ClinicalBertSum: RCT Summarization by Using Clinical BERT Embeddings

Stanford CS224N Custom Project

Mingyi Lu
Department of Electrical Engineering
Stanford University
mingyilu@stanford.edu

Xiaomeng Jin
Department of Computer Science
Stanford University
tracyjxm@stanford.edu

Zihan Wang
Department of Computer Science
Stanford University
wangzih@stanford.edu

Abstract

Text summarization is defined as generating a short, accurate, and fluent summary of a longer input text document, and it is an important task in modern NLP field, which is very useful in several real-world applications. In this paper, we proposed an extractive summarization model called ClinicalBertSum, which is based on BERT [1] and improve the performance on clinical datasets, e.g. PubMed [2]. BERT has dramatically improved performance on a wide range of NLP tasks. However, there are few research working on applying it to text summarization, especially on clinical domains. Our approach address this need by implementing state-of-the-art BertSum model [3] and releasing a pre-trained ClinicalBertSum model for clinical text.

1 Introduction

Over last couple of decades, with the advances in digital technology, the amount of digital biomedical data and resources grow exponentially. PubMed, which is the most well-known and wildly used platform for biomedical literature retrieval system, contains more than 30 million citations and abstracts of biomedical literature and the number is increasing by more than 3,000 every day [4]. From research literature to clinical notes, the types of biomedical data also become more diverse and the values behind them are massive. In order to release the full potential of the biomedical data, large scale data mining techniques are necessary. Due to the advances in natural language processing, the problem of doing text mining on massive biomedical data becomes solvable. The applications include named entity recognition for medical concepts [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17], hypothesis generation and knowledge discovery [18, 19, 20, 21, 22, 23, 24], text summarization for biomedical literature [25, 26, 27, 28, 29], biomedical terminology extraction [30, 31, 32, 33, 34], text classification for diagnosis [35, 36, 37, 38, 39]. These applications become useful tools for biomedical study and research.

By taking a look from NLP perspective, the most impacting research during last two years is the invention of BERT [1]. BERT, which is the abbreviation of bidirectional encoder representations from transformers, uses a masked language model to be pre-trained by using bidirectional transformers. The novelty of BERT is that it learns the bidirectional representations of the words instead of unidirectional ones. The learned representations of BERT outperforms many downstream NLP tasks, such as QA and NER, compared with its peers. BERT is a real breakthrough in NLP research and brings many more extensions and applications to the field. One BERT extension is to fine-tune BERT with task-specific dataset. For example, in order to have a better embedding for biomedical data,

BioBERT, which was trained on PubMed and PMC data, has been created [40]. Similar models include ClinicalBERT, SciBERT and BlueBERT [41, 42, 43].

Among the applications of text mining in biomedical research, text summarization is an important tool to summarize long literature into short ones and it let the researchers use less time to capture the most important content in the readings. The goal of our research is to do automatic summarization on the clinical notes and medical abstracts. Our research uses BERT as our base model for extractive summaries and leverages the performance of BERT by using different variations.

2 Related Work

2.1 Text Summarization

Text summarization is the task to automatically produce a brief summary of a paragraph or an article which preserves the key information [44]. There are two major categories of text summarization: extractive summarization and abstractive summarization. Extractive summarization is to select sentences from the original text to construct a summary of the document. Abstractive summarization is to create a new paragraph by using natural language generation to summarize the original document. Normally, abstractive summarization methods are more difficult and complex than the extractive summarization methods, but they can produce a more flexible and concise summary. Text summarization can also be classified as single and multiple-document summarizations [45]. In single-document summarization, just one document is used for the algorithm to summarize. But for multiple-document, there are many documents used for generating the summary.

Before the use of neural network, text summarization systems are mostly extractive. It normally follows a typical pipeline: content selection, information ordering and sentence realization [46]. The content selection step is to select the sentences which will be include in the summary. The ways of selecting sentences can be categorized into two main classes: sentence scoring function and graph-based algorithms. Sentence scoring function is based on the presence of topic keywords and sentence features. Graph-based algorithms treats each sentence as a node and the sentence pair as edge. The weight of edge is proportional to sentence similarity. Graph algorithms are used to select the sentences which is central in the paragraph and meaningful to include in the summary. Information ordering is to determine the ordering of selected sentences and sentence realization is to do specific editions on the summary to make it more precise and readable.

In 2015, the neural network has been first used in text summarization [47]. It formalized the single-document abstractive summarization as a translation task and then applied standard sequence to sequence and attention neural machine translation model to solve the problem. The down side of seq2seq and attention systems is that it does not work well for copying over details. Then, the exploration on neural text summarization leans to finding efficient copying mechanisms. Copy mechanism is to use attention to enable a text summarizaiton model to copy words and phrases from the input to output. Some examples of copy mechanism and its variants are [48, 49, 50].

2.2 BERT and BERT Variations

In the field of NLP, the most impacting research during last two year is the invention of BERT [1]. BERT, which is the abbreviation of bidirectional encoder representations from transformers, uses a masked language model to be pre-trained by using bidirectional transformers. The novelty of BERT is that it learns the bidirectional representations of the words instead of unidirectional ones. It has been trained on Wikipedia and a book corpus. The learned representations of BERT outperforms many downstream NLP tasks, such as QA and NER, compared with its peers.

Though BERT is the state-of-art model for general natural language processing, it does not guarantee satisfactory results for the other domains. Therefore, there are variants of BERT, which was fine-tuned on other dataset. BioBERT is a pre-trained biomedical language representation model for biomedical text mining [cite]. It was trained on PubMed and PMC and tested on various tasks, such as biomedical named entity recognition, relation extraction and question answering. It shows a better performance on these tasks compared with original BERT model. Another variation of BERT is SciBERT, which is a pre-trained model based on BERT and fine-tuned on a large corpus of scientific text [51]. BERT can also be used in clinical domain. ClinicalBERT is an example of BERT trained on clinical notes [42]. Other BERT model includes BlueBERT and so on [43].

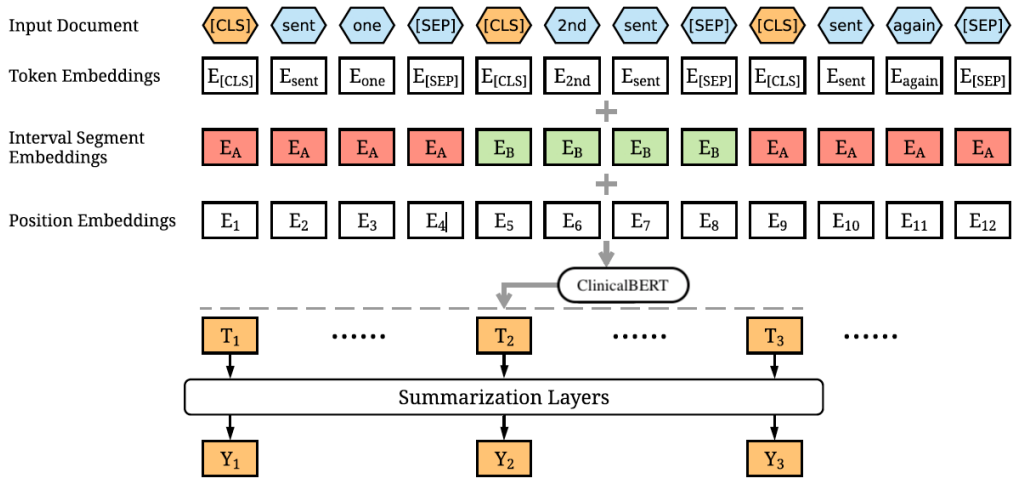


Figure 1: The architecture of our ClinicalBertSum model. Figure comes from the original BertSum [3] paper and is slightly changed to fit our model.

3 Approach

The objective of our research is to apply fine-tuned BERT model for medical abstract summarization. Our approach to tackle this problem has three major components: 1) a fine-tuned BERT model on clinical notes, ClinicalBERT; 2) a find-tuned BERT model on scientific data, SciBERT; 3) a BERT-based text summarization model, BertSum [52, 51, 53].

We have introduced BioBERT, which was trained on English Wikipedia and BooksCorpus, and fine-tuned by using biomedical corpora, such as PubMed Abstracts and PMC Fulltext articles. Addition to BioBERT, we used ClinicalBERT model which was build based on the BioBERT model and fine-tuned BioBERT by using clinical notes, which are from the MIMIC-III v1.4 database [54]. We used ClinicalBERT as a pre-trained model and apply the ClinicalBERT in the BERT-based summarization model. We call our model architecture as ClinicalBertSum. In the final report, we find a more powerful and appropriate tool than BioBERT, which is SciBERT, to help evaluate and analyze our ClinicalBERTSum model.

For summarization, we used the model BertSum as our primary model for extractive summarization [53]. BertSum is a fine-tuned BERT model, which works on the single document extractive and abstractive summarization. The model encodes the sentences in a documents by combining three different types of embeddings: token embeddings, interval segment embeddings and position embeddings. After encoding the original sentence, the embeddings are fed into the BERT model to obtain the sentence vectors. The sentence vector then has been used as input to the additional summarization layers to generate summaries by computing the final predicted score, $\hat{Y}_i$. We used the inter-sentence transformer as our summarization layer:

$$\widetilde{h}^l = LN(h^{l-1} + MHAtt(h^{l-1}) \tag{1}$$

$$h^l = LN(\widetilde{h}^l + FFN(\widetilde{h}^l)) \tag{2}$$

where h^0 is the position embedding of sentence vector T, LN is the larger normalization operator and $MHAtt$ is the multi-head attention operator. The output layer is a sigmoid classifier:

$$\hat{Y}_i = \sigma(W_o h_i^L + b_o) \tag{3}$$

The output of the classifier is used for selecting the sentences for summarization. During training, the BERT model is updated with the summarization layer and output layer.

The novelty of our work is to leverage the utility of ClinicalBERT to summarize medical literatures and abstracts. Our work is mainly based on ClinicalBERT and BertSum. Both works are available on github[1][2].

4 Experiments

4.1 Data

The dataset we are using is a modified version of a medical abstract dataset from PubMed, PubMed 200k RCT [2]. This dataset is originally used for sequential sentence classification. It focuses on the medical abstracts and especially on the randomized controlled trials (RCTs), which are normally the best source of medical evidence. In this dataset, each sentence had been labeled with a specific class, which indicates the section where the selected sentence from. There are five classes, which include *Objective*, *Background*, *Conclusions*, *Methods* and *Results*. All the selected abstracts follows two criteria: 1) belong to RCT; 2) must be structured. There are 195,654 abstracts satisfying the both criteria and split into three datasets, train (190,654), validation (2,500) and test (2,500).

In our setup, we slightly modify the use of the dataset. Our task is to summarize medical abstracts by using BERT-based models. Therefore, we combined the sentences with following classes of each abstract into one paragraph: *Objective*, *Background*, *Methods* and *Results*. We used the synthesized paragraph as the main body of text to be summarized. Furthermore, the sentences with class, *Conclusion*, were used as the reference summary of the medical abstract.

4.2 Evaluation Method

We are using ROUGE score to evaluate the performance of our model. ROUGE is mostly common-used for evaluating the summarization ability.

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [55] is a set of metrics for evaluating machine translations and automatic summarizations. We will report the following F1 score of ROUGE scores:

ROUGE-1: the overlap of the unigrams among the system and reference summarization.

ROUGE-2: the overlap of the bigrams among the system and reference summarization

ROUGE-L: the longest common subsequence statistics.

4.3 Experimental details

Our model configurations basically follows the settings from original BertSum paper [53]. To fine tune the pre-trained BERT or clinicalBERT model, we set Adam Optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$. Following this paper's setting [56], Learing rate has warming-up on first 10,000 steps:

$$lr = 2e^{-3} \cdot \min\left(step^{-0.5}, step \cdot warmup^{-1.5}\right)$$

All models are trained on a GPU (GTX 2080 Ti) for 50,000 steps. We use gradient accumulation every two steps, so our batch size approximates to 36. We saved model checkpoints every 1,000 steps, and select the highest three checkpoints based on performance on validation set to test. The training time for each model costs around 7.5 hours, and validation spends around 1.5 hours.

4.4 Results

Table 1 summarizes our results on both CNN/DailyMail and RCT dataset. The first line in the table displays the results of training on the CNN + DM dataset and the ROUGE score of CNN + DM testing set. This is to confirm the correctness of our reproducing the trained model. The second line is the results of our model baselines. The third line summarizes our approach on applying BERT to clinical notes summarization. The performance of our model outperforms the baseline, but there

[1] ClinicalBERT: https://github.com/kexinhuang12345/clinicalBERT

[2] BertSum: https://github.com/nlpyang/BertSum

Table 1: Comparison between different training sets

Training Set	Testing Set	ROUGE 1	ROUGE 2	ROUGE L
CNN + DM	CNN + DM	43.25	20.20	39.65
CNN + DM	RCT	31.79	10.35	25.78
RCT	RCT	33.58	11.87	27.41

Table 2: Comparison between different training sets

Pre-trained BERT model	Training Set (BertSum)	ROUGE 1	ROUGE 2	ROUGE L
BERT-uncased	CNN + DM	43.25	20.20	39.65
	RCT	31.69	10.30	25.73
SciBERT	CNN + DM	42.69	19.66	39.06
	RCT	33.70	11.77	27.57
ClinicalBERT	CNN + DM	42.98	20.03	39.38
	RCT	33.58	11.87	27.41

are still gaps from the CNN/Dailymail dataset trained and tested on bertSum model. To explore this problem, we arrange groups of experiments, summarized in table 2.

We choose pubMed dataset described in the above section to be our training and testing dataset. Table 2 summarizes our experiments of this dataset on three different pre-trained BERT models: The Original pre-trained BERT (uncased), The Clinical BERT and the SciBERT [51]. After fixing the training data, we see that the original BERT-uncased, which is used in bertSum, performs poorer than our clinical BertSum, and also poorer than SciBERT. SciBERT trained on research papers from semantic scholar, which includes a lot of papers from biomedical and medical fields, and it is reported to be achieves SOTA results on some BioBERT results on biomedical tasks, therefore, we use it as an extension of bioBERT.

5 Analysis

In this section, we will analyze our result not only based on the ROUGE evaluation metric, but also by analyzing the generated summaries. This is because we find the disadvantages of ROUGE metrics, and even if we have already tried to figure out this problem, e.g. finding a better evaluation metric, however, it is beyond the scope of this final project. Thus, to better present our result, we will choose examples from generated data directly.

5.1 Selected Examples from Summaries

- **Example 1. From the reproduced result of BertSum.**

  Target Summary: *the 79th masters tournament gets underway at augusta national on thursday<q>rory mcilroy and tiger woods will be the star attractions in the field bidding for the green jacket at 2015 masters<q>mcilroy , justin rose , ian poulter , graeme mcdowell and more gave sportsmail the verdict on each hole at augusta<q>click on the brilliant interactive graphic below for details on each hole of the masters 2015 course<q>click here for all the latest news from the masters 2015*

  Generated Summary: *to help get you in the mood for the first major of the year , rory mcilroy , ian poulter , graeme mcdowell and justin rose , plus past masters champions nick faldo and charl schwartzel , give the lowdown on every hole at the world-famous augusta national golf club .<q>the masters 2015 is almost here .<q>click on the graphic below to get a closer look at what the biggest names in the game will face when they tee off on thursday .*

- **Example 2. From clinical BertSum.**

  Target Summary: *Amifostine is effective in reducing the incidence and severity of acute mucositis , acute and late xerostomia in head and neck cancer patients .*

Table 3: Comparison of summarization datasets. Statistics of CNN and DailyMail dataset are from BertSum paper [3].

Dataset	**Avg. doc length**		**Avg. summary length**	
	words	sentences	words	sentences
CNN	760.50	33.98	45.70	3.59
DailyMail	653.33	29.33	54.65	3.86
RCT	267.6	10.1	41.6	1.8

Generated Summary: *Amifostine has a potential role for salivary gland protection in head and neck cancer patients who had radiotherapy .<q>The salivary gland function returned to normal at a rate of 36.3 % in the Amifostine group versus 9.1 % in the control group .<q>The efficacy of the treatment was determined by a questionnaire evaluating dryness of mouth and the oral comfort , the RTOG/EORTC acute/late radiation morbidity scoring criteria , collection of the whole saliva and the 99mTc-pertecnetate scintigraphy of the salivary glands .*

These are two testing examples that generated by our model. The first example is a piece of CNN article summary. The second one is a clinical abstract (RCT) summary. From the result table we can tell that the performance of CNN testing set is better than RCT according to the ROUGE score.

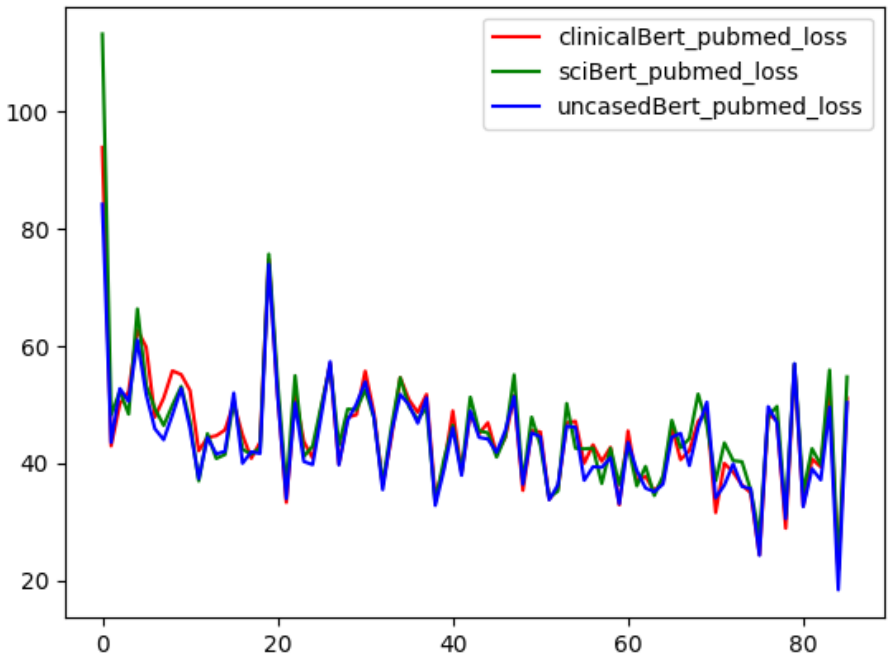
5.2 Limitation of ROUGE

In our observation, we find that even if our clinical summary is reasonable and looks like an efficient summary of the original text, the ROUGE score is still low. In addition, compared to the CNN/Daily mail summary generated by BertSum, our clinical summaries are more readable and represent the main claims of documents. Therefore, the fairness of ROUGE score is doubted. Schluter claimed in his paper [57] that 1) It is theoretically computationally hard to achieve perfect score for extractive memorization; 2) For short summaries, ROUGE scores are generally rather low, and it seems to get higher scores for longer summary datasets. We counted our document length and the average length of CNN/DailyMail dataset, and the result is shown in Table 3.

We claim that without balance of document length of different dataset, we are not able to directly compare results between datasets. Thus, we should not compare the ROUGE score of BertSum model trained on CNN/Daily Mail dataset and ClinicalBertSum model trained on PubMed dataset, which means the result in Table 1 does not means our result are much poorer than baseline. Therefore, the experimental results of different models fixed on the same PubMed dataset are more convincing, and our analyze mainly focus on Table 2.

5.3 Different Pretrained model performances

As we can see from Table 2, we selected three different pretrained BERT models for initialization. According to the ROUGE score, we can see that both SciBERT and ClinicalBERT outperformed than loading the original BERT-uncased model. The original BERT has been trained on a large corpus and doesn't have any specific focusing areas. However, SciBERT and ClinicalBERT have been trained on the scientific papers, clinical notes, and biomedical tests, the performance of loading these two pre-trained BERT models are definitely better.



The plot consists of the training loss of using three different pretrained-models: clinicalBert, sciBert, uncasedBert. From the plot we can see that the training procedures are very similar.

6 Conclusion

In our work, we have introduced a new method to summarize clinical and medical abstracts, ClinicalBertSum. The model has a better performance on RCT data than using the original BERT model. Furthermore, we compared the synthesized summaries with the true summaries and analyzed the limitations of ROUGE evaluation metric.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[2] Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*, 2017.

[3] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics, 2019.

[4] Johanna McEntyre and David Lipman. Pubmed: bridging the information gap. *Cmaj*, 164(9):1317–1319, 2001.

[5] Adrian Benton, Shawndra Hill, Lyle Ungar, Annie Chung, Charles Leonard, Cristin Freeman, and John H Holmes. A system for de-identifying medical message board text. *BMC bioinformatics*, 12(S3):S2, 2011.

[6] Francisco Carrero, José Carlos Cortizo, and José María Gómez. Building a spanish mmtx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 346–353. Springer, 2008.

[7] Oscar Ferrández, Brett R South, Shuying Shen, and Stéphane M Meystre. A hybrid stepwise approach for de-identifying person names in clinical documents. In *Proceedings of the 2012 workshop on biomedical natural language processing*, pages 65–72. Association for Computational Linguistics, 2012.

[8] Karin Kipper-Schuler, Vinod Kaggal, James Masanz, Philip Ogren, and Guergana Savova. System evaluation on a named entity corpus from clinical notes. In *Language resources and evaluation conference, LREC*, pages 3001–3007, 2008.

[9] Yu-Kai Lin, Hsinchun Chen, and Randall A Brown. Medtime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46:S20–S28, 2013.

[10] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[11] Kirk Roberts and Sanda M Harabagiu. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18(5):568–573, 2011.

[12] Kirk Roberts, Bryan Rink, and Sanda M Harabagiu. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, 20(5):867–875, 2013.

[13] Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158, 2014.

[14] Ozlem Uzuner, Jonathan Mailoa, Russell Ryan, and Tawanda Sibanda. Semantic relations for problem-oriented medical records. *Artificial intelligence in medicine*, 50(2):63–73, 2010.

[15] Yefeng Wang and Jon Patrick. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the workshop on biomedical information extraction*, pages 42–49. Association for Computational Linguistics, 2009.

[16] Yunqing Xia, Xiaoshi Zhong, Peng Liu, Cheng Tan, Sen Na, Qinan Hu, and Yaohai Huang. Combining metamap and ctakes in disorder recognition: Thcib at clef ehealth lab 2013 task 1. In *CLEF (Working Notes)*, 2013.

[17] Xiaodan Zhu, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Berry De Bruijn. Detecting concept relations in clinical text: Insights from a state-of-the-art model. *Journal of biomedical informatics*, 46(2):275–285, 2013.

[18] John A Baron, Stephen Senn, Michael Voelker, Angel Lanas, Irene Laurora, Wolfgang Thielemann, Andreas Brückner, and Denis McCarthy. Gastrointestinal adverse effects of short-term aspirin use: a meta-analysis of published randomized controlled trials. *Drugs in R&d*, 13(1):9–16, 2013.

[19] Roy J Byrd, Steven R Steinhubl, Jimeng Sun, Shahram Ebadollahi, and Walter F Stewart. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International journal of medical informatics*, 83(12):983–992, 2014.

[20] Tyler S Cole, Jennifer Frankovich, Srinivasan Iyer, Paea LePendu, Anna Bauer-Mehren, and Nigam H Shah. Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for ehr-based research. *Pediatric Rheumatology*, 11(1):45, 2013.

[21] Nigel Collier. Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global public health*, 7(7):731–749, 2012.

[22] Norris H Heintzelman, Robert J Taylor, Lone Simonsen, Roger Lustig, Doug Anderko, Jennifer A Haythornthwaite, Lois C Childs, and George Steven Bova. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association*, 20(5):898–905, 2013.

[23] Ahmad P Tafti, Jonathan Badger, Eric LaRose, Ehsan Shirzadi, Andrea Mahnke, John Mayer, Zhan Ye, David Page, and Peggy Peissig. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics*, 5(4):e51, 2017.

[24] Yuan Yang, Pengtao Xie, Xin Gao, Carol Cheng, Christy Li, Hongbao Zhang, and Eric Xing. Predicting discharge medications at admission time based on deep learning. *arXiv preprint arXiv:1711.01386*, 2017.

[25] Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. Summarization from medical documents: a survey. *Artificial intelligence in medicine*, 33(2):157–177, 2005.

[26] Noemie Elhadad, M-Y Kan, Judith L Klavans, and Kathleen R McKeown. Customization in a unified framework for summarizing medical literature. *Artificial intelligence in medicine*, 33(2):179–198, 2005.

[27] Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, and Thomas C Rindflesch. Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of biomedical informatics*, 42(5):801–813, 2009.

[28] Thomas C Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. Semantic medline: An advanced information management application for biomedicine. *Information Services & Use*, 31(1-2):15–21, 2011.

[29] Han Zhang, Marcelo Fiszman, Dongwook Shin, Christopher M Miller, Graciela Rosemblat, and Thomas C Rindflesch. Degree centrality for semantic abstraction summarization of therapeutic studies. *Journal of biomedical informatics*, 44(5):830–838, 2011.

[30] Yu-Ching Fang, Hsuan-Cheng Huang, Hsin-Hsi Chen, and Hsueh-Fen Juan. Tcmgenedit: a database for associated traditional chinese medicine, gene and disease information using text mining. *BMC complementary and alternative medicine*, 8(1):1–11, 2008.

[31] Stephen Luther, Donald Berndt, Dezon Finch, Matthew Richardson, Edward Hickling, and David Hickam. Using statistical text mining to supplement the development of an ontology. *Journal of biomedical informatics*, 44:S86–S93, 2011.

[32] Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966, 2009.

[33] Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884, 2012.

[34] Boya Xie, Qin Ding, Hongjin Han, and Di Wu. mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, 29(5):638–644, 2013.

[35] Elisabeth Metais, Didier Nakache, and Jean-François Timsit. Automatic classification of medical reports, the cirea project. In *Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey*, pages 354–359, 2006.

[36] Serguei Pakhomov, Susan A Weston, Steven J Jacobsen, Christopher G Chute, Ryan Meverden, Veronique L Roger, et al. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care*, 13(6 Part 1):281–288, 2007.

[37] Rajakrishnan Vijayakrishnan, Steven R Steinhubl, Kenney Ng, Jimeng Sun, Roy J Byrd, Zahra Daar, Brent A Williams, Christopher Defilippi, Shahram Ebadollahi, and Walter F Stewart. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of cardiac failure*, 20(7):459–464, 2014.

[38] Meliha Yetisgen-Yildiz and Wanda Pratt. The effect of feature representation on medline document classification. In *AMIA annual symposium proceedings*, volume 2005, page 849. American Medical Informatics Association, 2005.

[39] Guido Zuccon, Amol S Wagholikar, Anthony N Nguyen, Luke Butt, Kevin Chu, Shane Martin, and Jaimi Greenslade. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. *AMIA Summits on Translational Science Proceedings*, 2013:300, 2013.

[40] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[41] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[42] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[43] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

[44] Halizah Basiron, Yogan Jaya Kumar, Sing Goh Ong, Hea Choon Ngo, and Puspalata C Suppiah. A review on automatic text summarization approaches. *Journal of Computer Science*, 12:178–190, 2016.

[45] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

[46] Constituency Parsing. Speech and language processing. 2009.

[47] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[48] Yishu Miao and Phil Blunsom. Language as a latent variable: Discrete generative models for sentence compression. *arXiv preprint arXiv:1609.07317*, 2016.

[49] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[50] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

[51] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

[52] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*, 2019.

[53] Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019.

[54] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[55] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[57] Natalie Schluter. The limits of automatic summarisation according to ROUGE. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 41–45. Association for Computational Linguistics, 2017.