DeepNet Based Summary Evaluation

Priyank Pathak , Susmit Wagle and Prof. Harish Karnick

[1]Department of Computer Science – IIT KANPUR

{priyankpathak50,waglesmit.uec21}@gmail.com {hk}@iitk.ac.in

Abstract. *Currently, Text Evaluation is limited to use of ROUGE, BLEU, F1Scores, which don't match with human judgment scores. Taking an image from computer Vision as an analogy for the document in a Linguistic space, we try to implement similar approaches for it in the hope of surpassing results of BLEU scores and ROUGE scores for text evaluation. Treating summarization process as a caption generation problem as implemented in computer Vision, thus a joint modal learning model for evaluating Headlines concerning the whole document (Twenty News Dataset, CNN-DailyMailDataset, and DUC Dataset)*

Keywords*: Text Evaluation, Summarization, NLP, Vision.*

1. Introduction

The project is based on the idea of taking advantage of the similarity between images and documents for caption/headline generation, thus the implementation of various Vision techniques from Vision. The presentation primarily focuses on the indulgence of DeepNets in summary evaluation.

Text evaluation of a summary refers to the Comparison of a summary with respect to its original/reference document and giving it a score, on how accurately it covers the topic. The reasons NLP is difficult is: a) No common notion of smallest unit in literature as opposed to a pixel in Vision. It can be characters, words, sentences or even paragraphs [1]. If sentence vector really the best strategy? (3-word sentences and 91 words having the same dense representation). b) No Universal tokenizer.Common tokenizers from Keras, NLTK have shortcomings. The common notion, "internet is full of articles, offers sustainable training data", is a myth, since hours and hours of pre-processing is needed, unlike Image.

The primary question one needs to ask when working with summarization is, what makes a good summary "good"? [2]
1) Coverage of important topics from the reference document
2) Good Coherent Structure
3) Generic if "Abstractive"
4) Non-redundant
5) Good Readability

While summarizing is still an open problem, the following are the question one needs to ask oneself before evaluating the summary. 1) What are the critical points/sentences in a reference document? 2) More sentences from "central idea" better than covering the entire document? 3) How to give an "Abstractive" summary, written with no "Extractive" words from the reference document, the same score as that from the extractive

one (generalized)? (for a valid scoring metric) 4) Can incorporating "Literary devices" help? 5) How to ensure grammatical readability? 6) How to keep a check on the coherent structures which prevent misinterpretations, even if two sentences are taken from the document without any context. 7) Assigning weights to sentences based on their importance/idea conveying power in the reference document? 8) How can the golden summary help if the scoring is taking place between the reference document and the generated summary? 9) Although Word order is important but is longest subsequence matching really the best method? 10) How does Humans create Summaries or do they Judge them? 11) What are current methods of evaluation? 12) What dataset to use for evaluation?

Our contribution

- Use of cross modality of caption generation for evaluation, using a Neural Network learning based architecture for evaluation.

2. Previous Works

We have gone through various document representation techniques and evaluation methods before starting our approach.

- Character based NLP processing and CNN Understanding Text from Scratch. [1]
- Detailed analysis of evaluation of text summarization [2]
- Modified Rouge Scores for Scientific Documents, Revisiting Summarization Evaluation for Scientific Articles [3]
- Details on Rouge Scores [4]
- Limits of Rouge Scores [5]
- Updated Rouge Scores [6]
- Fine Tuning used [7]
- Main architecture [8]

"The original ROUGE metrics show high correlations with human judgments of the quality of summaries on the DUC 2001-2003 benchmarks. However, these benchmarks consist of newswire data and are intrinsically very different than other summarization tasks such as summarization of scientific papers. We argue that ROUGE is not the best metric for all summarization tasks" [3]

- Jackknifing procedure : Given M references, we compute the best score over M sets of M-1 references. The final ROUGE-N score is the average of the M ROUGE-N scores using different M-1 references. The Jackknifing procedure is adopted since we often need to compare system and human performance and the reference summaries are usually the only human summaries available. Using this procedure, we are able to estimate average human performance by averaging M ROUGE-N scores of one reference vs. the rest M-1 references. [5]
- The metric returns an average of ROUGE scores over multiple reference summaries in order to avoid bias [5]
- Lot of different methods of evaluation (Rouge-1 , Rouge-2 , Rouge-N , Rouge-Skip... etc.)
- Compute the correlation between ROUGE assigned summary scores and human assigned summary scores. The intuition is that a good evaluation measure should assign a good score to a good summary and a bad score to a bad summary. The ground truth is based on human assigned scores.

With the DUC data, *Pearson's product moment correlation coefficients, Spearman's rank order correlation coefficients, and Kendall's correlation coefficients* between systems' average ROUGE scores and their human assigned average coverage scores using single reference and multiple references was computed.

Though utterance level was not reported in [4], the averaging was done to judge to the "peer".i.e. For all the summaries the scores were computed with respect to the peer where as we have computed with respect to the document that is we are judging the summaries with respect to the sumamries. Altough the Spearman Correlation makes the most sense i.e. for a given document and list of summaries we need to rank the summaries in the same order in wehich humans have ranked them.

ROUGE does not capture synonymous concepts, ROUGE expects system summaries to be identical to reference summaries, and ROUGE scores do not capture topic or subset coverage [2] (Solved by allowing synonym capture and topic capturing) Dataset used **TAC** [6] More problems highlighted in [5]

3. Approaches

3.1. Caption Generation on Joint Modal System

Use of CNN from the paper(CNN architecture for sentence classification [8]). It's the state-of-the-art for 10 class sentiment classification and has outperformed every model for our 18 class classification. Its role is equivalent to what VGG has in Vision, so similarly can be modified for various tasks. Feeding Document to CNN enables us to makes the size of input arbitrary long, a major drawback for LSTM (current headline generation method are restricted to LSTMs, so a CNN model can produce the same result while removing the input size constraints.)

Different Architectures were tried where one Neural Network was feed-ed the parent document and the other document was feed-ed the summary (once golden and once the garbage). The assumption when the model is feed-ed the original golden summary instead of the document the model is supposed to rate the summaries equally well, hence showing that the document is independent of the input. The architecture baseline used is shown in figure[1]. Model is trained on Mixture of CNN-DailyMail dataset and Twenty News Dataset and fine-tuned on DUC Dataset-2001.

The DUC dataset has much larger document size compared to Twenty News Dataset or CNN-DailyMail. The Max sized documents was **4633 words** and minimum size summary allowed was **10 words**. The tokenizer was custom regular expression based, while observing DUC2004 dataset. Github link for ***tokenizer***

The model initially shared the weight, but later the weights for the Document Neural Net and Summary Neural Net were separated. The differentiator was also removed since the model started to get confused and mean value soon got fixated on 0.5. The K max pool was insufficient to match the discrepancy in the text (readability check, where the words were randomly jumbled to create an incoherent unreadable English). Soon the Neural Net for readability were also separated because of 2 reasons:

- The K-max pool is independent of word position, so incoherency will never be detected, so the New Neural Net doesn't have K-max pooling only convolution.

- The idea of primarily using the CNN based approach was to filter essential words from the documents. Thus an incoherent document yet containing critical bits from the original document should be scored perfect one. The garbage summary was scoring high because of the coherency, and incoherent summary was scoring high because of topic overlap.

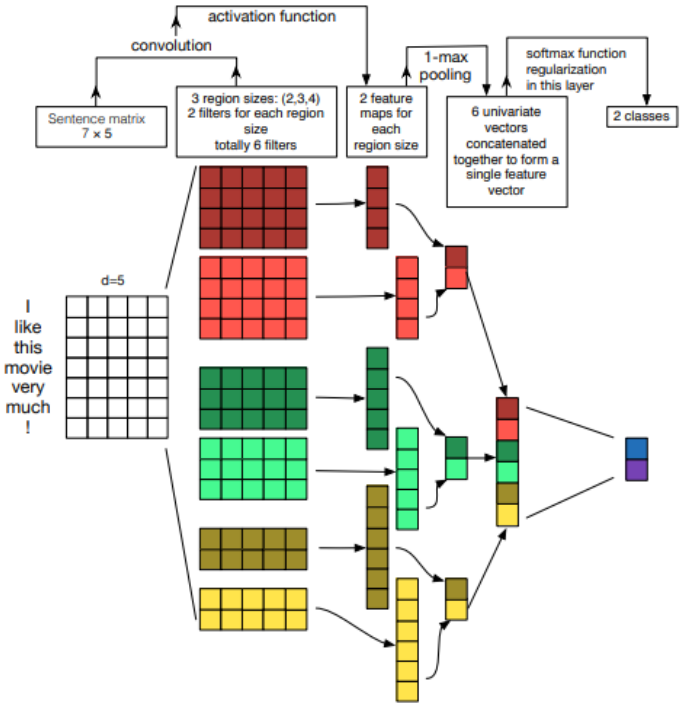

Figure 1. CNN architecture taken from A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification [8]

Model Details

- Dataset used is Twenty News Dataset (only those documents whose summary length was greater equal to 10 words) and CNN-Dialy Mail Dataset.
- Model has been later finetuned on DUC Dataset (2001) training documents and Duplicate Summaries.
- Different embedding for document and summary didn't perform that well so were kept same for Documents and Summary.
- The Readability Model was fed the embedding of the documents and summaries using the convolutional layers whose weights were shared for all input. The output was supplied to a Fully connected layer which again shared the weights. The output of the incoherent summary concatenated with the original document was given a score 0, and garbage summary and golden summary concatenated with the original document was given a score 1.
  The Idea was to take help from the original document in scoring the readability scores.

- The separate topic model consisted of applying filters of size [1,2,3...7] of 300-dimensional embedding. The weights were initially shared for the summary and document but were later separated for the document and summary.

- the output from each convolution were initially reduced using the global K-means but were replaced with stridden K means of size 5. All the outputs were merged/-concatenated.
- dropout was applied on the output of [document, golden summary] and [document, garbage/incoherent summary] which was later passed to the Fully connected layer of size Hidden layer of 50 nodes and then to 20 nodes and finally to 1 node. *The problem was treated as a classifier problem where the model was trained to differentiate between golden and garbage summary.* Note : Incoherent summary was still given a topic score of 1.

With Differentiator The figure shows how differentiator was used to give scores.

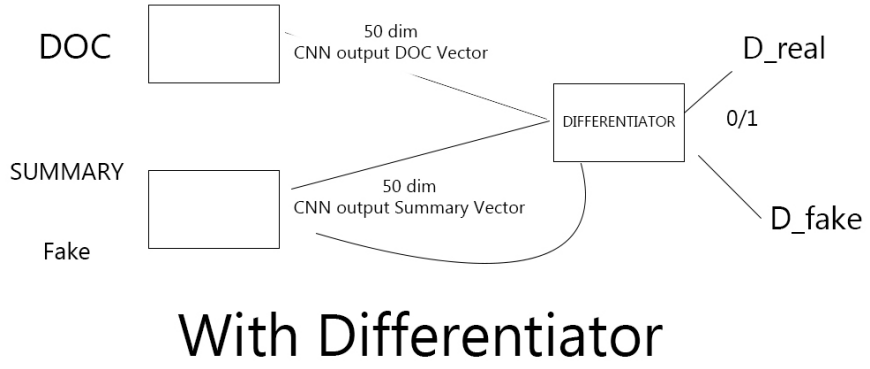

Figure 2. Differentiator based Model

3.2. Fine Tune

The whole implementation is taken from [7] Each layer has a different learning rate, where the last layer (L-th layers) has the maximum learning rate and each subsequent layer has a factor reduced by 2.6. The readability Neural Net has a factor of 10 instead since the loss shoots up to NaN for finetuning with the factor of 2.6

$$\mu_i = \mu_{i-1}/2.6$$

3.2.1. Slanted triangular learning rates (STLR)

Learning rate first linearly increases the learning rate and then linearly decays it according to the following update schedule,

$$freq_{cut} = [0.1 * (Number\ of\ Epochs)]$$

$$\text{p} = \begin{cases} epoch/freq_{cut} & epoch < freq_{cut} \\ 1 - ((epoch - freq_{cut})/freq_{cut} * (ratio - 1)) & epoch > freq_{cut} \end{cases}$$

$$\text{lr} = \text{lr}_{max} * (1 + p * (ratio - 1)/ratio)$$

Number of fine tuning epochs were set to 12 so cuttof frequency was $1.2 \approx 2\text{lr}_{max}$ for the topic model is 1e-2 and readability model 1e-4 ratio is taken to be 32

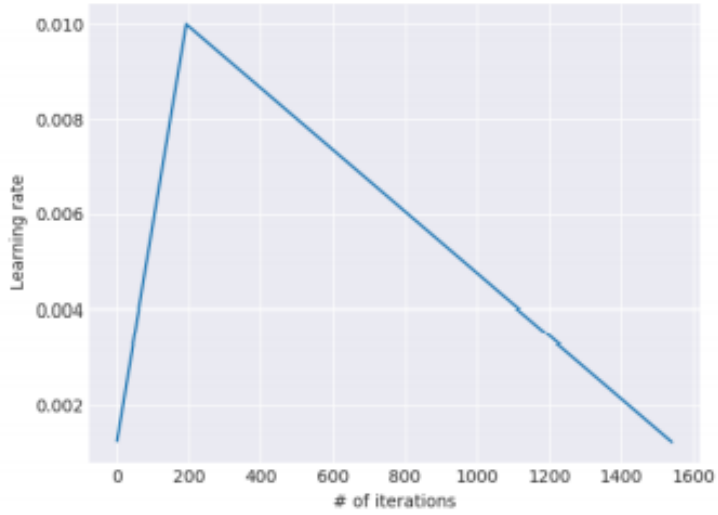

Figure 3. Funetuning elarning rate curve taken from: Universal Language Model Fine-tuning for Text Classification [7]

3.2.2. Intermediate layers Fine tuning

Intermediate layers were fine-tuned using concatenation of max pooling of layer and avg pooling of layer and Final layer and were passed to fully connected layer to train from scratch, where this model was learned to give exact scores, and the previous final layer was used as a classifier for golden and garbage summary.

3.3. Details of DUC 2001 Dataset

The dataset contains Train Documents (30 topics, each topic containing few documents and 100 words summaries (one document and one summary) along with other types of summaries like multi-document summaries), Trial Documents (4 topics each having the same format as that of Train Folder)
The Test Folder contains 30 topics each topic having some documents, corresponding **golden summaries** and **duplicate summaries**. Then generated summaries known as "**peer summary**", The baseline summaries are created by picking up the top 50-100 words. The coverage score is as follows :

The Sentences are called units, in the golden summary are taken as reference units. The peer/duplicate summaries units are also created (will denote it by generated units). Some units from the golden summary and generated units overlap and are termed as "marked units" and non overlapping units are termed as unmarked units.

First, three scores are Grammaticality, Cohesion (how sentences are related) And Organization (is the idea concise, even though may be wrong). , a.k.a readability scores.(a golden summary is possibly not even referenced)

S1: that ought to be in the model in place of something there. The idea that peer content expresses that is not present in the golden summary. (peer summary is better than the golden summary)

S2 : Unmarked PUs that don't deserve to be in the model, but are related to the subject. Topic overlap but not that much.

S3: unrelated to the subject of the model. Higher this number, higher the garbage content of peer summary.

Number of model units Number of peer units Number of unique peer units marked (overlapping content, higher this is, better is the summary) treat it as (number of unique peer units marked / Number of peer units) can penalize the garbage content in summary.

For each model unit : Number of peer units marked for this model unit The extent to which marked PUs express meaning of the current MU "-" if no PUs were marked or Sorted list of marked PU IDs

All scores $\in 0, 1, 2, 3, 4)$

Proposed Topic overlap scores :

$$\frac{ofuniquepeerunitscovered}{Numberofpeerunits} * \sum_{n=1}^{ofmodelunits} (\frac{contentoverlap}{Numberofmodelunits * 4)}$$

Total score will be a decimal from the range [0-1] (of unique peer units covered) / Number of peer units: fraction of peer units there were overlapping with the golden/ model summary.

(content overlap) / (Number of model units * 4): 4 denominators is just normalizing constant, scaling the marks to one. If a peer unit is not overlapping much with the model unit, then the overlapping content will be low for that model unit. Number of model units penalizes the peer if a model unit is missed, and

If all peer units are overlapping with the model units, that is, no garbage unit, and all overlap score are 4, from all model units are covered then only the overall score will be 1

Sample Golden and duplicate summary whose coverage score was very low :

$b_{bA}P890227 - 0016.txt$:

In 1988 the nation's more than 700 tornadoes caused 32 deaths, down from 59 a year earlier and the long-term average of 99. Meteorologists attribute the decline in deaths to increased public awareness of the storms. 1988 also witnessed the largest one-day outbreak of twisters in fourteen years when 57 tornadoes tore through four states on Mother's Day. Few in winter, tornadoes generally increase in March leading to an April average of 109, a average May peak of 166 and tapering in June to an average of 150. The typical tornado is about 50 yards wide and travels about two miles on the ground.

$b_{hA}P890227 - 0016.txt$:

In the transition from winter to spring, hot weather stirs the air and helps form thunderstorms and tornadoes, violently twisting winds reaching down from

thunderclouds. Tornadoes increase in March and peak in May, averaging 166 that month, but they can occur any month. November's average is 23 but last year unusually warm and wet weather helped trigger 121. Tornadoes are most likely between noon and sunset and least likely just before sunrise. Typically they are 50 yards wide and travel two miles on the ground, usually heading northeast, and often come in groups. Tornado deaths are down due to increased public awareness.

4. Results

4.1. Baseline results for ROUGE Scores

While Rouge-1 fails to account for incoherency, Rouge-L seems to be best among all matrices

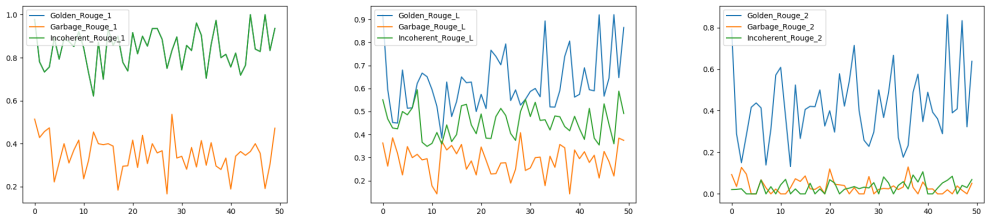

Figure 4. Rouge scores on incoherent, golden and garbage summary. Left most (Rouge1), middle (Rouge2) and Right Most (Rouge-L)

4.2. Differentiator based approach for evaluation

The model failed to clearly distinguish between garbage and golden summary. Instead it got confused and saturated to 0.5. Typical GAN maneuver, perfectly trained confuses between garbage and golden summary.

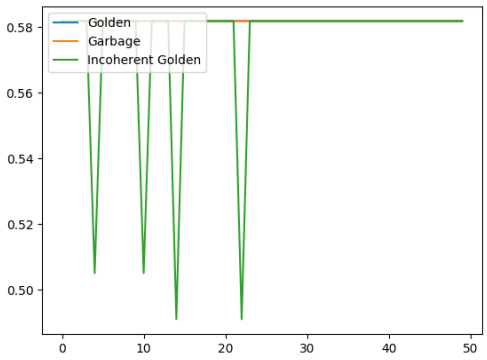

Figure 5. Model has got confused between garbage summary and golden summary.

4.3. No Need of golden summary

Coverage Score average pearson correlation			
original document	**p value**	**golden summary**	**p value**
0.01234295432	0.5057081329	0.01302110616	0.5064524149
0.1019048873	0.4440830296	0.1478918492	0.455998677

4.4. Analysis of How human treats a summary

The results shows even two humans can't have unique summaries for a given document. Human scores for duplicate summary compared to golden summary.

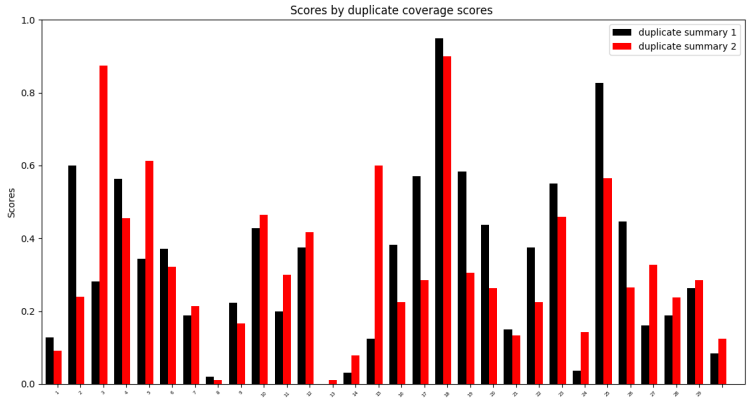

Figure 6. Randomly picked Sampled of original documents and duplicate documents for comparison of coverage scores, 1 represents complete overlap coverage score of one

4.5. K Means on Model

Strided K Means, performs better than global K-means, helps preserving more number of important concepts from the original documents rather just using one.

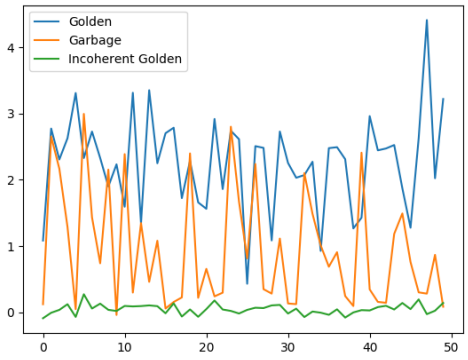

Figure 7. Results on Model having strided K means,

4.6. Separation of Model into Readability and Topic Coverage

Kindly refer to Figure 8. The model is now able to differentiate the golden garbage and incoherent summary, yet retaining high readable scores to the golden and garbage summary and at the same time penalizing the topic scores for the garbage summary.

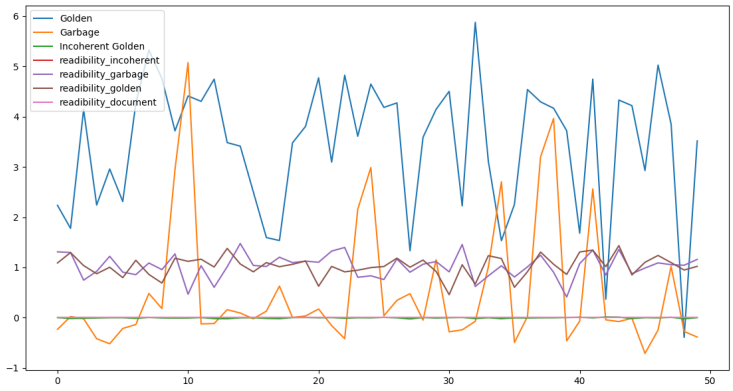

Figure 8. The different scores are from different model For Topic Coverage and Readability scores

4.7. Similarity loss function

Replacing Loss function to similarity loss function helped the model learn the difference yet couldn't perform well on given exact scores or scores in a similar manner. Rather training the model for scoring one for golden summary and 0 for garbage summary, the similarity loss function tends to separate the gap between scores of garbage and golden summary.

$$delta = Score_{real} - Score_{fake}$$

$$loss = log(1 + e^{-10 * delta})$$

Perhaps this model performed the best.

4.8. Comparison with Rouge Scores

Both the relations are uploaded here: **Pearson Correlation** and **Spearman Correlation**

Kindly refer to figure 10 and Figure 11.

4.9. Comparison with Rouge Scores

Pearson Correlation

5. Future Proposals

- Training on Sequential Degradation for making the model learn the exact topic coverage rather 0-1 classification.
- Time Stamps check, to prove the deepnets are faster than Rouge Scores, on CPU cores.

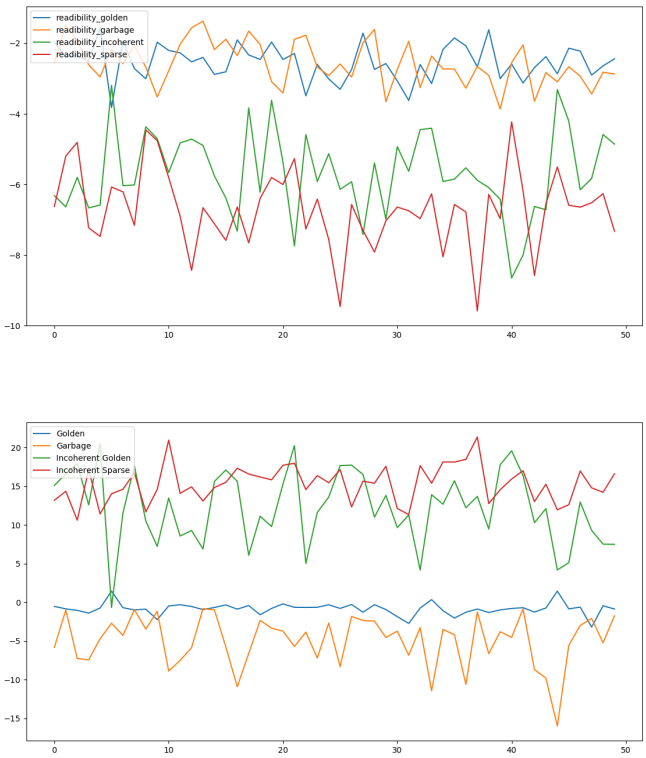

Figure 9. The different scores are from different model For Topic Coverage and Readability scores

- Correlation with Rouge Scores, to show Model is best matches to Rouge-L scores (Ideal Rouge Scores).
- Word replacement with Paraphrase/synonyms, to prove the use of embedding ensure that the synonymous well handled.
- Other Comparisons matrices: *LexRank , Latent Semantic Analysis, Maximal Marginal Relevance, Citation based summarization, Using frequency of the words*, etc.
- Indulgence of more DUC Dataset DUC-2002 and Duc-2003, for extensive comparisons.ons.

References

[1] Text Understanding from Scratch
https://arxiv.org/pdf/1502.01710.pdf

[2] EVALUATION MEASURES FOR TEXT SUMMARIZATION, Detailed analysis.
http://www.cai.sk/ojs/index.php/cai/article/viewFile/37/24

[3] Revisiting Summarization Evaluation for Scientific Articles
https://arxiv.org/pdf/1604.00400.pdf

[4] Rouge Scores
http://www.aclweb.org/anthology/W04-1013

Pearson Coorelation		DUC -2001 (Single references)					
		Grama (ref: golden summary)		Coherence (ref: golden summary)		Org (ref: golden summary)	
Recall	Rouge-L	0.09113345412	0.4874803897	0.1433698397	0.4862240495	0.1667306399	0.429525328
	Rouge-1	0.1023588783	0.4846399136	0.1801280144	0.4651876693	0.1968104906	0.4369594289
	Rouge-2	0.1056955313	0.5145372628	0.1706432337	0.4734532265	0.2039262786	0.4237958486
Precision	Rouge-L	0.1169101075	0.5000302117	0.1984782931	0.5012272063	**0.2244595302**	0.4190289691
	Rouge-1	0.1073995773	0.4936795237	0.1073995773	0.4759162007	0.2086749664	0.4149610732
	Rouge-2	**0.1177924277**	0.5222318765	0.179503783	0.4816733211	0.2101181833	0.4332935386
F-Score	Rouge-L	0.1064209974	0.4961686293	0.1788777091	0.4990540512	0.2091839161	0.4227030799
	Rouge-1	0.1082442628	0.486109285	**0.2046985326**	0.4618599032	0.2178792257	0.4214505184
	Rouge-2	0.1115463632	0.5199267965	0.1782437894	0.4735722033	0.2110920287	0.4255920477
My model	Readability model	0.0148074365	0.4798261289	0.05628055412	0.5167398457	0.02931587397	0.469273564
	topic model	-0.00159963140	0.5197596263	-0.02230742045	0.4967031593	-0.05385048958	0.4528556399

		DUC -2001 (Single references)			
		Coverage (ref: golden summary)		Avg Read	
Recall	Rouge-L	0.5093470897	0.216222445	0.168884946	0.4336092672
	Rouge-1	0.571892982	0.1485124985	0.2012555508	0.4033737488
	Rouge-2	0.5667591648	0.1541194047	0.2047154907	0.4203102441
Precision	Rouge-L	0.5821817788	0.1461857657	0.2293646587	0.4183285931
	Rouge-1	0.615621866	0.1229614129	0.2179262547	0.4154829513
	Rouge-2	0.5955243676	0.1336240617	0.2136373172	0.433293582
F-Score	Rouge-L	0.5784148502	0.1495283319	0.2137660505	0.4215376675
	Rouge-1	**0.6346989795**	0.1106120286	**0.2250248384**	0.3957574135
	Rouge-2	0.5914254177	0.1371915675	0.2128783181	0.4249507411
My model	Readability model	0.01302110616	0.5064524149	0.04191710994	0.4425622232
	topic model	0.1478918492	0.455998677	-0.04047111442	0.4589216439

Figure 10. Pearson correlation for grammatically, coherency, organization, Average readability and Coverage Score

[5] Analysis of ROUGE Scores
http://www.aclweb.org/anthology/E17-2007

[6] Analysis of ROUGE Scores, modifications and limitations
https://arxiv.org/pdf/1803.01937.pdf

[7] Fine Tuning used for modifying the model according to the DUC Dataset
https://arxiv.org/pdf/1801.06146v4.pdf

[8] A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification
https://arxiv.org/pdf/1510.03820.pdf

		DUC -2001 (Single references)					
		Gram (ref: original doc)		Coherence (ref: original doc)		Org (ref: original doc)	
	Rouge-L	**0.2752631663**	0.3844898091	0.04617140902	0.5126535197	0.02335574825	0.4741372973
	Rouge-1	0.2391610512	0.4335552949	0.04567813673	0.5521310894	0.01447231155	0.4974779756
Recall	Rouge-2	0.241448023	0.4143323103	0.05861821197	0.5104475773	0.03071868905	0.4713626451
	Rouge-L	0.1432011056	0.5087961957	0.05011194464	0.5471209495	0.037295774	0.5021369784
	Rouge-1	0.1113770878	0.5395657316	0.02672176355	0.5121323271	0.01612596312	0.486398991
Precision	Rouge-2	0.146422497	0.5037474333	0.05666601956	0.5316195856	0.04399576824	0.4851428258
	Rouge-L	0.1408757622	0.5112364507	0.04979042017	0.5462869189	0.03702964149	0.5017891178
	Rouge-1	0.1253213781	0.5356696638	0.03014258429	0.510843233	0.01963805918	0.4835104009
F-Score	Rouge-2	0.1582213285	0.4962537308	**0.05987177634**	0.5268962711	**0.04740840711**	0.4808395557
My model	Readability model	0.0190213746	0.4768706779	0.05885612894	0.5191254735	0.02970310597	0.4715300338
	topic model	0.01723491025	0.5006542603	-0.01312166696	0.5174918193	-0.03425261011	0.4588922494

		DUC -2001 (Single references)			
		Coverage (ref: Original Doc)		Avg Read	
	Rouge-L	0.001055050966	0.4657440519	0.09864908541	0.4334170868
	Rouge-1	0.003578297523	0.4589626995	0.08936444516	0.4760688863
Recall	Rouge-2	-7.287664332	0.4660326924	**0.102852767**	0.424111642
	Rouge-L	0.1120506986	0.4745690131	0.07840686441	0.4977964394
	Rouge-1	0.09187527561	0.4728682702	0.0524940958	0.4845731133
Precision	Rouge-2	0.1089274285	0.4726248553	0.08515155083	0.4832667082
	Rouge-L	**0.1124842406**	0.4745896902	0.07741548604	0.4976079805
	Rouge-1	0.08993842953	0.4718164294	0.06006838384	0.4829482162
F-Score	Rouge-2	0.1072127208	0.4713373296	0.09175196695	0.4749341714
My model	Readability model	0.01234295432	0.5057081329	0.04380266776	0.4402223848
	topic model	0.1019048873	0.4440830296	-0.02018221921	0.4521634327

Figure 11. Pearson correlation for grammatically, coherency, organization, Average readability and Coverage Score, with respect to original document