-L	Y	N	P	0.423	0.490	0.389	R-4	N	Y	F1	0.231	0.246	0.209
R-L	N	N	F1	0.426	0.485	0.381	R-4	N	N	F1	0.248	0.244	0.202
R-L	Y	N	F1	0.429	0.484	0.381	R-3	N	Y	P	0.277	0.238	0.197
RWMD-R-C	-	-	-	0.484	0.428	0.336	R-3	N	Y	F1	0.254	0.224	0.185
R-1	Y	Y	P	0.446	0.428	0.336	R-4	Y	Y	R	0.193	0.217	0.177
R-SU4	Y	Y	P	0.364	0.427 ●	0.334	R-4	N	Y	R	0.165	0.216	0.181
R-SU4	N	Y	P	0.359	0.422	0.330	S-CENT	-	-	-	0.427	0.211	0.161
R-S4	Y	Y	P	0.357	0.411	0.322	R-SU4	N	Y	R	0.210	0.209	0.159
BLEU	-	-	-	0.312	0.408	0.319	R-4	Y	N	R	0.195	0.207	0.166
RWMD-M-C	-	-	-	0.477	0.407	0.318	R-SU4	Y	Y	R	0.206	0.197	0.150
R-S4	N	Y	P	0.352	0.406	0.320	R-4	N	N	R	0.170	0.195	0.159
R-1	N	Y	P	0.422	0.405	0.317	R-3	Y	N	R	0.192	0.193	0.152
WMD-R-E	-	-	-	0.424	0.401	0.314	R-3	Y	Y	R	0.187	0.190	0.151
R-1	Y	N	P	0.421	0.393	0.307	R-S4	N	Y	R	0.193	0.187	0.143
R-SU4	Y	N	P	0.348	0.393	0.307	R-2	Y	N	R	0.189	0.179	0.137
R-SU4	N	N	P	0.346	0.393	0.306	R-SU4	N	N	R	0.196	0.179	0.137
R-S4	N	N	P	0.342	0.384	0.300	R-2	Y	Y	R	0.186	0.178	0.137
R-S4	Y	N	P	0.343	0.384	0.300	R-3	N	N	R	0.171	0.178	0.142
R-SU4	Y	Y	F1	0.339	0.381	0.297	R-2	N	N	R	0.176	0.175	0.134
R-SU4	N	Y	F1	0.333	0.378	0.295	R-S4	Y	Y	R	0.188	0.175	0.134
R-2	Y	N	P	0.347	0.376	0.295	M-CENT	-	-	-	0.413	0.172	0.130
R-1	N	N	P	0.404	0.374	0.291	R-S4	N	N	R	0.189	0.172	0.131
R-2	Y	Y	P	0.353	0.372	0.295	R-SU4	Y	N	R	0.192	0.171	0.130
RWMD-M-E	-	-	-	0.410	0.371	0.290	R-2	N	Y	R	0.168	0.167	0.129
R-S4	Y	Y	F1	0.331	0.367	0.286	R-S4	Y	N	R	0.184	0.163	0.124
R-1	Y	Y	F1	0.416	0.364	0.283	R-3	N	Y	R	0.163	0.162	0.132
R-S4	N	Y	F1	0.326	0.363 ●	0.284	R-W	N	Y	R	0.194	0.148	0.112
R-SU4	Y	N	F1	0.326	0.352	0.274	R-W	N	N	R	0.188	0.139	0.105
R-SU4	N	N	F1	0.324	0.351	0.274	R-W	Y	Y	R	0.187	0.135	0.102
R-2	N	N	P	0.325	0.344	0.270	R-W	Y	N	R	0.183	0.129	0.098
R-S4	Y	N	F1	0.322	0.344	0.268	RWMD-L-E	-	-	-	0.215	0.127	0.097
R-S4	N	N	F1	0.320	0.344	0.268	R-L	N	Y	R	0.180	0.123	0.093
R-1	N	Y	F1	0.389	0.341	0.266	RWMD-L-C	-	-	-	0.282	0.116	0.089
R-2	Y	N	F1	0.324	0.340	0.266	R-L	Y	Y	R	0.170	0.107	0.081
R-1	Y	N	F1	0.396	0.339	0.263	R-1	N	Y	R	0.189	0.104	0.079
R-2	Y	Y	F1	0.322	0.332	0.261	R-L	N	N	R	0.165	0.103	0.077
R-2	N	Y	P	0.321	0.327	0.261	R-1	Y	Y	R	0.204	0.098	0.074
R-2	N	N	F1	0.304	0.316	0.247	R-L	Y	N	R	0.157	0.092	0.069
R-3	Y	N	P	0.312	0.315	0.254	R-1	N	N	R	0.189	0.089	0.067
R-1	N	N	F1	0.375	0.314	0.244	R-1	Y	N	R	0.199	0.088	0.066
R-3	Y	Y	P	0.313	0.305	0.248							

Table 6: Pearson (r), Spearman (ρ) and Kendall (τ) correlations of BLEU, 96 variants of ROUGE (R-*), centroid distance based metrics (CENT-C, CENT-E), WMD based metrics and methods described in this paper(S-ROUGE, M-ROUGE, S-CENT, M-CENT) with human assessment of the **Readability** metric in BIOASQ. The ROUGE variants are presented with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F). The results are sorted in descending order by ρ. Correlations marked with ● signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant according to pairwise Williams significance tests.

Metrics	Stem.	RWS	P/R/F	r	ρ	τ	Metrics	Stem.	RWS	P/R/F	r	ρ	τ
S-ROUGE	-	-	-	0.432	0.463 ●	0.366	R-3	Y	Y	P	0.278	0.230	0.189
R-W	Y	Y	P	0.371	0.425	0.339	S-CENT	-	-	-	0.399	0.227	0.174
R-W	N	Y	P	0.369	0.423	0.337	R-2	N	N	F1	0.269	0.225	0.176
R-W	Y	N	P	0.355	0.419 ●	0.334	R-3	Y	N	F1	0.269	0.219	0.177
R-W	N	N	P	0.351	0.414	0.331	R-2	N	Y	F1	0.267	0.216	0.172
R-L	Y	Y	P	0.385	0.412	0.328	R-4	Y	N	P	0.258	0.214	0.176
R-L	N	Y	P	0.384	0.411	0.328	R-3	Y	Y	F1	0.265	0.214	0.174
R-L	Y	N	P	0.371	0.402	0.321	CENT-E	-	-	-	0.353	0.210	0.162
R-W	Y	Y	F1	0.374	0.401	0.317	R-1	N	N	F1	0.285	0.208	0.163
R-L	N	N	P	0.367	0.398	0.319	R-4	Y	Y	P	0.253	0.203	0.170
RWMD-R-C	-	-	-	0.459	0.397	0.310	R-3	N	N	P	0.256	0.202	0.166
R-W	Y	N	F1	0.362	0.395	0.311	R-4	N	N	P	0.234	0.201	0.167
R-W	N	Y	F1	0.369	0.394	0.312	R-4	Y	N	F1	0.251	0.200	0.163
R-W	N	N	F1	0.357	0.389	0.307	R-4	Y	Y	F1	0.243	0.194	0.161
R-L	Y	Y	F1	0.388	0.384	0.305	R-4	N	N	F1	0.229	0.189	0.157
R-L	N	Y	F1	0.382	0.379	0.301	R-3	N	N	F1	0.245	0.183	0.149
R-L	Y	N	F1	0.378	0.377	0.299	R-4	N	Y	P	0.228	0.180	0.154
R-L	N	N	F1	0.373	0.373	0.295	R-4	N	Y	F1	0.222	0.175	0.149
R-1	Y	Y	P	0.401	0.366	0.288	R-3	N	Y	P	0.246	0.172	0.145
RWMD-R-E	-	-	-	0.406	0.358	0.281	R-SU4	N	Y	R	0.202	0.172	0.131
RWMD-M-C	-	-	-	0.437	0.352	0.273	R-SU4	Y	Y	R	0.200	0.167	0.127
R-SU4	Y	Y	P	0.326	0.340 ●	0.265	R-3	N	Y	F1	0.238	0.164	0.137
R-SU4	N	Y	P	0.322	0.332	0.260	R-2	Y	Y	R	0.185	0.156	0.120
R-1	Y	N	P	0.367	0.324	0.254	R-S4	N	Y	R	0.189	0.154	0.118
BLEU	-	-	-	0.289	0.319	0.248	R-4	Y	Y	R	0.179	0.153	0.126
R-S4	Y	Y	P	0.318	0.318	0.250	R-2	Y	N	R	0.183	0.148	0.114
R-S4	N	Y	P	0.315	0.316	0.249	R-3	Y	Y	R	0.180	0.148	0.118
R-1	N	Y	P	0.352	0.309	0.243	R-4	Y	N	R	0.179	0.148	0.119
R-2	Y	Y	P	0.323	0.307	0.244	R-3	Y	N	R	0.182	0.147	0.116
R-2	Y	N	P	0.320	0.306	0.242	R-4	N	N	R	0.154	0.147	0.120
RWMD-M-E	-	-	-	0.383	0.305	0.239	R-S4	Y	Y	R	0.186	0.146	0.111
R-SU4	Y	N	P	0.308	0.302	0.236	R-4	N	Y	R	0.167	0.145	0.122
R-SU4	Y	Y	F1	0.307	0.297	0.231	M-CENT	-	-	-	0.392	0.144	0.107
R-SU4	N	N	P	0.306	0.296	0.232	R-W	N	Y	R	0.189	0.139	0.105
CENT-C	-	-	-	0.374	0.291	0.224	RWMD-L-C	-	-	-	0.271	0.138	0.105
R-1	Y	Y	F1	0.356	0.290	0.227	RWMD-L-E	-	-	-	0.231	0.138	0.105
R-SU4	N	Y	F1	0.303	0.290	0.226	R-SU4	N	N	R	0.179	0.137	0.104
R-S4	Y	N	P	0.303	0.289	0.226	R-2	N	Y	R	0.167	0.136	0.106
R-S4	N	N	P	0.301	0.284	0.223	R-SU4	Y	N	R	0.178	0.134	0.101
R-S4	Y	Y	F1	0.299	0.277	0.217	R-W	Y	Y	R	0.187	0.134	0.101
R-1	N	N	P	0.326	0.274	0.215	R-2	N	N	R	0.160	0.131	0.101
R-S4	N	Y	F1	0.295	0.273	0.214	R-S4	N	N	R	0.173	0.131	0.100
R-2	Y	Y	F1	0.299	0.268	0.211	R-W	N	N	R	0.183	0.130	0.099
R-2	Y	N	F1	0.298	0.265	0.207	R-W	Y	N	R	0.183	0.129	0.098
R-2	N	N	P	0.286	0.259	0.204	R-3	N	N	R	0.158	0.128	0.102
R-SU4	Y	N	F1	0.291	0.257	0.201	R-S4	Y	N	R	0.171	0.126	0.095
R-SU4	N	N	F1	0.289	0.253	0.197	R-3	N	Y	R	0.164	0.120	0.098
R-1	Y	N	F1	0.326	0.252	0.198	R-L	N	Y	R	0.167	0.109	0.082
M-ROUGE	-	-	-	0.357	0.248	0.191	R-1	Y	Y	R	0.187	0.102	0.078
R-S4	Y	N	F1	0.287	0.246	0.192	R-L	Y	Y	R	0.161	0.100 ●	0.075
R-3	Y	N	P	0.282	0.245	0.198	R-L	N	N	R	0.155	0.092	0.068
R-S4	N	N	F1	0.285	0.243	0.191	R-L	Y	N	R	0.153	0.087	0.065
R-1	N	Y	F1	0.307	0.243	0.190	R-1	Y	N	R	0.173	0.083 ●	0.062
R-2	N	Y	P	0.282	0.239	0.192	R-1	N	Y	R	0.156	0.082	0.062
							R-1	N	N	R	0.144	0.063	0.047

Table 7: Pearson (r), Spearman (ρ) and Kendall (τ) correlations of BLEU, 96 variants of ROUGE (R-*), centroid distance based metrics (CENT-C, CENT-E), WMD based metrics and methods described in this paper(S-ROUGE, M-ROUGE, S-CENT, M-CENT) with human assessment of the **Precision** metric in BIOASQ. The ROUGE variants are presented with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F). The results are sorted in descending order by ρ. Correlations marked with ● signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant according to pairwise Williams significance tests.

Metrics	Stem.	RWS	P/R/F	r	ρ	τ	Metrics	Stem.	RWS	P/R/F	r	ρ	τ
R-W	Y	Y	F1	0.430	0.519	0.384	R-2	N	N	F1	0.301	0.301	0.220
R-W	N	Y	F1	0.423	0.513	0.378	R-3	Y	N	F1	0.295	0.293	0.220
R-W	N	Y	P	0.403	0.511	0.378	R-3	Y	Y	F1	0.291	0.285	0.216
R-W	Y	Y	P	0.406	0.510	0.378	R-2	N	Y	F1	0.293	0.284	0.210
R-W	Y	N	F1	0.411	0.510	0.375	R-1	N	N	F1	0.368	0.282	0.204
R-W	N	N	F1	0.406	0.506	0.372	R-4	Y	N	P	0.275	0.276	0.212
R-L	Y	Y	F1	0.452	0.501	0.370	R-4	Y	Y	P	0.272	0.267	0.209
R-W	N	N	P	0.380	0.499	0.370	R-4	Y	N	F1	0.271	0.260	0.199
R-W	Y	N	P	0.383	0.499	0.371	R-4	Y	Y	F1	0.263	0.257	0.200
R-L	N	Y	F1	0.443	0.495	0.365	R-3	N	N	P	0.270	0.257	0.197
R-L	N	Y	P	0.420	0.495	0.366	CENT-E	-	-	-	0.444	0.247	0.179
R-L	Y	Y	P	0.423	0.492	0.365	R-4	N	N	P	0.245	0.243	0.191
R-L	Y	N	F1	0.435	0.488	0.359	R-3	N	N	F1	0.263	0.239	0.182
R-L	N	N	F1	0.429	0.485	0.356	R-4	N	Y	P	0.236	0.236	0.188
R-L	Y	N	P	0.400	0.474	0.351	R-SU4	N	Y	R	0.248	0.233	0.166
R-L	N	N	P	0.397	0.474	0.351	R-4	N	N	F1	0.241	0.231	0.180
MEAN-ROUGE	-	-	-	0.428	0.472	0.346	R-4	N	Y	F1	0.226	0.230	0.182
RWMD-R-C	-	-	-	0.511	0.424	0.309	R-SU4	Y	Y	R	0.250	0.226	0.161
R-1	Y	Y	P	0.447	0.420	0.307	R-3	N	Y	P	0.259	0.219	0.171
BLEU	-	-	-	0.316	0.418	0.306	R-2	Y	Y	R	0.235	0.212	0.152
R-SU4	Y	Y	P	0.352	0.413	0.301	R-4	Y	Y	R	0.206	0.212	0.161
R-SU4	N	Y	P	0.347	0.408	0.298	R-S4	N	Y	R	0.234	0.211	0.150
RWMD-M-C	-	-	-	0.511	0.404	0.294	R-3	N	Y	F1	0.250	0.210	0.163
RWMD-R-E	-	-	-	0.449	0.397	0.29	R-3	Y	N	R	0.222	0.206	0.151
R-S4	Y	Y	P	0.344	0.392	0.286	R-2	Y	N	R	0.236	0.206	0.147
R-S4	N	Y	P	0.339	0.390	0.286	R-3	Y	Y	R	0.215	0.205	0.151
R-SU4	Y	Y	F1	0.342	0.379	0.274	R-W	N	Y	R	0.262	0.203	0.144
R-1	N	Y	P	0.402	0.378	0.276	R-S4	Y	Y	R	0.235	0.202	0.144
R-SU4	N	Y	F1	0.336	0.374	0.271	R-4	Y	N	R	0.212	0.202	0.151
R-2	Y	N	P	0.340	0.370	0.272	R-4	N	Y	R	0.172	0.199	0.156
R-SU4	Y	N	P	0.331	0.369	0.269	R-W	Y	Y	R	0.261	0.195	0.138
R-2	Y	Y	P	0.347	0.369	0.275	R-SU4	N	N	R	0.229	0.193	0.137
R-1	Y	N	P	0.411	0.368	0.268	R-W	N	N	R	0.251	0.190	0.135
RWMD-M-E	-	-	-	0.444	0.367	0.267	R-SU4	Y	N	R	0.229	0.188	0.134
R-SU4	N	N	P	0.328	0.367	0.267	R-S4	N	N	R	0.223	0.186	0.132
R-1	Y	Y	F1	0.438	0.366	0.266	R-4	N	N	R	0.179	0.185	0.141
R-S4	Y	Y	F1	0.334	0.359	0.260	R-W	Y	N	R	0.251	0.185	0.131
R-S4	Y	N	P	0.325	0.358	0.260	R-2	N	N	R	0.207	0.185	0.131
R-S4	N	N	P	0.323	0.357	0.260	R-2	N	Y	R	0.201	0.184	0.133
R-S4	N	Y	F1	0.328	0.356	0.258	RWMD-L-E	-	-	-	0.301	0.184	0.131
R-2	Y	N	F1	0.334	0.342	0.250	R-S4	Y	N	R	0.222	0.180	0.128
R-2	Y	Y	F1	0.334	0.340	0.250	RWMD-L-C	-	-	-	0.389	0.177	0.127
R-SU4	Y	N	F1	0.325	0.338	0.244	MEAN-CENT	-	-	-	0.519	0.174	0.123
CENT-C	-	-	-	0.492	0.336	0.243	R-3	N	N	R	0.190	0.173	0.128
R-SU4	N	N	F1	0.322	0.335	0.243	R-L	N	Y	R	0.250	0.170	0.120
R-1	N	N	P	0.376	0.330	0.240	R-L	Y	Y	R	0.247	0.159	0.113
R-S4	Y	N	F1	0.321	0.328	0.237	R-3	N	Y	R	0.178	0.156	0.118
R-S4	N	N	F1	0.318	0.327	0.237	R-L	N	N	R	0.233	0.146	0.103
R-1	Y	N	F1	0.410	0.324	0.235	R-1	Y	Y	R	0.289	0.146	0.104
R-2	N	N	P	0.306	0.324	0.238	R-L	Y	N	R	0.231	0.138	0.098
R-1	N	Y	F1	0.387	0.323	0.234	R-1	N	Y	R	0.246	0.125	0.089
R-3	Y	N	P	0.301	0.315	0.238	R-1	Y	N	R	0.276	0.121	0.086
R-2	N	Y	P	0.303	0.304	0.227	R-1	N	N	R	0.241	0.100	0.070
R-3	Y	Y	P	0.301	0.302	0.231							

Table 8: Pearson (r), Spearman (ρ) and Kendall (τ) correlations of BLEU, 96 variants of ROUGE (R-*), centroid distance based metrics (CENT-C, CENT-E), WMD based metrics and methods described in this paper(MEAN-ROUGE, MEAN-CENT) with the **Mean** of all four human assessment metrics in BIOASQ. The ROUGE variants are presented with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F). The results are sorted in descending order by ρ. Correlations marked with • signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant according to pairwise Williams significance tests.

The inability of the method to statistically outperform the next best metric in DUC-2004 should not be taken as a weakness indicator. It is apparent that the method achieves the strongest Spearman and Kendall correlations across the board compared to other methods described in this paper including the ROUGE variants. This result is invaluable when creating a system with the demand to evaluate summaries quickly and independently of human assessors. The fact that the method has consistent performance within the same datasets, leads us to believe that the factor affecting it is mainly the human assessment criterion. Since HAS, the criterion used in DUC-2004, is a constructed measure and quite esoteric, as opposed to those used in BIOASQ, we propose that the method is robust when used with strictly defined assessment criteria. This conclusion is reinforced when examining the performance of the method on the mean of all the human assessment measures in BIOASQ, which is a constructed measure as well.

Worth mentioning is the fact that the centroid-based methods described in section 4.2 are routinely outperformed by the methods utilizing the ROUGE variants and by the ROUGE variants themselves. This indicates that the use of sentence centroids as a representation of summaries in evaluation tasks may not be useful, though further experimentation is needed.

Furthermore, the single task learning approach seems to be able to achieve assessment scores closer to the human. The networks we defined were unable to exploit the shared information across the measures of BIOASQ, which may indicate either issues with the training methodology, or the nonexistence of meaningful shared information which may have led to difficulties in convergence.

Finally, the study provides some insights in the use of correlation metrics. While the choice of Sprearmans ρ as the one we based our analysis on was quite arbitrary, it was reinforced by the fact that the Kendall correlation values agree on magnitude with the Spearman values, despite them being much lower. This is not the case with Pearsons r, which tends to report very high correlations even in cases when the other two do not. That is certainly in part because of the arbitrary assumptions the Pearson correlation makes about the data and the fact that it is restricted to linear correlations, which may not be the case for the dataset in question. It is not clear which correlation measure is better for the task of summary evaluation, but the fact that two of the three we used seemed to agree we propose the use of either, especially when considering the limitations of Pearsons r.

7 Conclusions

In this study we evaluated the performance of different metrics on the task of summarization evaluation on two different datasets specific to the task. The metrics include 96 ROUGE variants, other distance based metrics and 4 new metrics that we developed for the task. The metrics were evaluated based on correlations with human assessment scores, and their statistical significance was tested using the Williams test. The results reveal that the choice of a metric is dependable on the specifics of the task, specifically on the way human evaluation was conducted. Consistently, the best performance was achieved by combining all other metrics and treating them as input to an MLP architecture. This method, described in the study, presents both the strongest correlation with human assessment overall and also statistically outperforms all other metrics.

Moving forward, we would like to test the methods presented in this study to more datasets to determine whether they generalise even more. The AESOP task seems an ideal dataset to explore. We would also like to explore whether the removal of some of the combined metrics would reduce performance. Finally, we believe that using embeddings within the computation of the ROUGE metrics is an area which might yield exciting results.

References

Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Berlin, Germany, pages 114–118. http://anthology.aclweb.org/W16-2915.

Yvette Graham. 2015. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Trinity College, Dublin, pages 128–137. http://aclweb.org/anthology/D15-1013.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization .

M. Kusner, Y. Sun, N. Kolkin, and K. Q. Weinberger. 2015. From word embeddings to document distances. In *In Proc. of the 32nd International Conference on Machine Learning*. Lille, France, pages 957–966.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. Association for Computational Linguistics, Barcelona, Spain, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. Edmonton, Canada, pages 71–78.

C.D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1925–1930. http://aclweb.org/anthology/D15-1222.

Paul Over and James Yen. 2004. An introduction to duc 2004 intrinsic evaluation of generic new text summarization systems. n Proceedings of the Document Understanding Conference (DUC).

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Association for Computational Linguistics, Montréal, Canada, pages 1–9. http://www.aclweb.org/anthology/W12-2601.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 131–136. http://www.aclweb.org/anthology/P13-2024.

G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M.R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. .