ExtraSum @ MultiClinSum: Extractive Summarization of English, Spanish, French and Portuguese Clinical Case Reports

Soukaina Rhazzafe[1,*], Simon Colreavy-Donnelly[1] and Nikola S. Nikolov[1]

[1]Department of Computer Science and Information Systems, University of Limerick, V94 T9PX Limerick, Ireland

Abstract

This paper presents an extractive summarization approach for multilingual clinical case reports submitted to the MultiClinSUM 2025 shared task. We focused on selecting the ten most important sentences from each report while preserving the original text to ensure factual consistency. Our method compares four extractive techniques: graph based, concept based, topic based and clustering based summarization, tested on English, Spanish, French and Portuguese. Our experiments show that the clustering based summarization using multilingual BERT consistently outperforms the other methods in all languages, with the strongest semantic similarity seen in English. This suggests that multilingual BERT embeddings are effective at capturing the central meaning of clinical texts across different languages.

Keywords

Extractive summarization, Clinical case reports, Clinical text summarization, Multilingual text, Sentence selection

1. Introduction

Clinical case reports are a valuable source of detailed medical knowledge, as they provide insights into the diagnosis, treatment and outcomes of individual patients. They typically include critical information such as patient demographics, clinical presentation, diagnostic process, treatment and follow-ups [1, 2]. These reports not only support clinical education and research but also form a foundation for developing automated tools for clinical language processing. However, given their length and complexity, summarizing these documents effectively and accurately remains a significant challenge, especially when semantic plausibility, relevance and factual accuracy are required for clinical utility [3, 4].

The MultiClinSUM 2025 shared task [2, 5] addresses this issue by providing resources and benchmarks for multilingual summarization of clinical documents, focusing on four languages: English, Spanish, French and Portuguese. As part of the BioASQ BioASQ Workshop [6], participants are invited to generate summaries of clinical case reports using any approach, with evaluations conducted independently per language. The task highlights the complexity of generating high quality summaries in this domain, where even human-generated summaries can vary significantly in content and clarity. This complexity is further compounded by the lack of clear summarization objectives in the literature [7], making the design and evaluation of summarization systems especially difficult.

Clinical case reports resemble discharge summaries in structure and content, making them particularly suitable for developing summarization models using publicly available data [2, 5]. However, many prior works seem to fail to clearly define their summarization objectives, leading to difficulties in evaluating the clinical relevance of the outputs [7]. Some studies describe their goals in vague terms such as "significant impressions" or "critical diagnoses," with mismatches between intended audiences and generated outputs. To address these gaps, there is a growing need for models that produce factually consistent, targeted and clinically relevant summaries [7].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.
soukaina.rhazzafe@ul.ie (S. Rhazzafe); simon.colreavy@ul.ie (S. Colreavy-Donnelly); nikola.nikolov@ul.ie (N. S. Nikolov)
https://github.com/soukainaRhazzafe/ (S. Rhazzafe)
0009-0006-5846-4897 (S. Rhazzafe); 0000-0002-1795-6995 (S. Colreavy-Donnelly); 0000-0001-8022-0297 (N. S. Nikolov)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Extractive summarization has traditionally been favored over abstractive methods in the clinical domain because it preserves the original phrasing and reduces the risk of hallucination or factual errors [8]. In our previous work [9], we proposed a hybrid summarization approach for Electronic Health Records (EHRs) that combined extractive and abstractive techniques to summarize ICU progress notes and predict patients' length of stay. Our concept based extractive strategy, in combination with a T5 model, showed promising results in capturing clinically relevant information while maintaining coherence.

Building on this foundation, in the current study we focus exclusively on extractive summarization techniques for the MultiClinSUM shared task. Our approach aims to identify the 10 most important sentences from each clinical case report in a way that retains factual integrity and relevance. We define importance based on how central a sentence is in the overall text and whether it covers key clinical concepts. To achieve this, we implement and compare four extractive methods: a graph based method based on semantic similarity and sentences raking using PageRank, a concept based method using QuickUMLS for medical term extraction and ranking, a topic based method using TF-IDF to identify topic-salient sentences and, finally, a clustering based method using multilingual BERT embeddings to extract centroid representative sentences.

By choosing purely extractive methods, we mitigate the risks of generating factually incorrect or misleading summaries, an especially critical concern in the clinical domain. Furthermore, our decision to retain sentences without altering their original text aligns with the need for high transparency, interpretability and clinical fidelity. This approach also bypasses the common limitations in previous work, such as unclear objectives and audience misalignment [7].

The rest of the paper is organized as follows: Section 2 describes our summarization pipeline and techniques, Section 3 presents evaluation results, Section 4 discusses the results further and finally Section 5 concludes with key findings and directions for future research.

2. Methodology

2.1. Extractive Summarization Techniques

Extractive summarization is a technique that generates summaries by selecting and concatenating the most important sentences from the input document. The number of selected sentences is typically limited by a compression rate, a length cutoff or a predefined threshold [10].

As shown in Figure 1, the extractive summarization pipeline begins with pre-processing, which depends on the task and may involve several text cleaning techniques [10, 11]. Common pre-processing steps include sentence boundary detection, typically based on punctuation like periods, stop word removal, eliminating frequent but non-informative words and stemming reducing words to their root form to emphasize meaning [12]. Another method includes replacing specific values or terms with placeholders to improve sentence comparability.

The processing step involves selecting key sentences from the input text. First, a representation of the text is created. Then, scores are assigned to each sentence based on features derived from that representation. Finally, the top ranked sentences are selected according to the summary size constraint and concatenated to form the summary [10]. This step, including the text representation models, is introduced in detail in the remainder of this section.

In the post-processing step, the selected sentences may be reordered to match their original sequence in the input text. Placeholders introduced during pre-processing are also replaced with the original content [10, 12].

There are various extractive summarization techniques available [10]. These methods differ in how they represent the input text and in how they score and rank sentences for selection. To address the MultiClinSum challenge, we implemented four such techniques. Some of these methods were adapted from our previous work [9], where they were applied in a different clinical summarization context.

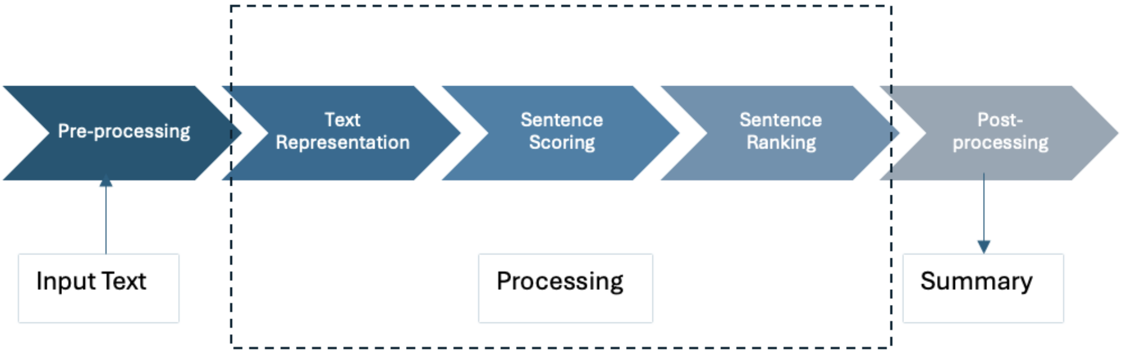

Figure 1: Extractive Summarization Pipeline [10].

2.1.1. Graph Based

Graph based extractive summarization techniques represent sentences as nodes in a graph, where edges between nodes capture the similarity between pairs of sentences [10]. This method measures sentence-to-sentence similarity to identify central content. The input text is first converted into a numerical representation, which is used to calculate sentence similarities and build the weighted graph. A ranking algorithm is then applied to the graph to identify the most important sentences for the summary.

To implement this technique, we represented sentences using Term Frequency-Inverse Document Frequency (TF-IDF) vectors [13], which assign weights to words based on their frequency and rarity in the text. The TF-IDF formula is shown in Equation (1), where TF_{ij} is the Term Frequency of the i-th word in the j-th document and IDF_i is the Inverse Document Frequency of the i-th word. $TF\text{-}IDF_{ij}$ is the TF-IDF value of the i-th word in the j-th document; n_{ij} is the number of occurrences of the i-th word in the j-th document, while N_i is the number of documents containing the i-th word, and D is the total number of documents.

$$TF\text{-}IDF_{ij} = TF_{ij} \cdot IDF_i = \frac{n_{ij}}{\sum_k n_{kj}} \log\left(\frac{D}{N_i}\right) \tag{1}$$

Pairwise cosine similarity between the TF-IDF vectors was then calculated to create a similarity matrix, forming the weighted edges of the graph. Cosine similarity was used because it measures similarity between two vectors and is well suited for comparing TF-IDF representations, which are continuous and sparse [14]. The PageRank algorithm [15] was then applied to rank sentences by their centrality within this graph. Finally, the top-ranked sentences were selected and ordered according to their original sequence in the document to create the summary.

2.1.2. Concept Based

Concept based summarization techniques focus on representing sentences using clinical concepts extracted from an external knowledge base, rather than relying solely on the words themselves. Sentences are represented by the set of clinical concepts they contain and their importance is determined by measuring the similarity between these concept sets [10, 16].

To implement this technique, we used QuickUMLS [17], an open-source tool that extracts clinical concepts from text by mapping them to the Unified Medical Language System (UMLS) metathesaurus [18]. For each sentence, clinical concepts were identified and represented as sets. We then computed pairwise sentence similarities using the Jaccard similarity coefficient [19] over these concept sets, creating a similarity matrix. Jaccard similarity was selected because it measures the similarity between two sets and is appropriate for comparing sentences represented as sets of extracted clinical concepts [14]. This matrix was used to build a weighted graph, where sentences are nodes connected by edges weighted by their concept similarity. Finally, the PageRank algorithm [15] was applied to rank sentences based on their importance in the graph structure, with the top-ranked sentences selected to form the summary.

However, the coverage of clinical terminologies across languages is uneven. While QuickUMLS provides support for several languages, its resources are significantly skewed toward English. Previous studies have reported that non-English counterparts of the UMLS lack between 65% and 94% of the term coverage available in English [20]. This limitation can reduce the number and consistency of extracted concepts in non-English texts, which may impact the effectiveness of this approach in multilingual settings.

2.1.3. Topic Based

Topic based summarization techniques focus on identifying the main themes or topics of the document by measuring how relevant each sentence is to the overall content [10]. Unlike the graph based method, which relies on sentence-to-sentence similarity, this approach measures sentence-to-document topical salience.

To implement this technique, we used TF-IDF vectors [13], as calculated by the formula is Equation (1), to represent sentences. Sentences with higher overall TF-IDF weights were considered more important for capturing the main topics. The top-scoring sentences were selected to form the summary.

2.1.4. Clustering Based

This technique identifies the most central and relevant sentences in a cluster [10]. Sentence centrality is based on the distance between a sentence and the centroid of the document cluster in vector space, with sentences closest to the centroid considered the most important.

To implement this technique, we used the BERT Summarizer [21] from the Python module "`bert-extractive-summarizer`". Particularly, we used the "`bert-base-multilingual-cased`" variant that has been trained on a corpus of raw Wikipedia texts in 104 languages [22]. This method applies BERT (Bidirectional Encoder Representations from Transformers) [23] to encode sentences into contextual embeddings that capture semantic meaning. Then, the embeddings are clustered using K-means and for each cluster the sentence nearest to the centroid is selected as a summary candidate. This process enables the extraction of sentences that collectively represent the document's main content while preserving the original wording. The number of sentences included in the final summary can be specified as a parameter, allowing flexible control over the summary length while requiring minimal manual tuning [24].

2.2. Evaluation Methods

To evaluate the text summarization systems, both ROUGE and BERTScore metrics were employed.

2.2.1. ROUGE Scores

ROUGE scores (Recall-Oriented Understudy for Gisting Evaluation) [25] are a set of metrics used to evaluate automatic text summarization systems. They compare the generated summaries to human-written reference summaries by counting overlapping elements, such as n-grams, word sequences or word pairs, between the system-generated summary and the reference texts. The more overlap, the higher the score, indicating better alignment with the human-written reference.

Specifically, the ROUGE-Lsum variant [26] was used, which interprets newline characters as sentence boundaries and computes the union of the Longest Common Subsequences (LCS) across sentence pairs.

This variant, commonly used in neural summarization research, is well suited for evaluating the overall structure and content fidelity of full summaries rather than individual sentence matches.

2.2.2. BERTscores

BERTScore [27] is an evaluation metric that measures how similar a generated text is to a reference text based on meaning, not just word overlap [28]. Unlike ROUGE, which relies on exact word matches, BERTScore uses contextual word embeddings from a pre-trained BERT model [27].

To calculate it, BERTScore first encodes both the candidate summary and the reference summary using BERT. Then, it compares the words in both texts by measuring how similar their embeddings are using cosine similarity. These alignments are used to calculate precision, recall and a modified F1 score, which is weighted using inverse document frequency (IDF) to reduce the impact of very common words [27].

BERTScore has been shown to correlate better with human judgments, especially in tasks like translation and paraphrasing, because it can recognize similar meanings even when different words are used [29].

3. Experiments and Results

Before applying the summarization methods, the clinical text provided in the challenge was preprocessed. This included converting all characters to lowercase, removing unnecessary spaces and replacing decimal points with temporary placeholders to avoid confusion with sentence boundaries. Specifically, to prevent splitting sentences at decimal points in numbers (e.g., "5.3"), all floats, detected with a regular expression, were temporarily replaced by the same number with "DOT" instead of the period (for example, "5DOT3"). Additional cleaning steps included normalizing whitespace and standardizing certain phrases, such as replacing "and/or" with "and or".

The experiments were conducted on the gold-standard training and test datasets provided in the challenge in four languages: English, Spanish, French and Portuguese [30]. For each language, the training set consisted of 592 full-text clinical case reports paired with human-written summaries. The test set contained between 3,396 and 3,469 full-text reports per language (English: 3,396, Spanish: 3,406, French: 3,469, Portuguese: 3,442).

For each language, the previously described extractive summarization methods were applied to extract the 10 most important sentences in the text to form a summary. The resulting summaries were then submitted to the challenge platform [5], where they were evaluated using BERTScore and ROUGE metrics.

We chose to select exactly 10 sentences for each summary to balance coverage and conciseness. The training texts varied widely in length, ranging from a minimum of 4 sentences to a maximum of 197, with an average of approximately 27 sentences. Choosing 10 sentences provides a consistent summary length that is sufficient to capture key information while maintaining brevity.

3.1. English

The results for the English clinical text are summarized in Table 1. The clustering based approach performed best overall. It achieved the highest BERTScore F1 (84.91%), indicating that the selected sentences were semantically close to the human written summaries. It also had the highest ROUGE F1 (24.67%), suggesting better surface level overlap.

Interestingly, concept based summarization tied with graph based in BERTScore (84.46%), but outperformed it in ROUGE. This likely reflects the advantage of using clinical concepts, that were extracted from QuickUMLS, which helped better match medical terms in the reference summaries. Topic based summarization was also competitive, with a slightly higher BERTScore than the graph and concept based, likely because TF-IDF helped surface more globally relevant sentences.

Table 1
Scores of English Clinical Text In Percentages (%)

System	BERTScore			ROUGE		
	Precision	Recall	F1	Precision	Recall	F1
Graph Based	86.56	82.50	84.46	38.66	16.41	21.99
Concept Based	86.67	82.39	84.46	38.59	17.72	23.13
Topic Based	86.47	82.75	84.55	37.34	17.37	22.55
Clustering Based	**87.06**	**82.90**	**84.91**	**38.89**	**19.40**	**24.67**

Table 2
Scores of Spanish Clinical Text In Percentages (%)

System	BERTScore			ROUGE		
	Precision	Recall	F1	Precision	Recall	F1
Graph Based	74.45	68.71	71.43	**42.53**	17.52	23.65
Concept Based	74.14	68.78	71.31	38.84	18.75	23.98
Topic Based	74.25	68.85	71.40	39.71	18.20	23.70
Clustering Based	**74.99**	**69.03**	**71.85**	39.17	**19.98**	**25.17**

The graph based method had the lowest ROUGE score, which could be because it relies heavily on sentence-to-sentence similarity. If key clinical terms are not repeated or well connected, this method might miss important isolated information.

3.2. Spanish

Again, in the case of Spanish, clustering based summarization had the strongest performance in both metrics. This suggests that sentence embeddings generated from multilingual BERT were effective in identifying central sentences in Spanish clinical texts. Results are shown in Table 2.

While concept based summarization performed well in ROUGE (23.98%), its slightly lower BERTScore indicates that although it picked sentences with overlapping terms, the semantic meaning might have been less aligned. This might be due to QuickUMLS's Spanish support being more limited compared to English, where only 189,563 medical concepts in Spanish were in the database as opposed to 585,453 in English, affecting concept extraction quality.

The topic based method did slightly better than the concept based in BERTScore, possibly because TF-IDF can still capture general topical content even when concept extraction is weaker.

3.3. French

For French, only the topic based and clustering based methods were applied, results are shown in Table 3. Similar to the previous languages, clustering based had the strongest performance. This confirms that encoding-based representations, such as those from BERT, work well across different languages and domains, including French clinical text. [1]

3.4. Portuguese

Finally, in Portuguese, the clustering based summarizer outperformed the topic based method. The margin here was smaller than in other languages, but still consistent with the overall trend, as the clustering based summarization gave the most balanced and accurate summaries across languages. Results for the Portuguese extractive summarization are shown in Table 4.

[1] Due to time constraints and focus on benchmarking strong baseline methods, we limited our submission for French and Portuguese to the two most promising techniques based on early development: clustering and topic-based summarization.

Table 3
Scores of French Clinical Text In Percentages (%)

System	BERTScore			ROUGE		
	Precision	Recall	F1	Precision	Recall	F1
Topic Based	74.23	69.16	71.56	**37.43**	17.57	22.66
Clustering Based	**75.20**	**69.59**	**72.24**	37.27	**19.25**	**24.13**

Table 4
Scores of Portuguese Clinical Text In Percentages (%)

System	BERTScore			ROUGE		
	Precision	Recall	F1	Precision	Recall	F1
Topic Based	73.91	68.52	71.07	**37.53**	17.58	22.66
Clustering Based	**74.66**	**68.76**	**71.54**	37.47	**18.92**	**23.88**

4. Discussion

Across all four languages, the clustering based extractive summarization method achieved the best performance in both semantic similarity (BERTScore) and lexical overlap (ROUGE). This suggests that sentence embeddings from multilingual BERT can effectively capture important content across different languages. Interestingly, even though the Multilingual BERT was not trained with an explicit cross-lingual objective, it still provides strong multilingual representations. This supports previous findings [22], that Multilingual BERT generalizes well across languages for various downstream tasks, despite being trained only on monolingual Wikipedia data without alignment between languages.

However, while Multilingual BERT showed consistent performance across languages, English summaries scored noticeably higher in BERTScore, with an average F1 of 84.60%, compared to 71.50% in Spanish, 71.90% in French and 71.31% in Portuguese. This indicates stronger semantic similarity between the generated and reference summaries in English, likely due to better sentence representations.

Interestingly, Spanish achieved the highest average ROUGE F1 score (24.13%), even surpassing English (23.09%). This suggests that extractive methods in Spanish may be better at capturing lexical overlap with human-written summaries.

Another interesting observation is that the topic based method performed the best in Spanish than the other languages. This suggests that TF-IDF tokenization may be more effective in Spanish for this task. One reason for this could be the difference in subword tokenization coverage. Generally, the tokenization performance of each language is correlated with the language's subword dictionary coverage. Spanish has broader coverage in this regard compared to English, which may have contributed to the stronger performance of TF-IDF in Spanish texts [31].

Despite these insights, the general BERTScore and ROUGE values across all languages remain relatively low. This highlights the inherent difficulty of extractive summarization in the clinical domain, where preserving factual accuracy, handling domain specific language and ensuring multilingual consistency are all critical and challenging tasks [3, 4].

5. Conclusion

In this work, we compared four extractive summarization methods on clinical case reports in English, Spanish, French, and Portuguese. Our results showed that the clustering based summarization, using multilingual BERT embeddings, consistently achieved the best performance in both BERTScore and ROUGE metrics across all languages. English summaries had the highest BERTS scores, probably due to stronger sentence representations, while Spanish had the highest ROUGE, which may reflect better lexical overlap in Spanish, possibly helped by more effective tokenization.

However, the overall scores were relatively low, especially in ROUGE. Moving forward, we could explore combining extractive and abstractive summarization. Abstractive methods may offer more fluent and shorter summaries but should be carefully designed to avoid clinical inaccuracies and hallucinations. A hybrid approach might give the best of both worlds, accuracy from the extractive summarization and readability from the abstractive summarization.

In addition, the decision to retain exactly ten sentences per summary was a heuristic choice aimed at providing consistent coverage across documents. This parameter was not optimized, and future work will include ablation studies and more systematic analysis of length sensitivity. In particular, we intend to investigate how varying summary length affects evaluation metrics in clinical contexts, to better understand the trade-offs between brevity and content coverage.

Acknowledgments

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

Declaration on Generative AI

In preparing this manuscript, the authors utilized OpenAI's ChatGPT to assist with grammar and spelling corrections, enhance writing clarity and rephrase text for readability. All AI-assisted content was carefully reviewed and revised by the authors, who assume full responsibility for the final content.

References

[1] H. Cohen, How to write a patient case report, American Journal of Health-System Pharmacy 63 (2006) 1888–1892. URL: https://doi.org/10.2146/ajhp060182. doi:10.2146/ajhp060182.

[2] Barcelona Supercomputing Center (BSC), Multiclinsum: Multilingual clinical summarization challenge — task information, https://temu.bsc.es/multiclinsum/task-info/, 2025. Accessed: 2025-06-13.

[3] B. Wallace, S. Saha, F. Soboczenski, I. Marshall, Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization, AMIA ... Annual Symposium proceedings. AMIA Symposium 2021 (2021) 605–614. Publisher Copyright: ©2021 AMIA - All rights reserved. Copyright: This record is sourced from MEDLINE/PubMed, a database of the U.S. National Library of Medicine.

[4] A. Elhady, K. Elsayed, E. Agirre, M. Artetxe, Improving factuality in clinical abstractive multi-document summarization by guided continued pre-training, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 755–761. URL: https://aclanthology.org/2024.naacl-short.66/. doi:10.18653/v1/2024.naacl-short.66.

[5] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of MultiClinSum task at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results., in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[6] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. Maria Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, F. P. Josiane Mothe, D. S. Paolo Rosso, G. Faggioli, N. Ferro (Eds.), Experimental IR

Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[7] L. Bednarczyk, D. Reichenpfader, C. Gaudet-Blavignac, A. K. Ette, J. Zaghir, Y. Zheng, A. Bensahla, M. Bjelogrlic, C. Lovis, Scientific evidence for clinical text summarization using large language models: Scoping review, Journal of Medical Internet Research 27 (2025) e68998.

[8] J. Liang, C.-H. Tsou, A. Poddar, A novel system for extractive clinical note summarization using EHR data, in: A. Rumshisky, K. Roberts, S. Bethard, T. Naumann (Eds.), Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 46–54. URL: https://aclanthology.org/W19-1906/. doi:10.18653/v1/W19-1906.

[9] S. Rhazzafe, F. Caraffini, S. Colreavy-Donnelly, Y. Dhassi, S. Kuhn, N. S. Nikolov, Hybrid summarization of medical records for predicting length of stay in the intensive care unit, Applied Sciences 14 (2024). URL: https://www.mdpi.com/2076-3417/14/13/5809. doi:10.3390/app14135809.

[10] W. S. El-Kassas, C. R. Salama, A. A. Rafea, H. K. Mohamed, Automatic text summarization: A comprehensive survey, Expert Systems with Applications 165 (2021) 113679. URL: https://www.sciencedirect.com/science/article/pii/S0957417420305030. doi:10.1016/j.eswa.2020.113679.

[11] C. P. Chai, Comparison of text preprocessing methods, Natural Language Engineering 29 (2023) 509–553. doi:10.1017/S1351324922000213.

[12] V. Gupta, G. S. Lehal, A survey of text summarization extractive techniques, Journal of emerging technologies in web intelligence 2 (2010) 258–268.

[13] C. Sammut, G. I. Webb (Eds.), TF–IDF, Springer US, Boston, MA, 2010, pp. 986–987. URL: https://doi.org/10.1007/978-0-387-30164-8_832. doi:10.1007/978-0-387-30164-8_832.

[14] N. Baruah, S. Gupta, S. Ghosh, S. N. Afrid, C. Kakoty, R. Phukan, Exploring jaccard similarity and cosine similarity for developing an assamese question-answering system, in: Proceedings of World Conference on Artificial Intelligence: Advances and Applications, Algorithms for Intelligent Systems, Springer Nature Singapore, 2023, pp. 87–98. doi:10.1007/978-981-99-5881-8_8.

[15] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1998) 107–117. URL: https://www.sciencedirect.com/science/article/pii/S016975529800110X. doi:https://doi.org/10.1016/S0169-7552(98)00110-X, proceedings of the Seventh International World Wide Web Conference.

[16] N. Moratanch, S. Chitrakala, A survey on extractive text summarization, in: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017, pp. 1–6. doi:10.1109/ICCCSP.2017.7944061.

[17] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, in: MedIR workshop, sigir, 2016, pp. 1–4.

[18] National Library of Medicine (US), Umls® reference manual, https://www.ncbi.nlm.nih.gov/books/NBK9684/, 2009. Updated 2021 Aug 20. 2, Metathesaurus.

[19] J. Hancock, Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient), 2004. doi:10.1002/9780471650126.dob0956.

[20] J. Hellrich, U. Hahn, Fostering multilinguality in the umls: a computational approach to terminology expansion for multiple languages, in: AMIA Annual Symposium Proceedings, volume 2014, 2014, p. 655.

[21] D. Miller, Leveraging bert for extractive text summarization on lectures, 2019. URL: https://arxiv.org/abs/1906.04165. arXiv:1906.04165.

[22] K. K, Z. Wang, S. Mayhew, D. Roth, Cross-lingual ability of multilingual bert: An empirical study, ArXiv abs/1912.07840 (2019). URL: https://api.semanticscholar.org/CorpusID:209183618.

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[24] K. Sethia, M. Saxena, M. Goyal, R. Yadav, Framework for topic modeling using bert, lda and k-means, in: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022, pp. 2204–2208. doi:10.1109/ICACITE53722.2022.9823442.

[25] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization

Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[26] Google LLC, rouge-score: A native python implementation of rouge, https://pypi.org/project/rouge-score/, 2022. Version 0.1.2, licensed under Apache-2.0. Implements ROUGE-N, ROUGE-L, ‘rougeLsum‘, bootstrap CI, Porter stemmer.

[27] A. Chen, G. Stanovsky, S. Singh, M. Gardner, Evaluating question answering evaluation, in: A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, D. Chen (Eds.), Proceedings of the 2nd Workshop on Machine Reading for Question Answering, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 119–124. URL: https://aclanthology.org/D19-5817/. doi:10.18653/v1/D19-5817.

[28] M. Hanna, O. Bojar, A fine-grained analysis of BERTScore, in: L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, C. Monz (Eds.), Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, Online, 2021, pp. 507–517. URL: https://aclanthology.org/2021.wmt-1.59/.

[29] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[30] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Giménez, L. Fernández, E. Farré-Maduell, M. Krallinger, Multilingual clinical summarization (multiclinsum) challenge datasets, 2025. URL: https://zenodo.org/records/15546018. doi:10.5281/zenodo.15546018.

[31] A. Wangperawong, Multilingual search with subword tf-idf, arXiv preprint arXiv:2209.14281 (2022).