Automatic Text Summarization using Document Clustering Named Entity Recognition

Senthamizh Selvan.R, Dr. K. Arutchelvan
Assistant Professor, Department of Computer and Information Science
Annamalai University, Chidambaram, Tamil Nadu, India

***Abstract*—Due to the rapid development of internet technology, social media and popular research article databases have generated many open text information. This large amount of textual information leads to 'Big Data'. Textual information can be recorded repeatedly about an event or topic on different websites. Text summarization (TS) is an emerging research field that helps to produce summary from a single or multiple documents. The redundant information in the documents is difficult, hence part or all of the sentences may be omitted without changing the gist of the document. TS can be organized as an exposition to collect accents from its special position, rather than being semantic in nature. Non-ASCII characters and pronunciation, including tokenizing and lemmatization are involved in generating a summary. This research work has proposed an Entity Aware Text Summarization using Document Clustering (EASDC) technique to extract summary from multi-documents. Named Entity Recognition (NER) has a vital part in the proposed work. The topics and key terms are identified using the NER technique. Extracted entities are ranked with Zipf's law and sentence clusters are formed using k-means clustering. Cosine similarity-based technique is used to eliminate the similar sentences from multi-documents and produce unique summary. The proposed EASDC technique is evaluated using CNN dataset and it shown an improvement of 1.6 percentage when compared with the baseline methods of Textrank and Lexrank.**

Keywords*—Named entity recognition; text summarization; k-means clustering; Zipf's law

I. INTRODUCTION

Internet technology paves the way for information resources. Excessive data is in text form and contains a lot of hidden information. Natural Language Processing (NLP) is one of the basic techniques used to extract hidden information from large amounts of textual data. Social media such as news channels and Facebook play an important role in generating high level text information. Reading huge amount of information is difficult task for human. Text summarization (TS) is the interesting research field that helps to generate summary from the text documents. Due to the data deluge and public consumption, information in the social media has redundant information in all normal dialects. For generating the summary, it is sometimes conceivable to delete words, expressions, rules and complete sentences without disrupting the significance of the message.

TS is the process of creating only brief outlines of the text without redundancies. Statistical analysis has its own limitations with the use of traditional ontological methods for deriving summary [1]. The content writers have different perspective and use their own writing styles. Hence, traditional ontology does not support to find the cognisance of the knowledge in the documents. This makes more complex for extracting the summary from the documents. Also, the text summary cycle may require an alternative overview of the information. Homogeneous depiction empowers a single representation to gather information from different form of resources using primary integration. This will help in summary productivity. In general, text summarization process is based on finite state automation [2].

These days, a huge number of research articles, news articles, blogs and forums are distributed in every field of study. It represents a great challengeable task for industry experts and researchers to know the latest developments in their specific fields. A new report reveals that many logical articles are copied at regular intervals [3]. The solution of logical essays overcomes this challenge by providing significant findings and commitments to the essay. In the scientific document summary generation, the summary generated by experts have help them to reduce the data collection (secondary research articles) work. It also reduces the work and time required to review any logical article. This was the essential inspiration to run this work.

In general, the text summarization can be done in two ways to get a summary of single or multiple documents. In the primary way, the paper's theory is considered an outline, although the problem with the theory is that it does not reveal the immeasurable significant commitments and findings of an article as a result of length limitations. The findings and confirmations made by the author of an article may be essential from the writer's point of view, but it may not be relevant to the local area. In addition, there is no data from all parts of the theoretical article [4]. The next method overcomes the weaknesses of the main system. Given a note sheet (sheet to be compressed), a note-based summary is generated using the notes in the note sheet. This approach has negative pages based on various authors composing reference texts, and any misconception by them may present mistakes in the last outline; As a result, facts, findings, and basic structure of the reference sheet may be missed. Reference Ecology [5] can take care of this problem. Here every sentence referring to the note-taking notes will be removed first. This arrangement of sentences is called reference systems. It can create the last summary using extracted sentences.

As it is discussed above, two methods are used to obtain the outline of the text: extractive and abstractive [6]. The Extractive Text Summarization (ETS) technique combines the

deletions obtained from the corpus to the frame outline. The Abstractive Text Summarization (ATS) technique creates new sentences from the data obtained from the corpus. The ETS technique is inapt for multiple document resolution. The cause is the likelihood of creating a one-page summary of a few sources [7]. Again, little effort has been made to summarize the evaluation records and differentiate between the components that affect the presentation of each method. Valuation logs contain ratings and preferences, e.g., websites or client surveys.

The goals of this research work are as follows:

- Proposing an entity ranking method to rank the important entities.
- To find the sentence similarity using Cosine similarity.
- Grouping the sentences of multi-document using k-means clustering.
- Generate the text summary using entity aware and document clustering.

The organization of this research article is as follows: the detailed overview of text summarization and its methods are discussed in this introduction Section I. The literature review of previous work on text summarization on various applications such as clinical summary generation, news articles summary generation and scientific article summary generation are discussed in the Section II. The proposed EADSC method is given with a big picture in the Section III. Results of the proposed work is discussed in the Section IV. Finally, this research article is concluded with the future work in section V.

II. Background Study

The introduction of the Transformer encoder-decoder models briefly sparked [8], [9] news [10] and logical articles [11] and significant improvements. By the way, their application for summary of medical notes has not been satisfactorily examined. A prototype in the light of Pointer-Generator-Networks [10], [12] has been proposed for a concise outline of radioactivity by combining materials in the medical specifications of UMLS [13] and RadLex [14]. They use inventions and impression pairs for abstract work, where inventions form communications and create objective definitions for creating records.

Sotudeh et al. [15] have proposed a two-level model that includes material selection and abstract summary for medical abstraction. Selector is ready to distinguish ontological terms from discoveries through medical metaphysics (Radlex) and create concise records. Two-LSDMs are used to encrypt inventions and LSDMs are used to create solutions following the LSDM-based decoder. Liang et. al [16] have developed a model for differentiating clinical symptoms in patients with diabetes and hypertension and developing obvious contractions of the disease. They examined a database of 3,453 medical records collected for 762 patients, outlining the difficulty in determining age as a punishment. The authors [17] have proposed a model that incorporated the syntax-based misdiagnosis and approval of the syntax medical idea for extracting the medical message. They conducted their experiments on the MIMIC-III [18] database.

Text report outline is the focus of much research in the research field of NLP. Different kind of methods have been used to solve this problem, such as passive semantic investigation [19], object visualization and poison models [20] and the meta-heuristic method [21]. The essence of live models is the control of information, because a lot of coded information is needed to solve a brief task [22]. In recent days, many researchers involved in developing graph-based models for generating summary from multi-documents. Sentences are considered as vertex or nodes, and the margins between the vertex indicate the similarity between the sentences. The key aspect of the graph-based approach is to determine the most focal sentence in the record. Part of the graph-based models are LexRank [23] and TextRank [24]. The logical solution from the bat made by Duffel et al is correct. [25].

The benefit of using the reference method to create the exterior of a research paper has been demonstrated by Elkis et.al. [26] and Hernandez et.al. [27]. Hong et.al. [28] have developed logical outline practices using various stabilized material wood. Cohen et al. [29] proposed a search-organized approach to separating significant sections of the reference sheet. Most experts are involved in reference contextualization to create a logical solution to the dissertation. The shared tasks in the TAC 2014 database, CL-SciSumm 2016 and the 2017 Logical Report solution [30] provided databases for local education for select reason. Although the CL-SciSumm 2016 and 2017 datasets compile the logical articles of the Computational Semantics section, the TAC database is linked to the Biomedical Article solution.

In a research work, the authors [31] have developed a group-based methods. The important notes are compiled first and more focused phrases are selected from different clusters to create the exterior. Li et al. [32] SVM classifiers are used to outline vocabulary and sentence proximity. The authors [33] have published a semi-ethnic model using similarities based on brain structure and tfidf, which scores relevant text that can be memorized for abstraction. The developers [34] proposed a revised TextRank calculation called TextSentenceRank to sort the sentences; here, a solution is designed in view of the stabilized sentences.

Baki et al. [35] Cosine similitude was used using the term frequency-inverse document frequency vector and a subgroup using SVM (Support vector Machine) [36]. The SVM is combined with Decision Tree (DT) to identify the reference length. The outline is created by removing the sentences considering the average notch in each quote. In another research work, the authors [37] have used TF as a vector space model that uses a characteristic approach to the representation of text and a non-negative structure. Lapalme G et al. [38], used similitude ability to identify reference text, including tips and very serious small fit to create the solution. Cao et al. [39] was used SVM model to rank the reference text for each reference, and the final solution was prepared by complex stabilization [40]. The authors [41] have worked on the tfidf comparison, jacquard similitude and proximity system were used to differentiate the reference length, which is further applied to the outline.

Chiruzzo et al. [42] have experimented the ACL Collection Note that uses a variety of embeds, such as Corpus Word Matching [43] and Google Newsword Installation. Jacquard used similarity [44], LDA [45] and more, with less emphasis on reference identity intimacy and outline age. Glavas et al. [46] the design of the reference summary used element-based features, level-based content, vector space likeness and unigram to obtain text-range. Dipankar et al. [47]. The cosine proximity was used to differentiate the reference length for the solution from the point of view of the scoring system.

The author proposed a work that, [48] reference length was extracted using a standard language model [49], printed content [50] and basic writing learning [51], which are additionally used to relate to the abstract design score. Cohen et al. [52] proposed a logical abstract technique. In another work, similar authors [5] proposed a technique for applying the reference context using question correction, word formation and direct learning. In another work, the researchers [53], have a separate reference structure using cosine likeness and jacquard comparison, and selects phrases in the light of different highlights from the note sheet to create a summary that is closely linked to the notes.

In another research work, authors have proposed a framework for logical resolution [54] to recognize the reference system using the mover's distance and the LDA model; also, they used a set age process (DPP) for the abbreviated age. The authors have proposed a framework [55] and used weighted democratic classifiers to extract the citation system. Clustering method was used to generate the summary. The researchers [56] have used the open NMT an application for the TS. Nghiem et al. [57] an adjusted two-way transformer was used to differentiate the reference system. Furthermore, they have proposed a semi-ethnic outline technique for the compression age. The table I represents the scope and drawbacks of the different automatic text summarization approaches.

TABLE I. SCOPE AND RESTRICTIONS OF THE CONVENTIONAL ATS APPROACHES

Reference	Year	Description	Limitations
[58]	2009	important ways to summarise the content and a taxonomy of summarising techniques	Missing from NLP is extractive, abstract, machine learning, and deep learning.
[59]	2014	Reviewed works from 2000 to 2013 and suggested a statistical approach.	excludes cognitive components
[60]	2014	A hybrid strategy can effectively use both extractive and abstractive Techniques	avoided difficult procedures
[61]	2016	Introduces the concepts of abstractive and extractive summarization.	only describe strategies and procedures
[62]	2017	Extractive approaches for summarising multilingual texts are presented.	There is a lack of a definite classification and concept of feature score.
[63]	2021	to manage several materials for comparison and summary based on recent research work	does not constitute a quick conversation

Li et al. [64] have distinguished between the Word 2 reference system for the CNN model and the determining point processes (DPP) for the text summarization. Cagliero et al. [65] have used a writing model for the reference system for individual sources. The short path amongst the selected text length is predicted and the summary is generated. The authors [66] have used a variety of classification and ballot systems to identify the reference system. To format the summary, they compile phrases and select high-quality phrases from each encounter. Researchers [67] have diagnosis of the use of intermediate brain systems and how to monitor the reference system; for a long time, abstraction was designed by selecting phrases such as notes.

III. MATERIALS AND METHODS

Extracting a summary from multi-documents are created by cutting key pieces of text from the collected documents. Statistical analysis is involved to find the important of sentences. The overall process involves in the proposed EASDC technique is given in the Fig. 1.

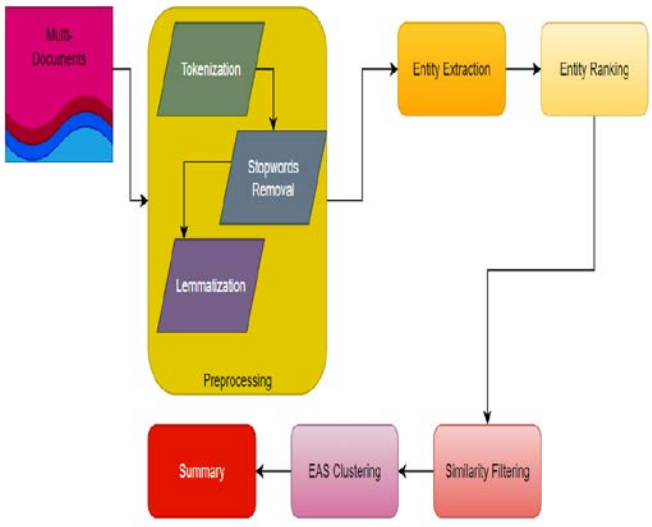

Fig. 1. Research Process of Proposed EASDC Work.

The proposed EASDC technique works using the following components: pre-processing, entity extraction, entity ranking, similarity filtering and EAS clustering. These components are well explained in the following sub-sections.

A. Pre-processing

In the text mining, preprocessing plays a major role to provide the documents in structured representation. Multi-document is defined by the following equation (1).

$$\begin{Bmatrix} \{d_1, d_2, d_3, \dots, d_n\} \in MD \\ if\ d > 1, MD\ is\ true \\ if\ d > 0\ and\ d \leq 1, MD\ is\ false \end{Bmatrix} \quad (1)$$

From the equation 1, d resembles document and MD resembles collection of documents. If d is lesser and equal to one, then it is single document. In English language, each sentence is identified by the full stop (.) at the end of a word. In preprocessing step, each sentence is tokenized using the regular

expression. The sentence tokenization is given in the equation (2), if number of sentences is lesser than five then the summarization process is not needed.

$$\begin{Bmatrix} \{sen_1, sen_2, sen_3, \dots, sen_n\} \in S \\ if\ sen\ \geq 5, S\ is\ true \\ else, S\ is\ false \end{Bmatrix} \quad (2)$$

The preprocessing phase involves the following technical process:

1) Tokenization: Each word is classified by a token in the sentence. As it is given in the equation 2, each sentence is classified by a token in a document. The sentences are tokenized and store into a desired format. In this proposed work, the tokenized sentences are stored in the array format and resembled as a sequence of tokenized-sentences. Removing non-ASCII characters are essential before proceeding the tokenization process. The non-ASCII characters are meaningless and it is not necessary for the text mining process.

2) Removal of stopwords: After tokenization, stopwords are important to remove the useless word in the sentences. For generating the summary, the subjective words are sufficient to find the important sentences. Hence the useless words such as the, of, a, an, in can be removed and generate new tokens.

B. Entity Extraction

Entities are the key terms that helps to identify the important sentences. The word ambiguity is the important challenge in the entity extraction. Named Entity Recognition (NER) is the important task in NLP to extract the important keywords. For example, consider the following sentence, Chennai super kings won the T20 match that was held in Chennai. In the given sentence, Chennai is location and Chennai super kings is an organization. The output of the entity extraction must be as follows: Chennai super kings (Organization) won the T20 match that was held in Chennai (location). The ambiguous words can be identified by the human easily, whereas the human generated computational application needs lot of training.

For this research work, the fastest Spacy 3.0 library is used to extract the entities. The extracted entities are converted into numerical using token-2-vector model that is present in the spacy library. Conditional Random Field (CRF) technique is used to tag the ambiguous word with its identified entities. The CRF can be calculated using the following equation (3).

$$p(y|X) = \frac{1}{Z(X)} \prod_{n=1}^{N} exp\{\sum_{m=1}^{M} \delta_m f_m(y_n, y_{n-1}, X_n)\} \quad (3)$$

where y is the part-of-speech of the current token and X is the observed entity. The weight of the words are resembles using δ_m f_m and y_n,y_(n-1),X_n resembles the features of the sentences. Z(X) is the total quantity of the named entities (NE). The weight estimation of the NE token is calculated using the maximum-likelihood estimation.

C. Entity Ranking

NE ranking is used to identify the importance of a NE present in the documents. In general, the frequency of NE across the document gives the importance of a NE (i.e., number of appearances of NE in the document). The actual frequency of word has to be calculated using the inverse proportional of entity in the whole document. Zipf's law is one of the popular methods to identify the rank of the word in the given document. Zipf's law states that, the rank frequency of word is inversely proportional to the rank in the frequency table. The Zipf's law is defined in the equation (4) given below,

$$Z(r, \beta) \propto \frac{1}{r^{\beta}} \quad (4)$$

where β≈1, r denotes the rank of word and Z(r,β) represents the frequency of entity in documents. The identified entities are ranked and organized using the above equation 4.

D. Similarity Filtering

Cosine similarity is calculated with the following equation (5) given below. It is used to find similarity in the sentences. From the equation (4), if the rank value is lesser than 0 is not considered for the document summary and it is eliminated from the tokenized entities. For the proposed EASDC work, the threshold value for the similarity index is set to greater than 75 percentage to get perfect similarity [68].

In the equation (5), sentence similarity is evaluated by comparing a sentence with all other sentences. The similarity matrix is created based on the values. The highly identical sentences based on the cosine value are removed from the documents. This reduced the redundant information from the multiple documents.

$$cos(s_i, s_j) = \frac{\sum_{k=1}^{n} s_i * s_j}{\sqrt{\sum_{k=1}^{n} s_i^2} X \sqrt{\sum_{k=1}^{n} s_j^2}} \quad (5)$$

where S is sentences and si and sj denotes the current and next sentence respectively.

E. EAS Clustering

Sentences in multiple documents are ranked based on the presence of entities. The entities are ranked by using the equation 4. The duplicate sentence is eliminated using the equation 5. The unique sentences are embedded using the Doc2Vec mechanism. It is the extension of word embedding. For sentence embedding, this research work uses the Distributed Memory version of Paragraph Vector (PVDM) technique. The embedded sentence is represented with a unique token and stored in a matrix format.

The document clustering is one of the popular methods that is used to group the documents. In the proposed EASDC research work, it used k-means clustering mechanism to grouping the sentences based on the cosine similarity. Cosine distance method is applied to find the distance between the similar sentences. The equation for the cosine distance is given in the equation 6.

$$cosine_{distance} = \frac{\sum_{n=0}^{N} s_n - p_x}{\sum_{n=0}^{N} (s_n)^2 * \sum_{n=0}^{N} (p_x)^2} \quad (6)$$

where sn is the sentence with the position n, where n is 1, 2, 3..., N. N is the total number of sentences. Px is the next sentence with point of observed x. The pseudocode of k-means-clustering pseudocode is given below.

Pseudocode 1: k-means clustering

Input: embedded sentences
Output: grouped sentences

```
Set number of k
Set centroids p1, p2, …px randomly
Repeat steps 4 and 5 till the end of iterations
for n in px:
find the nearest centroid
assign the point to that cluster
for j in cluster_k:
find new centroid by calculating the mean of centroids
End
```

The clusters are pre-defined based on the length of summary that has to be created. The vector of the sentences is used to find the distance between the sentences. The number of clusters is set based on the entities and the chosen topic.

The clustered sentences may have different topics, because, the cluster is formed based on the entities and topics that has discussed in the documents. Therefore, each cluster may have a higher probability of being different topics. The top sentence in each cluster is considered and helps to form the summary. Each sentence from a cluster is sufficient to exaggerate the important of the topic. If there is one cluster, then the sentences in the cluster is extracted to form the summary. In another situation, if similarity is found between the topics, the top sentence of each cluster is taken and the similarity of the sentences is calculated. The redundant sentence from those clusters is eliminated and summary can be generated. The summary generation is given in the pseudocode below.

Pseudocode 2: Summary Generator

Input: clustered sentences
Output: Summary

```
Load the clustered sentences as C
Set T_C = [ ] //list of topics, i.e., cluster head.
Set summary = [ ]
For j in C:
If [T_(C_(j ) ) and 〖 T〗 _(C_(j+1) )]>0.75 : # T_(C_(j ) ) denotes, jth sentence of a cluster C under topic T
For s in j:
Summary.append(s)
Break
If [T_(C_(j ) ) and 〖 T〗 _(C_(j+1) )]<0.75 :
Eliminate the redundant sentence T_(C_(j ) )
End
```

IV. Results and Discussions

The proposed EASDC method generates summary using document clustering and named entity recognition. The identified entities are ranked and create the cluster using the equation 5 and 6. The clustered sentences are involved in generating the summary. The summary of the multi-documents is restricted based on the required length. The sentence similarity plays major role in eliminating the redundant sentences. The sentences are embedded and it helps to diminish computational sparsity.

To evaluate the recital of the proposed EASDC research work CNN dataset is used. The entity extraction played a major role in the proposed research process. It turned out to be more difficult than we had anticipated. The DUC dataset looks to have a perfectly functioning XML structure, but we were unable to load it using numerous Python modules because of errors in the XML format. After that, we had to separate the lengthy text dumps into sentences. Our initial, basic implementation simply divided the text into paragraphs using periods. To speed up the loading of data, this data was saved to disc. This process can take over an hour to perform from scratch, but once finished, it doesn't need to be repeated.

The entities are extracted using the python spacy library. The named entities such as person, organization, location, GPE (geographical place), date, time and money. These entities are ranked based on their occurrences using Zipf's laws. The document-clustering using k-means cluster helps to grouping the sentences and generate the summary.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used to measure the efficient of proposed summarization technique, which is given in the equation 7.

$$ROUGE - N(c,r) = \frac{\sum_{r_i \in r} \sum_{n-gram \in r_i} Count(n-gram,c)}{\sum_{r_i \in r} numNgrams(r_i)} \quad (7)$$

Where N is the grams or tokens or words, c and r is candidate and reference respectively.

In this article, the following metrics are tested to evaluate the efficiency of the proposed EASDC technique:

- ROUGE - 1
- ROUGE - 2
- ROUGE - L

Table II denotes the outcome of comparison between EASDC with TextRank and LexRank methods. The proposed EASDC method outperformed well.

The comparative outcomes of the tested strategies for graphical representation are shown in Fig. 2. The LexRank algorithm yielded an average of 38.3%, compared to 39.06% for the TextRank algorithm. The proposed EASDC method performed better and generated 40.73%. It demonstrated a 1.67 percent improvement over the TextRank algorithm. The ROUGE scores improved with entity extraction-based extractive summarization.

TABLE II. Scope and Restrictions of the Conventional ATS Approaches

	LexRank	**TextRank**	**EASDC**
Rouge - 1	36.6	37.6	39.2
Rouge - 2	37.4	38.4	39.9
Rouge - L	40.9	41.2	43.1

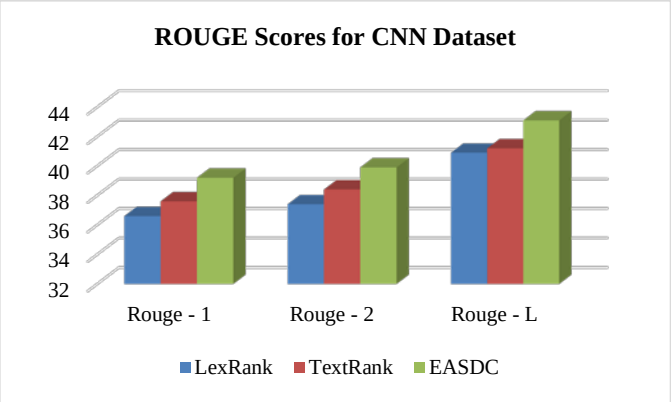

Fig. 2. ROUGE Scores of Proposed EASDC Technique and Existing Works.

V. Conclusion

The size-sensitive expansion of the World Wide Web has created better access to textual information. This work presents a technique for generating literary information that solves the problem of repetition and error in one of these ways. Implementing the proposed framework is not significantly different from business text summaries. The entity ranking and sentence similarity calculation helps to extract the unique sentences from the multiple documents. The extracted NE are then passed to the document clustering methods. By estimation, k-implies are a group calculation and high-level cluster calculations are used for incomparable effects. Similarly, the tendency to extract the sentence from each group is not based on random correlation rather to develop a particular calculation. The proposed EASDC technique shown an improvement of 1.67 percentage and 2.3 percentage compared to TextRank and LexRank algorithm respectively.

In order to further enhance the summary quality in the context of multidocument summarizing, we would like to investigate more strategies in the future, such as methods based on reinforcement learning. We also want to use our approach for additional tasks, such answering multidocument questions.

References

[1] G. Eason, Endres-Niggemeyer, Brigitte, Summarizing Information: Including CD-ROM 'SimSum', Simulation of Summarizing, for Macintosh and Windows. Springer Science & Business Media, 2012.

[2] Salton, Gerard, et al. 'Automatic analysis, theme generation, and summarization of machine-readable texts', Information retrieval and hypertext. Springer US, 1996. 51-73.

[3] 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.

[4] Atanassova I, Bertin M, Larivi`ere V (2016) On the composition of scientific abstracts. Journal of Documentation.

[5] Cohan A, Goharian N (2018) Scientific document summarization via citation contextualization and scientific discourse. Int J Digit Libr 19(2-3):287.

[6] Hahn, Udo, Inderjeet Mani, 'The challenges of automatic summarization', Computer 33.11, 2000, page: 29-36.

[7] Barzilay, Regina, Michael Elhadad, 'Using lexical chains for text summarization', Advances in automatic text summarization, 1999, page: 111-121.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint arXiv:1910.10683, 2019.

[10] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," arXiv preprint arXiv:1704.04368, 2017.

[11] I. Cachola, K. Lo, A. Cohan, and D. S. Weld, "Tldr: Extreme summarization of scientific documents," arXiv preprint arXiv:2004.15011, 2020.

[12] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, "Ontology-aware clinical abstractive summarization," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1013–1016.

[13] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," Nucleic acids research, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[14] C. P. Langlotz, "Radlex: a new method for indexing online educational materials," pp. 1595–1597, 2006.

[15] S. Sotudeh, N. Goharian, and R. W. Filice, "Attend to medical ontologies: Content selection for clinical abstractive summarization," arXiv preprint arXiv:2005.00163, 2020.

[16] J. Liang, C.-H. Tsou, and A. Poddar, "A novel system for extractive clinical note summarization using ehr data," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 46–54.

[17] W.-H. Weng, Y.-A. Chung, and S. Tong, "Clinical text summarization with syntax-based negation and semantic concept identification," arXiv preprint arXiv:2003.00353, 2020.

[18] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," Scientific data, vol. 3, no. 1, pp. 1–9, 2016.

[19] Gong Y, Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 19–25.

[20] Ma T, Nakagawa H (2013) Automatically Determining a Proper Length for Multi-document Summarization: A Bayesian Nonparametric Approach. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 736–746.

[21] Saini N, Saha S, Chakraborty D, Bhattacharyya P (2019) Extractive single document summarization using binary differential evolution: optimization of different sentence quality measures. PloS One 14(11).

[22] Louis A, Joshi A, Nenkova A (2010) Discourse Indicators for Content Selection in Summaization. In: Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue. Association for Computational Linguistics, pp 147–156.

[23] Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22:457.

[24] Mihalcea R, Tarau P (2004) Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 404–411.

[25] Teufel S, Moens M (2002) Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics 28(4):409.

[26] Elkiss A, Shen S, Fader A, Erkan G, States D, Radev D (2008) Blind men and elephants: what do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology 59(1):51.

[27] Hern´andez-Alvarez M, Gomez JM (2016) Survey about citation context analysis: tasks, techniques, and resources. Nat Lang Eng 22(3):327.

[28] Hoang CDV, Kan MY (2010) Towards automated related work summarization. In: Proceedings of the 23rd international conference on

computational linguistics: posters. Association for Computational Linguistics, pp 427–435.

[29] Cohan A, Soldaini L, Goharian N (2015) Matching citation text and cited spans in biomedical literature: a search-oriented approach. In: Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: human language technologies, pp 1042–1048.

[30] Jaidka K, Chandrasekaran MK, Rustagi S, Kan MY (2016) Overview of the CL-SciSumm 2016 shared task. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 93–102.

[31] Qazvinian V, Radev DR (2008) Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd international conference on computational linguistics, vol 1. Association for Computational Linguistics, pp 689–696.

[32] Li L, Mao L, Zhang Y, Chi J, Huang T, Cong X, Peng H (2016) Cist system for cl-scisumm 2016 shared task. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 156–167.

[33] Nomoto T (2016) NEAL: A neurally enhanced approach to linking citation and reference. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 168–174.

[34] Klampfl S, Rexha A, Kern R (2016) Identifying Referenced Text in Scientific Publications by Summarisation and Classification Techniques. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 122–131.

[35] Moraes L, Baki S, Verma R, Lee D (2016) Identifying referenced text in scientific publications by summarisation and classification techniques. In: Proceedings of the joint workshop on bibliometric enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 113–121.

[36] Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20(3):273.

[37] Conroy J, Davis S (2015) Vector space models for scientific document summarization. In: Proceedings of the 1st workshop on vector space modeling for natural language processing, pp 186–191.

[38] Malenfant B, Lapalme G (2016) RALI system description for CL-SciSumm 2016 shared task. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 146–155.

[39] Cao Z, Li W, Wu D (2016) Polyu at cl-scisumm 2016. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 132–138.

[40] Wan X, Yang J, Xiao J (2007) Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In: IJCAI, vol 7, pp 2903–2908.

[41] Li L, Zhang Y, Mao L, Chi J, Chen M, Huang Z (2017) Cist@ clscisumm-17: multiple features based citation linkage classification and summarization.

[42] Abura'ed A, Chiruzzo L, Saggion H, Accuosto P, Bravo Serrano `A (2017) Lastus/taln@ Clscisumm-17: cross-document sentence matching and scientific text summarization systems.

[43] Bird S, Dale R, Dorr BJ, Gibson B, Joseph MT, Kan MY, Lee D, Powley B, Radev DR, Tan YF (2008) The acl anthology reference corpus: a reference dataset for bibliographic research in computational linguistics.

[44] Jaccard P (1901) ´ETude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat 37:547.

[45] Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993.

[46] Lauscher A, Glavas G, Eckert K (2017) University of Mannheim@ CLSciSumm-17: Citation-based summarization of scientific articles using semantic textual similarity. CEUR workshop proceedings 2002:33–42. RWTH.

[47] Dipankar Das S, Pramanick A (2017) In: Proc. of the 2nd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL2017). Tokyo, Japan (August 2017).

[48] Karimi S, Moraes L, Das A, Shakery A, Verma R (2018) Citance based retrieval and summarization using ir and machine learning. Scientometrics 116(2):1331.

[49] Lv Y, Zhai C (2009) Positional language models for information retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, pp 299–306.

[50] Tian R, Miyao Y, Matsuzaki T (2014) Logical inference on dependency-based compositional semantics. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long Papers), pp 79–89.

[51] Blitzer J, McDonald R, Pereira F (2006) Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, Sydney, pp 120–128. https://www.aclweb.org/anthology/W06-1615.

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

[53] AbuRa'ed A, Bravo Serrano A`, Chiruzzo L, Saggion H (2019) LaSTUS-TALN+ INCO@ CL-SciSumm 2019. BIRNDL@ SIGIR, 224–232.

[54] Ma S, Xu J, Wang J, Zhang C (2017) NJUST @ CLSciSumm-17. BIRNDL@SIGIR.

[55] Ma S, Zhang H, Xu J, Zhang C (2018) NJUST@CLSciSumm-18. BIRNDL@SIGIR.

[56] Debnath D, Achom A, Pakray P (2018) NLP-NITMZ@ CLScisumm-18. BIRNDL@ SIGIR. pp 164–171.

[57] Zerva C, Nghiem MQ, Nguyen NT, Ananiadou S (2020) Cited text span identification for scientific summarisation using pre-trained encoders. Scientometrics 125(3):3109–3137.

[58] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, ''A comprehensive survey on text summarization systems,'' in Proc. 2nd Int. Conf.Comput. Sci. Appl., Dec. 2009, pp. 1–6.

[59] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. D. Fiol, ''Text summarization in the biomedical domain: A systematic review of recent research,'' J. Biomed. Informat., vol. 52, Dec. 2014,pp. 457–467.

[60] C. Saranyamol and L. Sindhu, ''A survey on automatic text summarization,'' Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 6,2014, pp. 7889–7893.

[61] N. Andhale and L. A. Bewoor, ''An overview of text summarization techniques,'' in Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA), Aug. 2016, pp. 1–7.

[62] M. Gambhir and V. Gupta, ''Recent automatic text summarization techniques: A survey,'' Artif. Intell. Rev., vol. 47, no. 1, pp. 1–66, Jan. 2017.

[63] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, ''Automatic text summarization: A comprehensive survey,'' Expert Syst. Appl., vol. 165, Mar. 2021, Art. no. 113679.

[64] Li L, Zhu Y, Xie Y, Huang Z, Liu W, Li X, Liu Y (2019) CIST@ CLSciSumm-19: Automatic Scientific Paper Summarization with Citances and Facets. In: BIRNDL@ SIGIR, pp 196–207.

[65] La Quatra M, Cagliero L, Baralis E (2019) Poli2Sum@ CLSciSumm-19: Identify, Classify, and Summarize Cited Text Spans by means of Ensembles of Supervised Models. In: BIRNDL@ SIGIR, pp 233–246.

[66] Ma S, Zhang H, Xu T, Xu J, Hu S, Zhang C (2019) IR&TMNJUST @ CLSciSumm-19. In: BIRNDL@ SIGIR, pp 181–195.

[67] Chiruzzo L, AbuRa'ed A, Bravo A`, Saggion H (2019) LaSTUSTALN+INCO@ CL-SciSumm 2019. BIRNDL@ SIGIR, pp 224–232.

[68] R. Senthamizh Selvan, Dr. K. Arutchelvan. (2021). An Effective Approach for Abstractive Text Summarization using Semantic Graph Model. Annals of the Romanian Society for Cell Biology, 13925–1393.