SEA: Summary Evaluation of Academic Publications with Unsupervised Methods

Shivani Jadhav¹, Sanket Joshi¹, Sabine Wehnert², Sayantan Polley¹, Erasmo Purificato¹,², Ernesto W. De Luca¹,², and Andreas Nürnberg¹

¹ Otto von Guericke University Magdeburg
² Georg Eckert Institute Leibniz Institute for International Textbook Research

Abstract. We live in an information paradox where the problem is often that we are drowning in information while starving for knowledge. Artificial Intelligence can address this problem using summarization techniques, such as extractive summarization, which can save time and effort. While there has been a plethora of research on generating summaries, we focus on the specific case where the problem is aggravated by the absence of ground truth. Also, there can be multiple candidates for a suitable summary when the text contains paraphrasing sentences and descriptions on different abstraction levels. Hence, datasets with only one correct ground truth may only paint a fraction of the whole picture. The lack of robust evaluation methods challenges the reliability of the current summarization techniques. In this paper, we discuss three alternative unsupervised methods as an alternative for supervised summary evaluation. First, we propose the “Relative Clustering Comparison Score”, consisting of three individual scores for cluster evaluation (Adjusted Rand Index, Mutual Information, and Completeness). Second, a further method assesses the ranking the sentences obtain based on cosine similarity, and it is called the “Relative Ranking Comparison Score”. Third, we employ a semantic similarity metric called BERTScore. We mostly observe a correlation of our regarded unsupervised metrics with human judgement and unveil challenges we face while working with academic paper summarization datasets.

Keywords: Summary Evaluation · Clustering · Unsupervised evaluation · Extractive text summarisation

1 Introduction

Text summarization is a technique to get a shorter version of a given document which describes the gist of that document. There are two types of text summarization, extractive text summarization and abstractive text summarization. Abstractive text summarization is a technique where a new
text is generated which may paraphrase the original text. Extractive text summarization on the other hand selects the most important sentence(s) from the document to use it or them as the summary. In our work, we focus on text summarization of scientific publications which may have important information clustered in some parts (e.g., the introduction) due to their common structure. Under the assumption that those works do not contain much redundancy, we deem extractive text summarization techniques to be a suitable choice, which reflects in the large amount of published research on this subject. Existing datasets exploit author-generated paper highlights offered by ScienceDirect\textsuperscript{3} as a basis for evaluation, since they often exhibit a high overlap with a sentence from the original paper content. We illustrate the anatomy of the publications within those datasets in Figure 1. All papers contain an Abstract, an Introduction, Highlights, and further sections.

![Fig. 1: Anatomy of the publications in the ScienceDirect datasets.](https://www.sciencedirect.com/)

Figure 2 depicts the general workflow leading to the unsupervised evaluation when using highlights, which first starts with data collection and data preprocessing with feature selection. Then the entire document is converted into various sections according to the anatomy. All Sections and in addition also only the Introduction are passed to the extraction phase. Then the extracted summary along with the abstract and highlights are passed on to the evaluation phase.
One of the major challenges to obtain good machine-generated summaries is a lack of a robust evaluation approach. The approaches that are widely used for evaluating the generated summaries require a reference summary, such as the aforementioned highlights. However, there are a lot of documents whose reference summary is not available or not the only valid option. There can be many good candidates for a generated summary, so that there is a need to evaluate the generated summaries in a more generalised manner. Hence, for evaluation, along with the existing evaluation approaches for summarizers, we are evaluating scoring techniques which do not require the reference summaries, i.e., completely unsupervised. In the following, we list our contributions:

1. We show that the spatial distribution of the highlights from commonly used datasets in extractive summarization of scientific papers is not well-suited for a supervised task on the whole content.
2. We adapt an unsupervised approach for extractive summary evaluation and discuss why this evaluation method can be more fruitful than the supervised one.

The remainder of this work is as follows. The second section gives a brief overview of related work for extractive text summarization and for evaluation of generated summaries. After that, we briefly describe the necessary background to understand the summarization and evaluation techniques we employed. The fourth section covers fundamental concepts of this work. Section 5 contains the procedure for data collection and pre-processing, followed by the procedure for generating the summaries and a brief description about how to evaluate them. This paper focuses with extractive text summarization. However, it is also possible to test the proposed evaluation methods on summaries generated by abstractive methods. The results are mentioned in section 6 along with an in-depth interpretation of them. We conclude and state future work in the last section.
2 Related work

In this section, we first discuss popular extractive summarization techniques. Then, we focus on research works performed on scientific papers, laying the foundations for the academic paper summary evaluation part.

Vodolazova et al. [15] describe a basic sentence scoring technique of using Term Frequency - Inverse Document Frequency (TF-IDF) after applying a standard pipeline for preprocessing on the text. They mention that certain features should be given more weight, such as upper-cased words or font style changes (e.g., bold, italic). Given our corpus of research papers, the impact of the words in the title and of keywords can also be amplified. In this work, we employ this sentence-level TF-IDF approach.

In 2004, Mihalcea and Tarau proposed an approach to extract sentences for the summary using the TextRank approach [10]. In this work, TextRank is used to perform firstly keyword extraction and then sentence extraction which can be used for text summarization.

A fair amount of work has also been done on the extractive text summarization of scientific papers. Collins et al. have introduced a dataset of scientific papers with 10,000 publications in the training set and 150 papers in the test set, which they named CSPubSum [3]. Moreover, they have built a deep learning architecture, SAFNet, which applies a Long Short-Term Memory Network (LSTM) as a sentence encoder in addition to an abstract vector and common features (e.g., the title score measuring the overlap between the title and the reference sentence). While such task-specific deep learning models generally require ground truth for training, we investigate in this work whether we could still evaluate their performance using unsupervised metrics based on similarity.

Cagliero and La Quatra also worked on text summarization of scientific articles wherein they focus on extracting the highlights from scientific articles using a regression-based approach with n-gram overlap features [2]. They compare their approach with different existing supervised and unsupervised approaches of summary generation on an existing dataset and also on their own datasets, among which we find AIPubSumm for our experiments.

Several different evaluation metrics for extractive text summarization have been presented and exploited during the years. ROUGE is the most commonly used metric to compare the generated summaries with the reference summaries, taking into consideration the exact words [8]. There are variants of ROUGE, namely ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. Kusner et al. proposed Word Mover’s Distance [6], based on the idea that distances between the word vectors are to some extent semantically meaningful. Recently, the scoring metric BERTScore was proposed [17]; it exploits the contextualised word embeddings and is based on computing...
cosine-similarity. In 2003, Ahmad et al. introduced the approach of evaluating the summaries by automatically categorizing them together with the full text into clusters and plotting them on 2Dmap [1]. The idea is that if the summaries and the full text are assigned to the same position on the map, then the summary can be considered as a good representation of the whole text. One of our proposed evaluation techniques, “Relative Clustering Comparison Score”, is inspired from this categorization approach. However, unlike our adaptation, Ahmad et al. exploit the labels to get the final score and check whether the generated summaries also predict the same labels as their respective content. Minimum edit distance (or word error rate) [7] has been used to evaluate the generated summaries either by calculating the edit distance at a word level (TER) [13], stem level (ITER) [12] or even at the character level (CHARACTER) [16], (EED) [14], requiring a reference summary. Our second proposed evaluation method, “Relative Ranking Comparison Score”, is based on the minimum edit distance, but without any need of a reference summary: we get the minimum edit distance by comparing the rankings of the content of a document and rankings of the generated summary of that document.

3 Background

After setting our scope in relation to the related work, we continue with fundamental concepts for extractive summarization and its evaluation.

3.1 Extractive Summarization

Sentence level TF-IDF Our first approach for computing the sentence score is using TF-IDF on the sentence level. Each sentence is considered as one document and the rest of the paper as a whole corpus. The TF-IDF scores for all words in a sentence are summed up and normalized according to the length of the sentence. Jones has used a keyword-based score as an additional feature for scoring the sentence with the idea that if a sentence has more keywords then that sentence is comparatively more important than other sentences [5]. We also adapted this idea and used the keywords from the keywords section. If the word is a keyword then we assign a higher weight. The idea of using keywords can also be extended to using the words from the title of the paper.

TextRank The second approach is TextRank [10]. The concept of Textrank is adapted from the concept of the Google PageRank algorithm which was used to rank the webpages. The idea behind TextRank is it assumes that
the rank of a sentence in a document is dependent on the importance of a sentence suggested by other sentences in terms of links. As TextRank is a graph-based approach, each sentence in the document is considered as one node, and the sentences of the graph and the weights are calculated using the similarity between the sentences and a weighted edge is formed between the sentences. We apply the PageRank algorithm on the weighted graph and is run until it converges. The scores for each node are obtained and as each node corresponds to one sentence, we get a score for each sentence.

![Diagram of summary extraction process using ensemble method]

**Fig. 3: Summary extraction process using ensemble**

**Ensemble method** For generating the summaries, we came up with one more approach that combines the score for each sentence given by the TF-IDF method and the score given by the TextRank method. Fig.3 describes the summary generation process using ensemble method. In the ensemble model we used TextRank which is strengthened using a pretrained embedding. We add both the scores and we get a cumulative score for every sentence, which is afterwards normalized. This method is an ensemble of the two approaches that are mentioned above.
LexRank LexRank is another approach based on Eigenvector centrality in graphs. The idea behind computing Eigenvector Degree Centrality is that every link to and from a node is considered as a vote that determines the overall value of the node. A sentence that has a connection to a high-scoring Eigenvector centrality score will contribute or in other words, is more important than a sentence that has a connection to a low-scoring Eigenvector centrality score. Basically, it calculates the impact each node has in the graph. We use LexRank to find the most central sentences that can be included in the summary.

3.2 Extractive Summary Evaluation

ROUGE For extractive text summarisation, ROUGE [9] is most relevant. N-gram recall between the generated summary and target summary is ROUGE-N. If N=1, it basically counts the number of common words in both generated and target summaries and calculates the recall based on that. There are various variants of ROUGE namely, ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. So we will calculate the ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S for the generated summary using the Highlight sentences. However, ROUGE being an n-gram based approach does not take into account the semantic dependencies between the words.

BERTScore The BERTScore [17] applies recent advancements in distributional semantics to evaluate machine-generated text. Using contextual word embeddings from a BERT model, each token of a reference sentence is encoded and compared to each token of a candidate sentence. For each token of the reference sentence, we compute the IDF-score and multiply it by the highest cosine similarity with a token from the candidate sentence. This is divided by the product of all IDF-scores. As a result, we obtain a metric with the same range of -1 to 1 as the cosine similarity, which can be optionally rescaled for the interval between 0 and 1.

4 Proposed approach

One of the most important aspects of research is the evaluation. After getting the summaries of the documents, we need to check how accurate the obtained summaries are. Considering our documents are research papers, we can use the abstract part of the research paper as a summary in itself. Also, as mentioned previously we have highlights which can have significant overlaps with sentences in the research paper. We pass the extracted summary along
One of the challenges with the summarization task is that there can be various candidate summaries for the same original content. Hence, we need something which would not just evaluate the extracted summary with respect to a given set of finite reference summary sentences. We are proposing two methods called "Relative Clustering Comparison Score" and "Relative Ranking comparison Score" which are based on the hypothesis:

"If the syntactic and semantic similarity relations among extracted summaries are preserved in the same way as the syntactic and semantic similarity relations among the documents, it can be stated that the extracted summaries are capturing the gist of the documents in a proper way."

**Clustering based Evaluation** [1] Aligning to the proposed hypothesis, this approach is based on the assumption that if the candidate summaries form the same clusters as the summaries drawn from all sections (excluding the abstract and author highlights) then the extracted summaries are good."

Figure 4 describes the clustering-based evaluation. First we need to generate the summary for all the research papers. We will vectorize the extracted summary, all sections, and the abstract. The vectorized abstract and vec-
torized all sections are clustered by setting the same number of clusters and the obtained clusters are compared. This is done to prove the hypothesis. Similarly, the vectorized extracted summary and all sections are clustered. This is done to evaluate the extracted summaries. We pass them pairwise so that we get the optimal number of clusters for the entire pair.

**Relative Minimum Edit Distance** Figure 5 describes this approach which is based on the assumption that if the similarity ranking obtained for a candidate summary of a document is the same as the similarity ranking obtained by considering all sections of the document, the extracted summary is considered to be good.

For a document from all sections, a ranking is given to all other documents based on the cosine similarity. Then we obtain the ranking for the vectorized extracted summary. The rankings are then compared using the minimum edit distance. An average minimum edit distance score is obtained. This score is converted it to a similarity score:

\[
\frac{1}{1 + \text{AvgMinimumEditDistance}}
\]

### 4.1 Proving the hypothesis

To prove the hypothesis we use the AIPubSumm dataset as we have domain knowledge in the field of Artificial Intelligence. The extracted summaries are evaluated using the proposed approaches "Relative Minimum Edit Distance" and the "Relative Clustering Comparison Score". The obtained scores are compared with the metrics that use a supervised setting, ROUGE-L and BERTScore.

Another way to prove the hypothesis is by using the abstract of the papers. Similar to what was done with the extracted summaries, the abstract of the AIPubSumm papers is evaluated using the proposed approaches, "Relative Minimum Edit Distance" and the "Relative Clustering Comparison Score". The domain experts evaluate the obtained summaries by marking the summaries as relevant and irrelevant. The precision and recall scores are obtained for papers with a high ROUGE-L score and low ROUGE-L score for further analysis.

### 5 Evaluation

#### 5.1 Experimental Setup

We discuss the experimental setup in three parts: data collection and preprocessing, summaries extraction, and generated summaries evaluation.
Data collection and preprocessing The experiments were run on two datasets which consist of scientific papers. In addition to all the usual sections of a research paper, along with the abstract, both the datasets have highlights which consist of important sentences of that paper according to its author. After analysing the datasets, it can be concluded that the highlights are not always exactly extracted from the paper but can also be phrased by the author. By performing crawling on the ScienceDirect website, as done by Collins et al. [3] and Caglierio et al. [2], the dataset is created by separating out Highlights, Abstract, Introduction and one with all sections of a paper (excluding Abstract and Highlights) for all the papers. For AIPubSumm, one paper was not considered as it has been redirected. So the experiment was performed with 65 papers from AIPubSumm and 150 papers from CSPubSum. All the text was preprocessed by first doing tokenization, followed by stopwords removal and lemmatization. The domain-knowledge based experiments were done using a subset of AIPubSumm dataset. As we are dealing with extractive text summarization, similar to the authors of these paper, highlights sentences can be considered as the ground truth.

Summaries extraction The summaries are generated first from all the sections of the paper and then only from the introduction part of the paper using unsupervised approaches. The number of sentences for the generated summaries are varied from 1 to 10 for evaluation purpose. The extractive text summarization techniques that are used are LexRank, TextRank, Sentence-level TF-IDF, TextRank with Glove Embeddings and an Ensemble of Sentence-level TF-IDF and TextRank.

Generated summaries evaluation After generating the summaries for all the papers using one of the aforementioned methods we use ROUGE-L and BERTScore for evaluating for the generated summaries with highlights as the reference summaries (ground truth).

Then the summaries are evaluated using the proposed unsupervised evaluation approaches “Relative Clustering Comparison Score” and “Relative Ranking comparison Score”. To validate the proposed evaluation approaches, we try to find the correlation between the scores given by existing evaluation methods and scores given by the proposed approaches.

The other way to prove the proposed hypothesis is by using abstract section of the documents on the assumption that abstract of the paper gives the overall gist of the paper. To do that, all sections and abstract of all the papers are clustered using K-means with and without applying Principal Component Analysis (PCA). The clustering labels of all sections are used as the ground truth and compared with the clustering labels obtained by the abstracts to get the Rand Index, Mutual Information and Completeness
"Relative Clustering Comparison Score". We also get the "Relative Ranking Comparison Score".

To evaluate the generated summary, we cluster all sections and the generated summaries of all the papers and obtain the “Relative Clustering Comparison Score” and “Relative Ranking Comparison Score”. We expect the scores to be at par with the scores obtained by using the Abstract.

5.2 Results

Unsupervised Methods Inclusion of keywords from the papers did not have any significant change in the overall performance of summary generation technique due to a few obvious reasons like keywords may be not be present in their exact form in the content of the paper, a particular keyword can be referred to by using it’s synonym or by a pronoun.

<table>
<thead>
<tr>
<th>$k$</th>
<th>TextRank</th>
<th>TextRank_Emb</th>
<th>$k$</th>
<th>TextRank</th>
<th>TextRank_Emb</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.1784</td>
<td>0.1879</td>
<td>3</td>
<td>0.1699</td>
<td>0.1798</td>
</tr>
<tr>
<td>4</td>
<td>0.1782</td>
<td>0.1898</td>
<td>4</td>
<td>0.1650</td>
<td>0.1851</td>
</tr>
<tr>
<td>5</td>
<td>0.1740</td>
<td>0.1926</td>
<td>5</td>
<td>0.1584</td>
<td>0.1902</td>
</tr>
<tr>
<td>6</td>
<td>0.1749</td>
<td>0.1836</td>
<td>6</td>
<td>0.1531</td>
<td>0.1845</td>
</tr>
</tbody>
</table>

Table 1: AIPubSumm

Table 2: CSPubSum

In Table 1 and Table 2, the ROUGE-L values for summaries generated using TextRank and TextRank that uses an embedding using only Introduction section with varying sentence count ($k$) is shown for AIPubSumm and CSPubSum dataset respectively. From the values it can be seen that TextRank performance is improved if an embedding is used. Hence, the ensemble uses TextRank_embedding to generate the summaries for Human evaluation.

The graphs in Fig. 6 show that the ensemble method which we are proposing gives the highest ROUGE score when the number of sentences are greater than 4. After doing the data profiling, we found that there are total 263 highlights from the 65 papers. The average number of sentences present in the highlights section is 4.04. Therefore, a good ROUGE score when $k=4$ and $k=5$ can be considered significant. Hence, for the Human evaluation we use the ensemble technique to generate the summaries.

We find that a significant amount of the highlights is generated from the introduction part of the research papers, see Figure 7.

Human Evaluation We evaluated 13 of the extracted summaries from the AIPubSumm dataset with the help of five academic researchers. All of
them have at least a Bachelor of Science in Computer Science, three possess a Master’s degree and substantial professional experience. The experiment was designed by selecting outstanding papers based on their ROUGE-L scores for both the upper and lower end of the spectrum and also on summaries generated from the Introduction or all sections, respectively. We show the results in terms of precision and recall in Table 4.

Table 4 shows the ROUGE-L F1 score, which is computed between the highlights and the extracted sentences. Below this score, we list four other scores from the human evaluation: F1, Precision, Recall and Fleiss’ Kappa. We can directly compare the F1 scores between the ROUGE-L F1 metric and the human judgement for Introduction-based summaries and all section-based summaries, respectively. Fleiss’ Kappa serves as an indicator for inter-annotator agreement. In addition, when we consider the other Table 4, we can see values in the range of 0.796 to 0.903. The BERTScore within this range can also serve as an indicator for semantic similarity. Considering both tables, for example we learn from the results that there is no clear winner regarding the basis for summaries. Although we have often a higher score for the summaries generated from all sections, the summaries from the Introduction also offer useful results. Consider document 3 in the introduction part, where we compute a ROUGE-L F1 score of 0.35, while the human
Table 3: Results of instance-based evaluation.

<table>
<thead>
<tr>
<th>Metric \ Doc.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROUGE-L F1</td>
<td>0.07</td>
<td>0.32</td>
<td>0.35</td>
<td>0.33</td>
<td>0.08</td>
<td>0.66</td>
<td>0.03</td>
<td>0.30</td>
<td>0.06</td>
<td>0.09</td>
<td>0.10</td>
<td>0.24</td>
<td>0.18</td>
</tr>
<tr>
<td>F1</td>
<td>0.07</td>
<td>0.38</td>
<td>0.84</td>
<td>0.33</td>
<td>0.36</td>
<td>0.76</td>
<td>0.30</td>
<td>0.47</td>
<td>0.20</td>
<td>0.32</td>
<td>0.00</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>Precision</td>
<td>0.12</td>
<td>0.36</td>
<td>0.72</td>
<td>0.32</td>
<td>0.36</td>
<td>0.72</td>
<td>0.28</td>
<td>0.44</td>
<td>0.40</td>
<td>0.30</td>
<td>0.12</td>
<td>0.20</td>
<td>0.28</td>
</tr>
<tr>
<td>Recall</td>
<td>0.05</td>
<td>0.40</td>
<td>1.00</td>
<td>0.40</td>
<td>0.27</td>
<td>0.80</td>
<td>0.33</td>
<td>0.50</td>
<td>0.13</td>
<td>0.35</td>
<td>0.00</td>
<td>0.32</td>
<td>0.25</td>
</tr>
<tr>
<td>Fleiss' Kappa</td>
<td>0.69</td>
<td>0.58</td>
<td>0.75</td>
<td>0.76</td>
<td>0.32</td>
<td>0.68</td>
<td>0.30</td>
<td>0.27</td>
<td>0.39</td>
<td>0.25</td>
<td>0.68</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>All Sections</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROUGE-L F1</td>
<td>0.24</td>
<td>0.41</td>
<td>0.26</td>
<td>0.19</td>
<td>0.11</td>
<td>0.40</td>
<td>0.15</td>
<td>0.19</td>
<td>0.12</td>
<td>0.18</td>
<td>0.35</td>
<td>0.50</td>
<td>0.44</td>
</tr>
<tr>
<td>F1</td>
<td>0.26</td>
<td>0.70</td>
<td>0.15</td>
<td>0.52</td>
<td>0.20</td>
<td>0.46</td>
<td>0.33</td>
<td>0.37</td>
<td>0.36</td>
<td>0.00</td>
<td>0.75</td>
<td>0.74</td>
<td>0.55</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
<td>0.68</td>
<td>0.12</td>
<td>0.44</td>
<td>0.16</td>
<td>0.44</td>
<td>0.44</td>
<td>0.40</td>
<td>0.32</td>
<td>0.20</td>
<td>0.64</td>
<td>0.72</td>
<td>0.60</td>
</tr>
<tr>
<td>Recall</td>
<td>0.20</td>
<td>0.72</td>
<td>0.20</td>
<td>0.65</td>
<td>0.27</td>
<td>0.48</td>
<td>0.27</td>
<td>0.35</td>
<td>0.40</td>
<td>0.00</td>
<td>0.90</td>
<td>0.76</td>
<td>0.50</td>
</tr>
<tr>
<td>Fleiss' Kappa</td>
<td>0.56</td>
<td>0.53</td>
<td>0.72</td>
<td>0.35</td>
<td>0.54</td>
<td>0.53</td>
<td>0.59</td>
<td>0.47</td>
<td>0.33</td>
<td>0.64</td>
<td>0.54</td>
<td>0.41</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 4: Unsupervised metrics for extracted summaries with all sections.

<table>
<thead>
<tr>
<th>Metric \ Doc.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT Intro</td>
<td>.79</td>
<td>.85</td>
<td>.86</td>
<td>.84</td>
<td>.83</td>
<td>.80</td>
<td>.86</td>
<td>.85</td>
<td>.86</td>
<td>.81</td>
<td>.84</td>
<td>.82</td>
<td>.84</td>
</tr>
<tr>
<td>BERT All</td>
<td>.84</td>
<td>.86</td>
<td>.84</td>
<td>.89</td>
<td>.83</td>
<td>.85</td>
<td>.87</td>
<td>.84</td>
<td>.85</td>
<td>.83</td>
<td>.82</td>
<td>.87</td>
<td>.90</td>
</tr>
</tbody>
</table>

judgement is significantly higher at 0.84 with a Kappa score of 0.75 and a BERTScore on the higher end with 0.886. This case clearly shows that the ROUGE-L F1 score is not sufficient to assess all appropriate extractive summary candidates.

An introspection of the generated summaries in the first example document from all sections indicates the merit of the proposed unsupervised methods. We obtained sentences in the extractive summaries, such as “We highlighted three key factors...” and “The error analysis from NAB..”. We argue that even though the current methods are unsupervised and extractive ones – the selected candidates capture the essence of what we can expect from summarization methods. Notice that even though there may not be a high similarity in matching tokens between the output and ground truth – the output seems to be satisfactory. This also corroborates the fact that often datasets with only one correct ground truth may only paint a fraction of the whole picture. This may be of primal importance for the community for research that attempt to learn supervised models exploiting ground truth. How true is the ground truth? Is there a single version of truth or there is room for equally convincing alternatives?

Table 5 consists of the Rand Index Score (RI), Mutual Information Score (MI) and Completeness Score (CO) for Summaries Generated using Introduction part (Intro), All Sections part (OC) and from the Abstract sec-
Table 5

<table>
<thead>
<tr>
<th></th>
<th>M Vec. TFIDF</th>
<th>GloVe100D</th>
<th>Google News</th>
<th>W2Vec FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI Intro</td>
<td>0.1796</td>
<td>0.1779</td>
<td>0.1359</td>
<td>0.3506</td>
</tr>
<tr>
<td>RI OC</td>
<td>0.45833</td>
<td>0.3695</td>
<td>0.3480</td>
<td>0.2207</td>
</tr>
<tr>
<td>RI Abst</td>
<td>0.2857</td>
<td>0.4268</td>
<td>0.6752</td>
<td>0.1454</td>
</tr>
<tr>
<td>MI Intro</td>
<td>0.2205</td>
<td>0.1095</td>
<td>0.1887</td>
<td>0.3802</td>
</tr>
<tr>
<td>MI OC</td>
<td>0.3584</td>
<td>0.3141</td>
<td>0.5006</td>
<td>0.3281</td>
</tr>
<tr>
<td>MI Abst</td>
<td>0.3200</td>
<td>0.4613</td>
<td>0.6193</td>
<td>0.2321</td>
</tr>
<tr>
<td>CO Intro</td>
<td>0.5336</td>
<td>0.4918</td>
<td>0.5618</td>
<td>0.6410</td>
</tr>
<tr>
<td>CO OC</td>
<td>0.6246</td>
<td>0.6316</td>
<td>0.7116</td>
<td>0.5975</td>
</tr>
<tr>
<td>CO Abst</td>
<td>0.6335</td>
<td>0.7221</td>
<td>0.7993</td>
<td>0.5678</td>
</tr>
</tbody>
</table>

We performed the experiment with four vectorizers namely TF-IDF, Glove 100D, Google news model, and a variant of fine tuned Word2Vec model.

We performed the clustering experiment on this 13 papers. The motivation behind doing so was we have best and least ROUGE scores for these 13 files, now we wanted to see whether or not these files have some correlation in an unsupervised way. We did that was using the abstract section of the paper and all sections from the paper. We performed clustering on both abstract part and all sections separately. We use four different vectorizers and then do clustering on it. The idea behind using different vectorizers was to see if which gives the best Rand Index, Mutual Information and the Completeness score. To perform the qualitative analysis on the obtained clustering we did manual verification of the files that formed same cluster and checked whether the clustering really made any sense. Not all the clusters formed were coherent. Few exceptions were observed even for the clustering that gave best Rand Index score. We got the best Rand Index, Mutual Information and Completeness score using Google News Model and worst score on the fine tuned Word2Vec model.

6 Conclusion and Future Work

In this work, we discussed a supervised evaluation approach for extractive text summarization, depicting some weaknesses in reference summaries extracted from author highlights. The two regardated datasets show a bias towards sentences from the introduction section despite other viable candidates. Since summary evaluation is generally a highly subjective task, we have studied human evaluations and compared them to the ROUGE-L F1 score from the highlights. We propose to incorporate unsupervised metrics into the evaluation process, since they have shown to reflect more nuanced
human judgement in cases where the reference summaries from the highlights do not suffice. Those metrics include the BERTScore, Relative Ranking Comparison Score and the Relative Clustering Comparison Score. For future work, we intend to test further clustering methods and to build upon the so far promising BERTScore with fine-tuned language models.

References