International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6395
***Article Received**: 02 February 2023 **Revised**: 07 March 2023 **Accepted**: 14 March 2023*

Data Mining Oriented Automatic Scientific Documents Summarization

Dr BJD Kalyani[1], Dr Jaishri Wankhede[2], Shaik Shahanaz[3]
[1]Associate Professor
Department of Computer Science and Engineering Institute of Aeronautical Engineering, Dundigal, Hyderabad.
kjd_kalyani@yahoo.co.in.
[2]Associate Professor Department of Coputational Intelligence
Malla Reddy College of Engineering and Technology, Hyderabad. jaishri.pravin2009@gmail.com
[3]Assistant Professor
Department of Computer Science and Engineering Institute of Aeronautical Engineering, Dundigal, Hyderabad.
Shaikshahanaz.in@gmail.com

Abstract— The scientific research process usually begins with an examination of the advanced, which may include voluminous publications. Summarizing scientific articles can assist researchers in their research by speeding up the research process. The summary of scientific articles differs from the abstract text in general due to its specific structure and the inclusion of cited sentences. Most of the important information in scientific articles is presented in tables, statistics, and algorithm pseudocode. These features, however, rarely appear in the standard text. Therefore, a number of methods that consider the value of the structure of a scientific article have been suggested that improve the standard of the produced summary. This paper makes use of clustering algorithms to handle CL- SciSumm 2020 and longsumm 2020 tasks for summarization of scientific documents. There are three well-known clustering algorithms that are employed to tackle CL- SciSumm 2020 and LongSumm 2020 tasks, and several sentences recording functions, with textual deduction, are used to retrieved phrases from each cluster to generate summary.

Keywords- DB-Scan, K-means, Data Pre-processing, Error Analysis, Summarization.

I. Introduction

The polls [1] conducted across a variety of industries show that obtaining a description of major achievements in the field of science is essential; nonetheless, obtaining such polls necessitate substantial human effort. Scientific summary [2] tries to solve this issue by giving a brief depiction of key results and contributions in scientific articles, decreasing the time necessary to read the complete text to grasp key contributions. The collection of citation texts [3] from various articles may offer a description of the situation major concepts, techniques, as well as the mentioned work's contributions, forming a brief overview referred study. These community-based summaries [4] highlight the paper's major contributions, examine the topic from different perspectives, and represent the work's influence.

Generally, there are several issues regarding citation texts. Because they were authored by various writers, prejudiced toward one another. The citation texts deficiency context includes technique specifics, data, assumptions, and outcomes [5]. More crucially, the citing authors may misinterpret the original paper's views and claims; some contributions to the cited work may be attributed to it which aren't in line with the intention of the original author. Another important issue is the alteration of claim epistemic value is a term used to describe the worth of knowledge argues that many assertions by the creator of the work may be asserted as well as facts in subsequent citations. As a result, citation sentences ought to be connected to particular sections of the reference article that accurately represent them. This is referred to as "citation contextualization" [6]. Citation contextualization is a difficult process owing discrepancies in vocabulary between the languages of the citing and referred authors. Scientists in diverse domains have a huge difficulty in keeping themselves up-to-date with the ever- changing scientific landscape. The number of publications published will have doubled in nine years, according to a bibliometric analysis [7]. It is the goal of the summary of a scientific publication to offer an overview of the reference material. All of the relevant information should be included in this concise overview. As a result, reading and comprehending the content is made easier for the reader.

DOI: https://doi.org/10.17762/ijritcc.v11i4.6395
***Article Received**: 02 February 2023 **Revised**: 07 March 2023 **Accepted**: 14 March 2023*

II. RELATED WORK

Scientific document summarization's sixth common duty CL-SciSumm 2020 [9]. This issue has been addressed in two ways in the literature. It is possible to think of the abstract as a summary of the article; however this method just provides the work's main topic. The summary's key ideas may not be adequately conveyed in the abstract of Yasunaga et al., [10]. Therefore, citation-based summarizing has already had already been used to overcome the problem of scientific summary of Qazvinian et al., [11]. It makes use of a number of citations that point back to the root source of information, citations, brief summaries of the contributions made by the reference publication in question. There are many different ways to summarize a document, but text summarization is one of the more used methods (s). In text summarization, there are two main approaches: extractive of Li et al., [12] and abstractive Fergadis et al., [13], both of it use human like behavior to construct new words based on information taken from a document.

III. METHODOLOGY

The proposed work stresses the necessity of embedding sentences in scientific writing. More emphasis is devoted to generic domains than particular task domains in many works, whereas task-specific domains receive less attention, such as speech recognition. The page rank comparison [14] and the Maximal Marginal Relevance (MMR) [15] in the content selection module. At least not in terms of scientific document summary tasks that have been previously studied and compared to deep neural representation and experimental verification of efficacy of proposed model. This paper presents a new approach for citation contextualization based on word entrenched and domain specific information to openly account for terminology variances and paraphrases between the citing and the cited authors, when a distributional space is used to map a set of words to dense vectors, the aim is that comparable words are positioned near to each other [16].

A. MATHEMATICAL MODELING

The Language Model (LM) to extract information [17] with the help of word embeddings to account for terminology changes. Given a query as citation text q and a document of reference span d, the LM gives a probability p(d|q) to a reference span d. We arrive at the following results under the assumption of term independence and a uniform document prior.

$$p(d|q) \propto p(q|d) = \prod p(qi\ |d)\ n\ i{=}1 \qquad (1)$$

Where qi (i = 1, ..., n) are the terms in the query. The Dirichlet Smoothing in LM, p(qi |d) is assessed with a smoothed maximum likelihood estimate:

$$p(q_1|d) = \frac{f(q_i,d)+\mu p(q_i|C)}{\sum_{w\in V} f(w,d)+\mu} \qquad (2)$$

The frequency function f, p(qi |C) provides the background probability of word qi in the collection C, V is the whole vocabulary, and is the Dirichlet parameter. In proposed model, word embeddings are employed to supplement the prior formulation (Eq. 2). In particular, we solve for p to get theprobability p(qi | d).

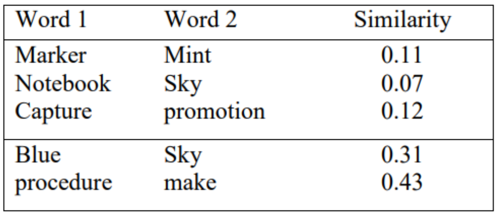
$$p(q_i|d) = \frac{\sum_{d_j\in d} s(q_i,d_j)+\mu(q_i|C)}{\sum_{w\in V}\sum_{d_j\in d} s(w,d_j)+\mu} \qquad (3)$$

Dj are words in the text d and s is an expression that expressesthe similarity between them.

$$s(q_i,d_j) = \begin{cases}\phi\left(e(a_i),e(d_j)\right), & if\ e(q_i).e(dj) > T\\ 0, & otherwise\end{cases} \qquad (4)$$

There is an embedding of word qi, a threshold, and a transformation function in the unit vector of the embedding e(qi). The dot product of the embeddings of the two terms qi and dj may thus be used to identify their similarities. Many unrelated words have dot products that are higher than zero, indicating syntactic and semantic relatedness [18], although this is not always the case. The noise they contribute is detrimental to the retrieval model's performance. The first step in tackling this problem is to set a threshold below which all similarity values are crushed to zero. This ensures that only phrases that are extremely relevant to the query are included in the retrieval model. To determine a reasonable value for, we compute the average and standard deviation of the point-wise absolute values of similarities between the pairs of terms in these samples using a random sample of words from the embedding model. To only take into account very high similarity values, we set to be two standard deviations bigger than the average similarities. When the terms have high similarity values, the results are not discriminative enough between linked and unrelated concepts. Table 1 The upper portion of the table shows pairs of random words, while the lowest portion shows similarity values for pairs of related words.

Table 1: Similarities Between Pairs of Words



Word 1	Word 2	Similarity
Marker	Mint	0.11
Notebook	Sky	0.07
Capture	promotion	0.12
Blue	Sky	0.31
procedure	make	0.43

Figure 1 shows that the most comparable phrases to the provided term aren't extremely discriminative. In other words, as one moves away from the most related terms, the similarity levels gradually decrease. To assess the impact of words that aren't connected to each other, we utilise a logit function (equation 5).

$$\emptyset(x) = log\left(\frac{x}{1-x}\right) \qquad (5)$$

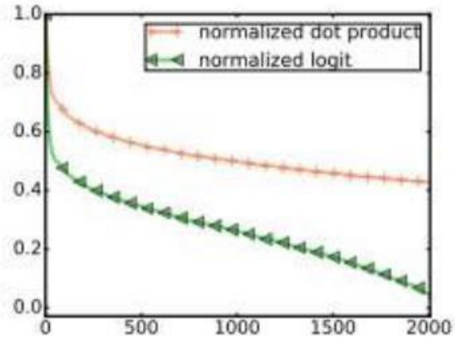


Figure 1: Normalized similarity Scores

As you can see from the graph, the x axis represents word indexes, while y represents similarity values. Original similarity values are shown in orange with + markers, whereas the logit transformed values are shown in green with triangle markers. In the logit function, fewer closely related terms are penalised. Expertise from a range of disciplines is included. To uncover the relationship between words, word embedding models are trained on a huge corpus. Word embeddings are subjected to a post-processing phase known as "retrofitting" [18] in this approach.

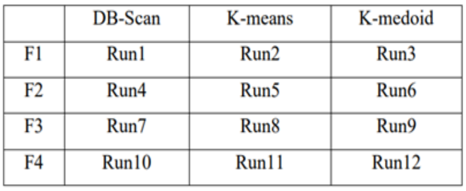
IV. IMPLEMENTATION AND RESULTS

This paper has presented three shared tasks, including CL-SciSumm 2020 [19], CL-LaySumm 2020 [20], and a third job called LongSumm 2020 [21]. On the NVIDIA T4 Tensor Core GPU, we fine-tune BERTBASE [22] on Masked Language Modeling and Sequence Classification tasks using the Transformer library2 for 2-6 epochs (with early stopping) before utilizing Adam to optimize our models. There are 40 publications in the 2018 training data set, but we only utilize 32 of them to fine-tune our model for sentence pair categorization.

A. *TASKS DESCRIPTION*

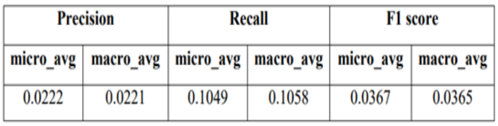
The first task is called "Task-1" (A) Using the reference paper (RP) and the citing papers as a guide, find all of the mentioned text spans in the paper (CPs). It is necessary to categorise each mentioned text span into preset aspects as part of Task-1 (B) (Hypothesis, Aim, Method, Results, and Implication). Use the citation in the reference document to write a brief description of it. The summaries created should not exceed 250 words in length. Table 2 provides a breakdown of the systems that were submitted. Although we utilized five clusters for K-means and K-medoid in this case, this is the default for DB scan. Table 3 and Table 4 lists the attributes that were utilized to choose the phrases inside clusters that would make up the summary for each run of Task-1 (a) and Task-1 (b) respectively. Task-2 has been broken down into twelve separate runs here, as in Table 5.

Table 2: CL-SciSumm Submissions



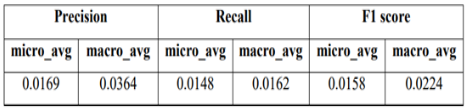
	DB-Scan	K-means	K-medoid
F1	Run1	Run2	Run3
F2	Run4	Run5	Run6
F3	Run7	Run8	Run9
F4	Run10	Run11	Run12

Table 3: Task-1 (A) Scores



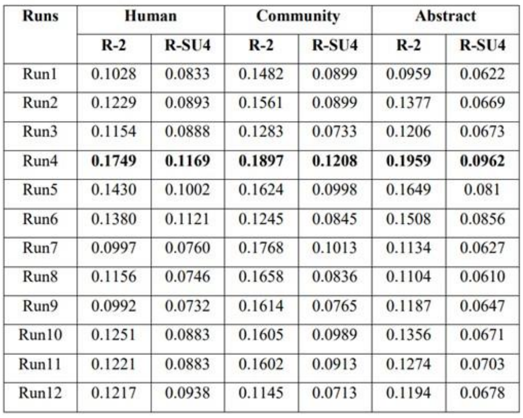
Precision		Recall		F1 score	
micro_avg	macro_avg	micro_avg	macro_avg	micro_avg	macro_avg
0.0222	0.0221	0.1049	0.1058	0.0367	0.0365

Table 4: Task-1 (B) Scores



Precision		Recall		F1 score	
micro_avg	macro_avg	micro_avg	macro_avg	micro_avg	macro_avg
0.0169	0.0364	0.0148	0.0162	0.0158	0.0224

Table 5: Task 2 scores



Runs	Human		Community		Abstract	
	R-2	R-SU4	R-2	R-SU4	R-2	R-SU4
Run1	0.1028	0.0833	0.1482	0.0899	0.0959	0.0622
Run2	0.1229	0.0893	0.1561	0.0899	0.1377	0.0669
Run3	0.1154	0.0888	0.1283	0.0733	0.1206	0.0673
Run4	**0.1749**	**0.1169**	**0.1897**	**0.1208**	**0.1959**	**0.0962**
Run5	0.1430	0.1002	0.1624	0.0998	0.1649	0.081
Run6	0.1380	0.1121	0.1245	0.0845	0.1508	0.0856
Run7	0.0997	0.0760	0.1768	0.1013	0.1134	0.0627
Run8	0.1156	0.0746	0.1658	0.0836	0.1104	0.0610
Run9	0.0992	0.0732	0.1614	0.0765	0.1187	0.0647
Run10	0.1251	0.0883	0.1605	0.0989	0.1356	0.0671
Run11	0.1221	0.0883	0.1602	0.0913	0.1274	0.0703
Run12	0.1217	0.0938	0.1145	0.0713	0.1194	0.0678

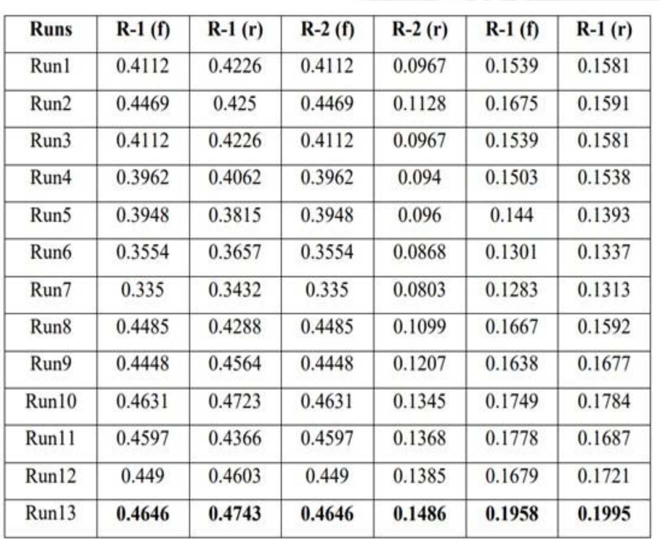
There is a 250-word limit on summaries in past scientific summarising attempts. However, the resulting summary for

***Article Received**: 02 February 2023 **Revised**: 07 March 2023 **Accepted**: 14 March 2023*

the current LongSumm shared task might be anything from 100 to 1500 words long. The training set for this dataset consists of 1705 articles linked to extractive summaries and 531 papers related to extractive summaries. It contains a collection of 22 files that are blindly checked and available at https://github.com/guyfe/LongSumm as in Figure 2, the scores are in Table 6. The lengthy summary scores for runs 1 through 12 were calculated using an extractive technique, whereas those for runs 13 were calculated using an abstractive approach.

Figure 2: The Task of Lay Summarization Architecture

Table 6: Longsumm run Scores



Runs	R-1 (f)	R-1 (r)	R-2 (f)	R-2 (r)	R-1 (f)	R-1 (r)
Run1	0.4112	0.4226	0.4112	0.0967	0.1539	0.1581
Run2	0.4469	0.425	0.4469	0.1128	0.1675	0.1591
Run3	0.4112	0.4226	0.4112	0.0967	0.1539	0.1581
Run4	0.3962	0.4062	0.3962	0.094	0.1503	0.1538
Run5	0.3948	0.3815	0.3948	0.096	0.144	0.1393
Run6	0.3554	0.3657	0.3554	0.0868	0.1301	0.1337
Run7	0.335	0.3432	0.335	0.0803	0.1283	0.1313
Run8	0.4485	0.4288	0.4485	0.1099	0.1667	0.1592
Run9	0.4448	0.4564	0.4448	0.1207	0.1638	0.1677
Run10	0.4631	0.4723	0.4631	0.1345	0.1749	0.1784
Run11	0.4597	0.4366	0.4597	0.1368	0.1778	0.1687
Run12	0.449	0.4603	0.449	0.1385	0.1679	0.1721
Run13	**0.4646**	**0.4743**	**0.4646**	**0.1486**	**0.1958**	**0.1995**

V. CONCLUSION

This paper has presented three shared tasks, including CL-SciSumm 2020, CL-LaySumm 2020, and a third job called LongSumm 2020. This work is broken down into three parts: Part 1(A), Part 1(B), and Part 2 for CL-SciSumm. While for Task 1 (A) we used WMD, for Task 1 (B) we used a similarity- based metric to determine the aspect of each quoted text span, respectively. The second task is based on clustering, which is followed by the extraction of relevant sentences from each cluster. Our research on LongSumm used clustering methods, and we discovered 13 alternative ways to produce a lengthy summary. LaySumm uses an encoder-decoder model based on deep learning to produce the lay summary using a well adjusted BERT language model's embedded representation.

REFERENCES

[1] Atanassova, I., Bertin, M., & Larivière, V. (2016). On the composition of scientific abstracts. Journal of Documentation.

[2] Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology, 66(11), 2215-2222.

[3] Cohan, A., & Goharian, N. (2017). Scientific article summarization using citation-context and article's discourse structure. ArXiv preprint arXiv: 1704.06619.

[4] Cohan, A., & Goharian, N. (2018). Scientific document summarization via citation contextualization and scientific discourse. International Journal on Digital Libraries, 19(2), 287-303.

[5] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[6] Mendoza, M., Bonilla, S., Nogoer, C., Cobos, C., & León, E. (2014). Extractive single-document summarization based on genetic operators and guided local search. Expert Systems withApplications, 41(9), 4158-4169s.

[7] Li, W., Xiao, X., Lyu, Y., & Wang, Y. (2018). Improving neural abstractive document summarization with explicit information selection modelling. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 1787-1796).

[8] Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-up abstractive summarization. ArXiv preprint arXiv: 1808.10792.

[9] Saini, N., Saha, S., Bhattacharyya, P., & Tuteja, H. (2020). Textual Entailment--Based Figure Summarization for Biomedical Articles. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1s), 1-24.

[10] Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., & Radev, D. R. (2019, July). Scisummnet: A large annotated corpus and contentimpact models for scientific paper summarization with citation networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7386-7393).

[11] Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., & Moon, T. (2013). Generating extractive summaries of scientific paradigms. Journal of Artificial Intelligence Research, 46, 165-201.

[12] Li, L., Xie, Y., Liu, W., Liu, Y., Jiang, Y., Qi, S., & Li, X. (2020, November). Cist@ cl-SciSumm 2020, longsumm 2020: Automatic scientific document summarization. In Proceedings of the First Workshop on Scholarly Document Processing (pp. 225-234).

[13] Fergadis, A., Pappas, D., & Papageorgiou, H. (2019). ATHENA@ CLSciSumm 2019: Siamese recurrent bi-directional neural network for identifying cited text spans. In BIRNDL@ SIGIR (pp. 256-262).

[14] Ou, S., & Kim, H. (2020, March). Ranking-Based Cited Text Identification with Highway Networks. In International Conference on Information (pp. 738-750). Springer, Cham.

[15] Parihar, A. S., Jain, A., & Gupta, A. (2020, June). Citation-Based Scientific Paper Summarization Using Game Theory. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp.

1157- 1161). IEEE.

[16] Zerva, C., Nghiem, M. Q., Nguyen, N. T., & Ananiadou, S. (2020). Cited text span identification for scientific summarisation using pre-trained encoders. Scientometrics, 125(3), 3109-3137.

[17] Altmami, N. I., & Menai, M. E. B. (2020). Automatic summarization of scientific articles: A survey. Journal of King Saud University-Computer and Information Sciences.

[18] Sotudeh, S., Cohan, A., & Goharian, N. (2020). On generating extended summaries of long documents. ArXiv preprint arXiv: 2012.14136.

[19] Vicente, M., & Lloret, E. (2020, October). A discourse-informed approach for cost-effective extractive summarization. In International Conference on Statistical Language and Speech Processing (pp. 109-121). Springer, Cham.

[20] Huang, R., & Krylova, K. (2020, November). Team MLU@ CL-SciSumm20: Methods for Computational Linguistics Scientific Citation Linkage. In Proceedings of the First Workshop on Scholarly Document Processing (pp. 282- 287).

[21] Mishra, S. K., Kundarapu, H., Saini, N., Saha, S., & Bhattacharyya, P. (2020, November). IITP-AI-NLP-ML@ CL- SciSumm 2020, CL-LaySumm 2020, LongSumm 2020. In Proceedings of the First Workshop on Scholarly Document Processing (pp. 270-276).

[22] Yu, T., Su, D., Dai, W., & Fung, P. (2020). Dimsum@ laysumm 20: Bart based approach for scientific document summarization. ArXiv preprint arXiv: 2010.09252.