Reference-free Medical Multi-document Summary Evaluation Metric via Contrastive Learning

Jimin Lee[1] and **Hwanhee Lee**[1†]
[1]Department of Artificial Intelligence, Chung-Ang University
{ljm1690, hwanheelee}@cau.ac.kr

Abstract

Despite the advancement of automatic summarization methods based on pre-trained language models, evaluating their effectiveness remains a challenge, particularly in the absence of a medical document reference-free summary evaluation metric. This paper proposes a novel reference-free evaluation metric for medical document summaries by employing contrastive learning using medical text-tailored data augmentation techniques. Our research showcases the metric's superior performance in assessing the quality of generated summaries without the need for comparison texts. Through extensive experimentation and analysis, this work makes significant strides in improving the reliability and usability of automatic medical document evaluation tools in medical document settings.

1 Introduction

In the rapidly evolving field of healthcare systems, the ability to quickly and accurately summarize medical documents can significantly aid professionals in keeping abreast of the latest developments and making informed decisions. Medical multi-documents are filled with specialized terminology, abbreviations, and jargon that can be challenging to interpret and summarize accurately (Abdollahi et al., 2021). Moreover, medical documents often contain nuanced information that requires a professional understanding of the context. Hence, evaluating the quality of medical summaries is complex compared to general document summary tasks. Moreover, medical document summary evaluation models struggle to capture medical details or nuances (Meystre and Haug, 2006). Because of these difficulties, researchers often require medical experts to assess the accuracy, completeness, and coherence of the summaries, making the evaluation process time-consuming and costly. For previous work on medical document summarization evaluation, widely used metrics such as ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020b), and Delta-Ei (DeYoung et al., 2020) have been used. Some metrics leverage sentence-bert (Reimers and Gurevych, 2019) and compute the cosine similarity between embeddings of target summary and generated summary to use as an evaluation metric.

Method	ROUGE	BERT-S	NLI
Correlation	-0.010	0.022	0.053

Table 1: Low correlation coefficients with human judgments for the widely used metrics on medical document summarization tasks.

There have been many trials to capture and summarize key points of related medical multi-documents and evaluate their quality. However, our findings show these metrics are not well-suited for medical document summarization evaluation, as n-gram similarity metrics such as ROUGE and models fine-tuned on datasets like SciFact (Wadden et al., 2022) have disappointingly low correlations with human evaluation scores proposed in the previous work (Wang et al., 2023), under 0.2 as shown in Table 1. This suggests traditional metrics do not adequately reflect human judgment for summarization tasks. Due to the nature of medical text data, the meaning of a sentence can completely change with just one-word alteration and the relationships between words are crucial. Capturing these minor changes is significantly important in evaluating the medical summaries.

In this work, we propose a reference-free medical multi-document summary evaluation metric which is not dependent on the reference ground truth medical summaries and use a document-summarization pair to evaluate the quality of medical document summaries. We develop our method upon the RoBERTa-large (Liu et al., 2019) model.

†Corresponding author.

We fine-tune model via contrastive learning, where the model is trained to distinguish between the ground-truth summaries and precisely augmented various negative medical summaries. The model learns to differentiate the real summary from the augmented summary, enhancing its ability to accurately understand and evaluate summary like medical professionals. We evaluate our proposed metric on the human evaluation medical dataset. Our experimental results show high correlations with human evaluations which outperform previous medical summary evaluation metrics.

2 Related work

EDA (Wei and Zou, 2019) introduces simple and effective data augmentation techniques for the texts. EDA proposes various data augmentation techniques such as synonym replacement, random swap, and random deletion methods which enormously improve performance on text classification tasks. Our work applied these data augmentation techniques which can help models capture subtle characteristics of medical documents. Automated metrics for medical multi-document summarization disagree with human evaluations(Wang et al., 2023) proposed human-evaluation datasets from MSLR (DeYoung et al., 2021) shared task. They introduced a dataset of human-assessed summary quality facets from the multi-document summarization for Literature Review and found out inefficiency of conventional metrics for summarization evaluation. Our work uses the human evaluation scores which can validate our model's performance. UMIC (Lee et al., 2021) introduces a metric that does not require reference captions to evaluate image captions. Umic uses negative captions and fine-tuning the UNITER (Chen et al., 2020) model via contrastive learning. Inspired by this idea, we adopt a method to train the negative dataset through contrastive learning. Using the various data augmentation techniques and contrastive learning approach, we propose a new metric for medical document summary evaluation.

3 Methods

We develop a new medical document sumarmization metric through the following three steps. First, we augment the MSLR Cochrane dataset using 6 different methods. Then, we generate entailment scores of the original summary and augmented summary. We formulate the final dataset to this form: *(original summary, augmented summary, entailment score)*. (Sec 3.1) Finally, we train the metric as a contrastive learning approach by learning the representation of medical document summaries by comparing similar *(positive medical document)* and dissimilar *(negative medical document)* pairs of data points. (Sec 3.2)

3.1 Medical Summary Augmentation

Our augmentation approach is aimed not only to include obvious differences but also to contain biomedically minor, subtle differences from the original medical summaries. These methods are meticulously designed to generate synthetic examples of low-entailment-scoring summaries, thereby enriching the dataset for contrastive learning. Our method lies in using pairs of *(original summary, augmented summary, entailment score)* as inputs for contrastive learning. The original summary is a concise representation of medical documents, while the augmented summary is a modified version of the original, which includes augmented sentences aimed at enhancing the richness of the training data for contrastive learning. The entailment score quantitatively represents the degree of informational overlap between the original and augmented summaries, essentially measuring how well the augmented summary retains the critical information from the original. In this work, we use the following data augmentation techniques to generate new data pairs as in Table 2 for training the proposed metric. Our final training dataset consists of 22,350 augmented summaries.

Synonym Replacement We first use Synonym Replacement (SR), which is designed to generate semantically similar sentences by replacing certain words with their synonyms. This method helps us to diversify the linguistic expression within our dataset without deviating from the original meaning. The words are from WordNet lexical databases (Miller, 1995) which have words to change. For each selected word, look up synonyms that fit the context of the sentence. The next step is replacing the original words with their synonyms. For instance, we substituted 'low-quality' with 'inferior-quality,' 'comparing' with 'contrasting,' and 'advanced' with 'progressed.'.

Random Deletion We also apply Random Deletion (RD), which randomly removes words to simulate different forms of summarization compression; For each sentence in the dataset, words are selected

Data Augmentation Approach
Original Summary: We found only low-quality evidence comparing ultra-radical and standard surgery in women with advanced ovarian cancer and carcinomatosis.
Positive Augmentation Methods
Synonym Replacement (SR): We discovered solely inferior-quality proof contrasting extreme and conventional surgery in females with progressed ovarian cancer and carcinomatosis.
Paraphrase (PAR): Our research yielded only substandard evidence when evaluating the effectiveness of extreme versus traditional surgical approaches in females diagnosed with severe ovarian cancer and carcinomatosis.
Negative Augmentation Methods
Random Deletion (RD): We found low-quality evidence comparing radical and standard surgery in women with ovarian cancer.
Random Swap (RS): Women found only low-quality evidence comparing ovarian and ultra-radical surgery in we with carcinomatosis and advanced standard cancer
Antonym Replacement (AR): We found only high-quality evidence comparing conservative and advanced surgery in women without early ovarian cancer and carcinomatosis.
NER Swap (NER): They found only middle-quality proof comparing ultra-radical and standard surgery in children with advanced ovarian torsion and carcinomatosis

Table 2: Examples of generated medical document summaries through the proposed data augmentation approaches.

at random for deletion. The selection process is uniform, where each word has an equal chance of being deleted. The selected words are removed from the sentence.

Random Swap We adopt Random Swap (RS), which randomly selects pairs of words within a sentence and swaps their positions, which can increase linguistic diversity without significantly altering the semantic integrity of the information conveyed. In the original sentence, two specific swaps are performed. The term 'ovarian' was swapped with 'standard', and the terms 'women' and 'we' are also exchanged.

Antonym Replacement We apply Antonym Replacement (AR), a data augmentation technique that replaces specific words or expressions in text with words or expressions with the opposite meaning. This approach helps our model improve its ability to understand and distinguish opposite concepts. The first step involves identifying words within a sentence that can be replaced with their antonyms. The target word is replaced with its antonym. For example, we replace 'with' with the antonym of 'without'.

Paraphrasing We also use paraphrasing (P), which rephrases sentences to add structural diversity. We used the Pegasus-paraphrase model (Zhang et al., 2020a) which shows the state-of-the-art performance in the paraphrasing multiple dataset task. This approach involves rephrasing sentences to create semantically similar but structurally distinct variants. By doing so, we intended to enhance the generalizability and comprehension capabilities of our summarization model. This paraphrased sentence maintains the core meaning of the original but is reconstructed with different vocabulary and grammatical structures, which contributes to a richer environment for our summarization evaluation metric.

Named Entity Replacement We finally utilize Named Entity Replacement (NER Swap or NS), which substitutes named entities with others of the same category to enrich the model's ability to handle factual information. In our study, we implemented a novel approach to data augmentation for the Cochrane dataset's training set by leveraging the d4data/biomedical-ner-all (Raza et al., 2022) model which recognizes 84 bio-medical entities. This pre-trained model identifies and categorizes biomedical entities into various entity group cate-

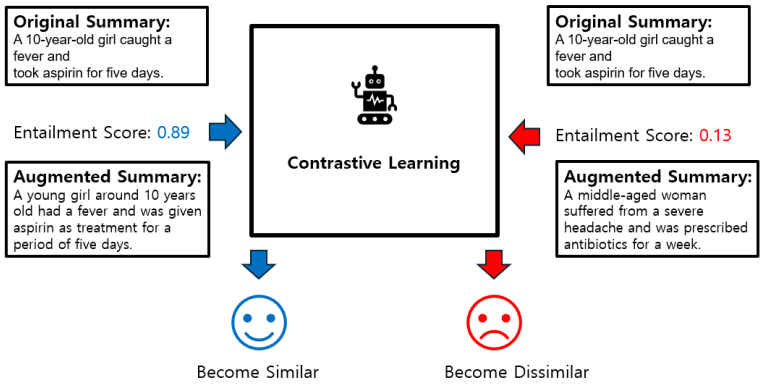

Figure 1: Proposed contrastive learning framework for training medical document summary metric.

Algorithm 1: Flow of NER swap algorithm

Data: Set of documents $\mathbf{D} = \{d_k\}_{k=1}^{T}$
Set of NERs $\mathbf{N} = \{n_{i,k}\}_{i,k=1}^{L,T}$
Result: NER swapped Medical Summaries $\mathbf{S}^*$
$\mathbf{S}^* \leftarrow []$; ; // Initialize the output list
for $k = 1$ **to** T **do**
  for $j = 1$ **to** LEN$(NER_{d_{:,k}})$ **do**
    $u \leftarrow$ Random$(N_{j,T})$; ; // Select a random number u from $N_{j,T}$
    d_k.Replace$(NER_{d_{j,k}}, n_{j,u})$; ; // Replace NER in document
    $\mathbf{N}$.Remove$(n_{j,u})$; ; // Remove used NER
  $\mathbf{S}^*$.Append(d_k); ; // Append modified document to list
return $\mathbf{S}^*$

gories such as *lab value*, *detailed description*, *therapeutic procedure*, *disease disorder*, *medication*, *diagnostic procedure*, and *sign symptom*. Upon analyzing the training dataset, we gathered words belonging to these seven distinct entity groups. The augmentation process involved swapping entities within the same category across different data instances. For instance, if 'headache' was labeled as a Sign symptom in the first training entry and 'nausea' was labeled similarly in the fifth, we swapped these two terms to create a new, augmented training data instance. This method of intra-group entity swapping aims to enrich the dataset by diversifying the context in which each term is used, potentially improving the robustness of models trained on this augmented dataset.The swapping technique was carefully designed to maintain the integrity of the medical context, ensuring that the swapped entities were contextually appropriate. This augmentation strategy not only augments the size of the training data but also introduces a level of variance that can prevent overfitting and enhance the generalization capabilities of downstream models.

Computing Entailment Score By using the augmented dataset, we pair the original summary and the augmented summary. Then, we get the entailment score of the pairs by using the PubMedBERT (Gu et al., 2021) fine-tuned on the MS-MARCO (Bajaj et al., 2018) dataset. This model outperforms other Bert model in biomedical tasks, especially in understanding biomedical words and expressions. The composition of the final dataset before the contrastive learning is an original summary, augmented summary, and entailment score.

3.2 Contrastive Learning

Contrastive learning is a method that extracts features by minimizing the representation distance between similar samples and maximizing the distance between representations of different samples. By employing contrastive learning, we can train the metric without needing specific labels, thereby reducing the effort required from human annotators. We construct a dataset based on the original summary (q_i) and positive augmentation methods, SR and PAR, and negative augmentation methods, RD, RS, AR, and NER.

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^{m} \quad (1)$$

Baselines	**Correlation coef**
ROUGE	-0.010
BERTScore	0.022
Delta-EI	-0.080
STS	0.066
ClaimVer	0.142
NLI	0.053
BioMistral-7B	0.069
LLama3-8B	-0.04
Orca-7B	0.131
Ours	
RoBERTa-base	0.015
+SR	0.08
+SR +RD	0.095
+SR +RD +RS	0.123
+SR +RD +RS +AR	0.376
+SR +RD +RS +AR +NS	**0.418**
+SR +RD +RS +AR +NS +P	0.415
RoBERTa-large	0.204
+SR	0.08
+SR +RD	0.387
+SR +RD +RS	0.480
+SR +RD +RS +AR	0.485
+SR +RD +RS +AR +NS	0.484
+SR +RD +RS +AR +NS +P	**0.519***
Distill-RoBERTa-base	-0.014
+SR	0.119
+SR +RD	0.177
+SR +RD +RS	0.193
+SR +RD +RS +AR	0.339
+SR +RD +RS +AR +NS	0.364
+SR +RD +RS +AR +NS +P	**0.357**

Table 3: Results of correlation coefficients between automated metrics and human evaluation(PIO) across different data augmentation methods

$\mathcal{D}$ is the training data that consists of m instances. We utilize an in-batch negative function for positive samples (p_i^+) and negative samples (p_i^-). During training, each of the two positive samples is included separately, resulting in the creation of its own loss function. The mathematical formulation of the L_{cs} is as follows:

$$L_{cs}(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) \quad (2)$$
$$= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}},$$

where $\text{sim}(q_i, p_i^+)$ is a cosine similarity between q_i and p_i^+. This approach intuitively aligns with the goal of medical document summary evaluation. For positive augmented datasets, our model learns to minimize the distance between their representations, recognizing them as different expressions of the same fundamental information. On the other hand, if a negative augmented summary diverges significantly, by introducing unrelated information or omitting critical details, our model increases the representational distance, highlighting the loss or distortion of information.

4 Experiments

4.1 Implementation Details

For paraphrase data augmentation, we use the Pegasus-paraphrase. We use *PubMedBert* (Gu et al., 2021) fine-tuned on the MS-MARCO dataset as an entailment scoring model. This model is used to calculate the entailment score of the Cochrane dataset (Thakur et al., 2021) and the augmented datasets. In our comprehensive study, we have undertaken a detailed comparison between human evaluation scores and embeddings generated by models for summaries. To achieve this, we utilize 'RoBERTa-base',' DistillRoBERTa - base' (Sanh et al., 2019), and 'RoBERTa-large' (Liu et al., 2019) as our chosen embedding model. This model embeds text to measure the similarity between the original summary and the augmented summary. For the test data, we use the MSLR Cochrane dataset comprised of 597 sets (ground truth summary, model-generated summary) from 6 systems and 8 quality faces of human evaluation scores. In these 8 faces, we use PIO alignment as human evaluation score which stands for population, intervention, and outcome.

4.2 Performance Comparison

The comprehensive experimental results presented in Table 3 underscore significant enhancements in the correlation between automated metrics and human evaluations when leveraging various data augmentation strategies. The experimentation with three different models, RoBERTa-base, RoBERTa-large, and DistilRoBERTa-base, revealed that model scale plays a crucial role in achieving higher performance metrics.Our findings specifically highlighted that the RoBERTa-large model, when subjected to a combination of all six augmentation methods, yielded the highest correlation coefficient of 0.519. This standout performance can be attributed to the larger model's capacity to process and integrate complex variations in training data more effectively, thereby producing assessments that are more aligned with human judgment. Among the individual augmentation techniques, antonym replacement (+AR) was particularly effec-

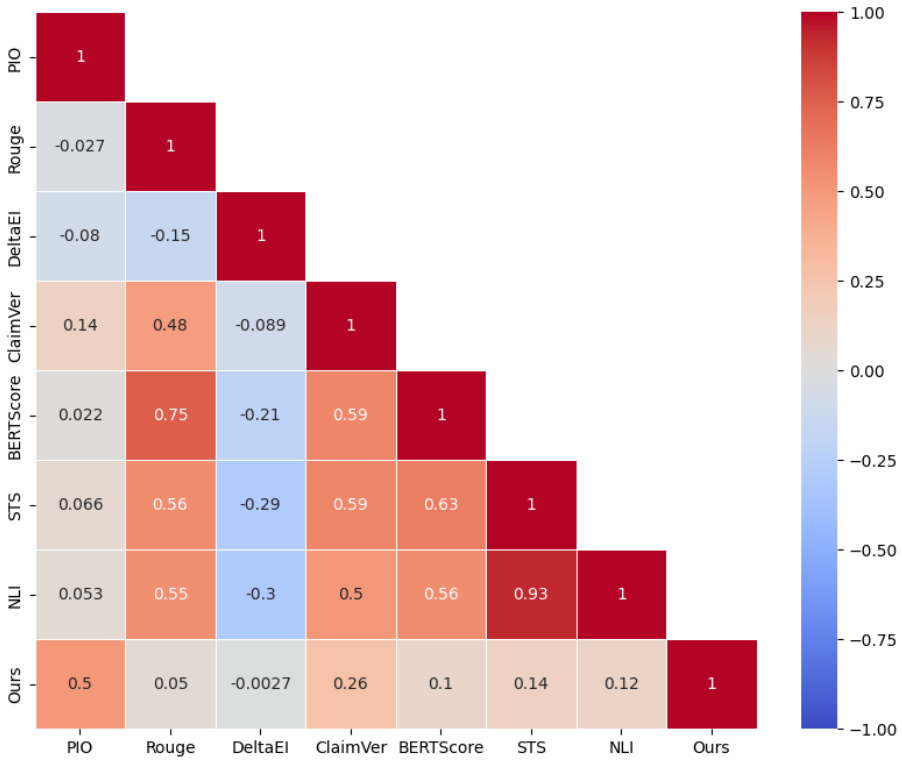

Figure 2: Correlation coefficients between previous metrics and proposed metric

tive, with notable increments in correlation values across all models. This suggests that introducing contrastive contexts into the training data can significantly enhance a model's ability to understand and evaluate nuances more akin to human reasoning. For instance, the RoBERTa-base model demonstrated a marked improvement from a base correlation of 0.015 to 0.376 under the +AR condition, underscoring the profound impact of this method. Furthermore, the progression of improvements from other methods like synonym replacement (+Syn) and random deletion (+RD) also reflects their utility in diversifying the linguistic landscape the models are trained on, although they were less impactful compared to antonym replacement. The correlation coefficient produced by the full-dataset fine-tuned model outperforms the widely used metrics as shown in table 1. These insights collectively point to the efficacy of targeted data augmentation in bridging the gap between automated evaluations and human perceptions, suggesting a path forward for enhancing the reliability and human-likeness of model outputs. The nuanced understanding of text provided by these augmentation methods fosters a deeper comprehension of context, which is essential for models tasked with evaluating text in a manner that resonates with human interpretations.

Correlation between previous metrics and the proposed metric as shown in Figure 2. Human evaluation (PIO) displays weak correlations with metrics except for our proposed metric (Model eval). STS and NLI show a high correlation which just leverages sentence bert. Our metric exhibits weak correlations with most metrics except for DeltaEI, suggesting it captures somewhat similar aspects of evaluation.

4.3 Case Study

As illustrated in Table 4, the document and summary exhibit nearly identical content, indicating a very high correlation between them. However, existing metrics such as ROUGE-1, BERTScore, and NLI report low correlation coefficients between the document and summary, failing to recognize their semantic overlap accurately. In contrast, our proposed metric successfully evaluates the high correlation between the document and the summary. By accurately capturing the degree to which the summary preserves the document's original meaning, our method shows its effectiveness in assessing summary quality. This underscores the superiority of our metric over traditional methods in evaluating medical document summaries, particularly in cases where semantic preservation is critical.

Document	Summary
A 65-year-old male with a history of hypertension and diabetes was admitted to the hospital with chest pain. An ECG revealed ST-segment elevation, and the patient was diagnosed with acute myocardial infarction. He underwent an emergency coronary angioplasty and stenting. The patient was started on aspirin, clopidogrel, and statins.	The patient, a 65-year-old male with hypertension and diabetes, was admitted for chest pain and diagnosed with acute myocardial infarction. He received coronary angioplasty and stenting, and was prescribed aspirin, clopidogrel, and statins.
Metric	**Score**
REMDoC	0.938
ROUGE-1	0.132
BERTScore	0.589
NLI	0.340
LLaMa3-8B	0.320

Table 4: Case study on evaluating the medical document summary using the example from MSLR dataset.

5 Conclusion

This paper proposes a reference-free medical multi-document summary evaluation of the metric via contrastive learning to the pre-trained language models. Experimental results demonstrate that our evaluating model outperforms existing metrics, indicating strong alignment with human evaluations. We expect that our findings contribute significantly to the medical community by facilitating more valuable research.

Acknowledgement

This research was supported by the Chung-Ang University Research Grants in 2023. This research was partly supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program (Chung-Ang University)).

References

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li, and Michael Narag. 2021. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. *Artificial Intelligence in Medicine*, 120:102167.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. *Preprint*, arXiv:1909.11740.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms^2: Multi-document summarization of medical studies. In *EMNLP*.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. Umic: An unreferenced metric for image captioning via contrastive learning. *Preprint*, arXiv:2106.14019.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Stéphane Meystre and Peter J Haug. 2006. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of biomedical informatics*, 39(6):589–599.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777*.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey E. Kuehl, Erin Bransom, and Byron C. Wallace. 2023. Automated metrics for medical multi-document summarization disagree with human evaluations. *Preprint*, arXiv:2305.13693.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *Preprint*, arXiv:1901.11196.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.