OntG-Bart: Ontology-Infused Clinical Abstractive Summarization

Sajad Sotudeh
sajad@ir.cs.georgetown.edu
IR Lab, Georgetown Univeristy
Washington, D.C., USA

Nazli Goharian
nazli@ir.cs.georgetown.edu
IR Lab, Georgetown Univeristy
Washington, D.C., USA

ABSTRACT

Automating the process of clinical text summarization could save clinicians' reading time and reduce their fatigue, acknowledging the necessity of human professionals in the loop. This paper addresses clinical text summarization, aiming to incorporate ontology concept relationships via a Graph Neural Network (GNN) into the summarization process. Specifically, we propose a model, extending BART's encoder-decoder framework with GNN encoder and multi-head attentional layers for decoder, producing ontology-aware summaries. This GNN interacts with the textual encoder, influencing their mutual representations. The model's effectiveness is validated on two real-world radiology datasets. We also present an ablation study to elucidate the impact of varied graph configurations and an error analysis aimed at pinpointing potential areas for future improvements.

CCS CONCEPTS

• **Computing methodologies** → *Information extraction*; *Natural language generation.*

KEYWORDS

text summarization, neural networks, clinical text summarization, abstractive summarization

ACM Reference Format:
Sajad Sotudeh and Nazli Goharian. 2023. OntG-Bart: Ontology-Infused Clinical Abstractive Summarization . In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3573128.3609346

1 INTRODUCTION AND RELATED WORK

Clinical notes summarization is the process of extracting relevant information from patient's clinical notes and generating a summary that captures the most important aspects of the patient's medical history and current condition. The increasing amount of unstructured data in electronic health records (EHRs) such as Radiology reports has led to the need for efficient methods to extract and summarize important information from clinical notes. Radiology reports are a common type of clinical notes, documenting the observations made by radiologists during the imaging process. These reports typically consist of multiple sections, including the FINDINGS and IMPRESSION sections. The FINDINGS section describes the results of the imaging examination in detail. The IMPRESSION provides the overall interpretation of the FINDINGS and is of paramount importance to other practitioners, such as referring physicians and surgeons, as it helps them make informed decisions regarding patient care. Automating the process of generating IMPRESSION given the FINDINGS has become increasingly important and useful in recent years, attracting attention from researchers [6, 7, 12, 16, 23, 24]. In this sense, an automated summarization system would not only save clinicians' time and reduce fatigue [4, 9] but also provide summaries that can be easily reviewed and edited by the practitioners.

Traditional clinical summarization methods relied on rule-based methods, information extraction techniques, and machine learning strategies [14]. Recent literature, however, emphasizes the use of deep neural network techniques for more effective summarization. For instance, Zhang et al. [23] proposed to encode a separate section of the report (i.e., BACKGROUND) to aid the summarization system in the decoding process. Moreover, MacAvaney et al. [12] proposed the recognition of ontological concepts in the FINDINGS text and incorporating them to enhance the decoder's performance. Sotudeh et al. [16] retrieved medical ontologies from FINDINGS, and then incorporated the concept of ontology saliency for improved summarization. Despite these advancements, our proposed method distinguishes itself by extending ontological concepts with their attributes from the knowledge base, a facet unaddressed in prior studies. Also, recent efforts aimed at enhancing the factual accuracy of generated impressions [13, 24, 25] suggest promising avenues for the further refinement of this procedure. In summary, our contributions are twofold:

- Proposing a novel graph-based clinical report summarizer leveraging a graph encoder to reveal pertinent ontology relationships.
- Evaluating the proposed system, conducting an ablation study, and error analysis of the system-generated results, thereby laying the groundwork for future improvements.

2 GRAPH-BASED CLINICAL SUMMARIZATION

Overview. This work examines the integration of external ontological resources into summarization framework. The methodology involves the construction of a graph featuring ontological concepts from radiology reports, enriched by associated attributes retrieved from the ontology resource, all of which are transformed into nodes. These nodes interconnect as edges, representing parent and child concepts, definitions, and synonyms. This graph is then fed into the Graph Neural Network (GNN) [19] module, yielding an encoded representation for each node. An "interaction node" is employed to facilitate the knowledge sharing between the GNN and BART [10]

DocEng '23, August 22–25, 2023, Limerick, Ireland

ACM ISBN 979-8-4007-0027-9/23/08…$15.00
https://doi.org/10.1145/3573128.3609346

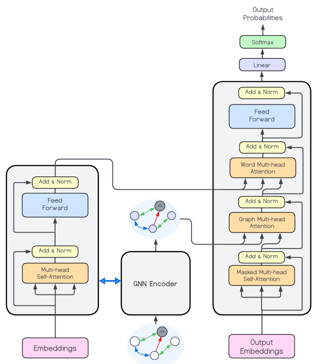

Figure 1: The detailed illustration of our model. As indicated, our model adds two separate modules, namely GNN Encoder [19] and Graph Multi-head Attention layers [18] within the decoder to the standard Transformers network.

encoder. This strategy promotes information exchange between the two modules, leading to updates in node representations that the decoder subsequently applies in hierarchical decoding. The general architecture of our model is shown in Figure 1. We name our model "**Ont**ology **G**raph-based **Bart**", abbreviated as OntG-Bart.

Graph construction. We extract clinical concepts from the RadLex ontology[1] base given the findings text; then we extend each concept with its attributes including: (1) definition, (2) child concepts, (3) parent concept, and (4) synonym(s). We then create a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ for each clinical note, where $\mathcal{G}$ comprises N nodes, each depicting an ontology concept and any related attributes. Graph edges are determined based on five types of relations derived from the retrieved attributes for each concept as follows: *concept-to-concept (C2C):* Nodes are linked by a C2C edge if both represent findings-mentioned concepts. *concept-to-child (C2Child):* A concept node is connected to its ontology-retrieved children nodes via a C2Child edge; *concept-to-parent (C2P):* A C2P edge is established between a concept node and its singular parent node; *concept-to-definition (C2Def):* A C2Def edge links the concept node and its definition node, indicating the relationship; *concept-to-synonym (C2Syn):* A C2Syn edge is set up between the concept node and its synonym nodes, signifying synonymy. We also introduce a unique "Interaction node", linked to concept nodes, to compile critical graph data and affect Bart representations via information exchange in encoder layers. Initial node embeddings also utilize the efficient TransE [1] method for relational information representation.

Ontology-infused encoder. In our summarization encoder, we leverage a modified GreaseLM [22] model to infuse medical ontology knowledge into the summarization encoder. The design facilitates the fusion and information exchange between the summarization encoder and the medical ontology GNN encoder. Our approach integrates the summarization encoder with a graph neural network via an Interaction node and <s> token representation by Bart encoder. This enables the encoder's tokens to gain insights from the knowledge graph, leading to a unified framework that effectively reasons over input text and medical ontology. The Bart encoder takes in the findings word embeddings $\{h_0^{(0)}, h_1^{(0)}, ..., h_J^{(0)}\}$ and then processes it through encoder layers including multi-head self-attention layers and linear layers, to output token representations for each layer:

$$\{h_0^{(\ell)}, h_1^{(\ell)}, ..., h_J^{(\ell)}\} = \texttt{Enc-Layer}(\{h_0^{(\ell-1)}, h_1^{(\ell-1)}, ..., h_J^{(\ell-1)}\}) \quad \text{for} \quad \ell = 1, 2, ..., N \tag{1}$$

$\texttt{Enc-Layer}(\cdot)$ indicates a single layer of Bart encoder, $h_0^{(\ell)}$ is the token representations associated with <s> token (i.e., Bart encoder's interaction token) at layer ℓ. The graph encoder is designed to process the information (i.e., ontology concepts and relations) embedded in the input graph, alongside the textual Transformers encoder that takes in the findings text. To initialize the graph nodes, we first compute the node embeddings $\{e_0^{(0)}, e_1^{(0)}, ..., e_K^{(0)}\}$ for the retrieved ontology concepts using TransE as discussed in Section 2. We then implement a Graph Neural Network (GNN) encoder [19] to process the graph and update the node representations. To be more specific, each layer of the GNN model takes in the current representations of the node embeddings $\{e_0^{(l-1)}, e_1^{(l-1)}, ..., e_K^{(l-1)}\}$ to facilitate the information flow between nodes in the graph, based on the adjacency matrix and output a pre-computed encoding for each node:

$$\{e_0^{(\ell)}, e_1^{(\ell)}, ..., e_K^{(\ell)}\} = \texttt{GNN}(\{e_0^{(\ell-1)}, e_1^{(\ell-1)}, ..., e_K^{(\ell-1)}\}) \quad \text{for} \quad \ell = 1, 2, ..., M \tag{2}$$

$\texttt{GNN}(\cdot)$ is an implementation of Graph Attention Networks (GATs) [19], and $\{e_0^{(\ell)}$ denotes the interaction node representations at the ℓ-th layer. Once the Bart and GNN encoder layers processed the findings and graph's information, we utilize a multi-layer perceptron module to fuse the information through the interaction node's gate and further pass them to the upper layers, $[\tilde{h}_0^{(\ell)}; \tilde{e}_0^{(\ell)}] = \texttt{MLP}([h_0^{(\ell)}, e_0^{(\ell)}])$ where $\tilde{h}_0^{(\ell)}$ and $\tilde{e}_0^{(\ell)}$ are the fused interaction nodes in layer ℓ. We then update the representation of the interaction token and node with the fused representations and then pass them to the next layer. Once the graph node representations are obtained, we employ a hierarchical decoder to attend to the encoded graph and textual representations in order to generate the impression.

Hierarchical ontology decoder. We incorporate another multi-head attention layer ($\text{MAT}^{\mathcal{G}}$) to attend to the graph's node representations, and place it before the Bart's multi-head attention layer (MAT) to enable hierarchical decoding. In this sense, the impression tokens (gold impression in training, and generated impression at inference) are first informed of the knowledge embedded in the graph, and then passed to the upper layers that take care of final impression generation. Specifically, the decoder inputs are processed by a self-attention layer, followed by a multi-head cross-attention layer that attends to the graph encoder outputs (node representations), as follows:

$$\mathbf{MAT}^{\mathcal{G}}(Q, K, V) = \texttt{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \qquad Q' = \mathbf{MAT}^{\mathcal{G}} \cdot V \tag{3}$$

where Q, K, V are linear transformations denoting query, key, and value matrices, respectively. $\text{MAT}^{\mathcal{G}}(\cdot)$ denotes the cross-attention operation on graph $\mathcal{G}$ encoder outputs. The final output of the

[1] http://www.radlex.org/Files/radlex3.10.xlsx

graph cross-attention layer is represented by Q' which is the dot product of the attention weights with the value matrix (i.e., graph encoder outputs). Q' is then passed to the upper multi-head attention layer, shown as MAT$(\cdot)$, that attends to the Bart's encoder outputs (i.e., V') in a similar way as in Equation 3, producing ontology-aware decoder output $\mathbf{y}$. The decoder finally employs a linear projection with Softmax function to sample the next IMPRESSION token sequence from the vocabulary distribution: $\mathbf{y^o} = \mathbf{W_{att}}(\mathbf{MAT}(Q', K', V') \cdot V') + \mathbf{y}, P_{\text{vocab}} = \mathtt{softmax}(\mathbf{W_{out}}\mathbf{y^o})$. We attempted to alter the sequence between **MAT** and $\mathbf{MAT}^{\mathcal{G}}$ in our experiments. However, due to the lack of substantial improvement in performance, we decided not to pursue this approach further.

3 EXPERIMENTAL SETUP

Research questions. The following research questions will be specifically addressed in the experiments:

- **RQ1:** Does incorporating a variety of relationships from medical ontologies into the summarizer result in improved performance in terms of Rouge and BertScore automatic evaluation metrics?
- **RQ2:** How significantly does each relationship, extracted from the ontology, contribute to the performance enhancement of the summarization model?

Datasets and ontology. We used two real-world radiology datasets: MIMIC Chest X-ray (MIMIC-CXR)[8] and one from a large urban hospital, totaling 117,324 (MIMIC-CXR) and 41,066 (urban hospital) reports after preprocessing[12, 23]. We split the data into training, validation, and testing sets in an 8:1:1 ratio. The RadLex ontology, with over 40k radiological terms, was used in our study.

Baselines. We compare our model to a number of extractive and abstractive baselines listed below:

- **LSA & LexRank** [3, 17]: These extractive methods utilize mathematical and graph principles: Singular Value Decomposition (SVD) and graph-based ranking, to evaluate sentence saliency.
- **Pointer-Generator (PG)** [15]: A seq2seq summarizer that copies FINDINGS words directly under high attention, thus integrating the advantages of both extractive and abstractive summarization.
- **Ont. PG** [12]: An abstractive summarizer that leverages ontology concepts (from RadLex), encoded via an auxiliary encoder, to steer the decoding process.
- **OntologyABS** [16]: A PG model extension that integrates significant ontological terms, determined by a content selector, into the seq2seq network during summarization.
- **Bart** [5]: An abstractive summarizer pre-trained via text corruption and reconstruction, improving summarization performance by learning robust text representations.
- **Pegasus** [21]: An advanced abstractive summarizer that develops a summarization-specific pre-training self-supervised objective by masking key sentences in a text and generating an output sequence from the remaining text.
- **GraphRadSum** [7]: The previous SOTA abstractive summarizer in the clinical notes domain that forms a graph from a clinical report, models the graph entities and relationships with a GNN encoder, and optimizes the GNN using contrastive learning loss to emphasize FINDINGS' keywords.

Implementation details. We employed the default hyperparameters for the execution of the baseline models. For our system, we set the batch size to 6 and the learning rate to $3e-5$. The model was trained over 15 epochs, and we selected the best checkpoint based on the highest Rouge-1 score on the validation set.

Table 1: Rouge (RG) and BertScore (BS) results on the test set of the MIMIC-CXR and Urban Hospital datasets. * shows statistically significant improvement (paired t-test, $p < 0.05$) as compared to the best-performing baseline (i.e., Bart).

	MIMIC-CXR				Urban Hospital			
Model	RG-1	RG-2	RG-L	BS	RG-1	RG-2	RG-L	BS
LexRank [3]	14.74	7.15	12.99	85.07	20.84	10.62	18.12	84.30
LSA [17]	18.28	9.30	16.75	85.93	22.85	11.69	19.66	84.90
PG [15]	51.47	39.26	50.22	87.01	37.17	22.36	35.45	85.49
Ont. PG [11]	51.92	39.66	50.89	87.82	38.42	23.29	37.02	87.16
OntologyAbs [16]	53.69	40.92	51.99	88.46	39.01	23.64	37.28	88.14
Pegasus [21]	56.41	44.88	54.01	89.91	47.71	31.87	41.99	89.47
Bart [10]	58.43	45.81	55.32	92.59	49.29	32.78	43.19	90.02
GraphRadSum [7]	57.38	45.12	54.91	92.88	48.98	32.36	42.91	90.29
OntG-Bart	**59.66***	**46.12**	**56.33***	**93.02**	**50.67***	**33.44**	**44.21***	**90.91**

4 EXPERIMENTAL RESULTS

Automatic evaluation. Table. 1 presents the experimental performance of the baselines along with our model in terms of Rouge and BertScore evaluation metrics. Comparing abstractive summarization systems with one another, amongst the traditional seq2seq neural summarizers, the OntologyABS model outperforms the PG network, but as expected it falls behind the pre-trained contextualized summarizers. Among the pre-trained Transformers-based abstractive summarizers, the Pegasus model underperforms the Bart system. This is likely due to its reliance on an extractive objective (i.e., Gap Sentence Prediction), which is shown to be less effective in clinical notes domain [12, 16, 23]. The Bart summarization model has the highest score amongst the baselines, however, it falls behind OntG-Bart which outperforms all methods, achieving statistically significantly improved scores on Rouge-1 and Rouge-L metrics. Comparing Bart and GraphRadSum models whose performances are similar to each other in some metrics (with Bart being slightly better), we notice that Bart performs better at IMPRESSION generation process (i.e., decoding) due to its strong pre-trained decoder. This is not the case in GraphRadSum which trains the decoder from scratch. Overall, our model gains relative improvements of 2.2%, 2.7% (Rouge-1), 0.8%, 2.0% (Rouge-2), and 1.8%, 2.3% (Rouge-L), 1.0% (BertScore) on MIMIC-CXR and Urban datasets, respectively, as compared to the Bart baseline, addressing our first research question, indicating the positive impact of our approach.

Ablation study. The results of the ablation study (Table 2) show a minor effect on performance when excluding concept2concept and concept2child relations. This is likely due to the fact that the Bart system has already encoded the relationships between concepts through its self-attention mechanism. Additionally, it may be difficult to identify relevant children nodes for a specific finding. Adding relevancy scores of children nodes to the report could address this latter issue [2, 20]. Excluding concept-to-definition and concept-to-synonym relationships significantly reduced BertScore, showing their importance in semantic space understanding. Pre-training the model with these relationships could potentially enhance performance. Changes in model complexity are negligible across ablation configurations due to the dominance of summarization-specific layers, with each relationship addition contributing around 20k trainable parameters, a minuscule fraction of the total.

Table 2: Ablation study in terms of Rouge and BertScore metrics on the MIMIC-CXR and Urban Hospital datasets.

Ablation Setting					MIMIC-CXR				Urban Hospital			
C2C	C2Child	C2P	C2Def	C2Syn	RG-1	RG-2	RG-L	BS	RG-1	RG-2	RG-L	BS
–	–	–	–	–	58.43	45.81	55.32	92.59	49.29	32.78	43.19	90.02
✓	✓	✓	✓	✓	**59.66***	46.12	56.33*	**93.02**	**50.67***	**33.44**	**44.21***	**90.91**
–	✓	✓	✓	✓	59.51*	**46.18**	56.31*	92.89	50.39*	33.41	44.20*	90.84
✓	–	✓	✓	✓	59.49*	45.89	56.01	92.85	50.31*	33.32	44.05	90.85
✓	✓	–	✓	✓	58.81	45.92	55.87	92.66	50.08	33.06	43.91	90.74
✓	✓	✓	–	✓	59.31	46.05	**56.35***	92.69	50.11	33.09	**44.21***	90.62
✓	✓	✓	✓	–	58.92	45.88	55.82	92.71	50.09	33.12	43.96	90.66
-	-	✓	✓	✓	59.22	46.01	55.99	92.99	50.35	33.34	44.12	90.85
-	✓	✓	-	✓	59.15	45.89	55.91	92.75	50.41	33.39	44.18*	90.59
-	✓	✓	✓	-	59.01	45.81	55.78	92.78	50.28	33.22	44.06	90.61
-	-	✓	-	✓	59.21	45.93	56.06	92.81	50.36	33.09	43.99	90.66
-	-	✓	✓	-	59.05	45.89	55.98	92.86	50.41	33.12	44.08	90.68
-	-	-	✓	✓	58.97	45.78	55.91	93.01	50.21	33.01	43.89	90.81

5 ERROR ANALYSIS

To pinpoint OntG-Bart's limitations, we compared 50 cases generated by our model, the Bart baseline, and the corresponding human-written summary from each dataset (i.e., 100 cases in total). We present three types of the most common errors, listed as follows with corresponding error rates in percentages: **(1) High Informational Volume Graph (rate: 46%)**: when the report graph has a high volume of information, our model focuses more on the concepts and the connections between them, causing misalignment with human-written impressions (e.g., the model generated "*1. ET tube in appropriate position. 2. Nonspecific interstitial prominence in the lower lungs bilaterally.*", while the gold impression is "*ET tube in appropriate position*"). Incorporating ontological word significance [5, 16] could address this; **(2) Factual Inconsistency (rate: 34%):** the model sometimes exhibits factual inconsistencies in underperformed cases due to loss of contextual information when nodes are selected (e.g., the model negated "*pneumonia*" in *"No acute cardiopulmonary process; specifically, no evidence of pneumonia"*, whereas the human-written impression is "*Resolution of the left lower lobe pneumonia*"). Constructing more sophisticated graphs that can better preserve the contextualized information of the concepts could resolve this; **(3) Inclusion of Unmentioned Information (rate: 24%):** the summarizer includes information present in the graph but not mentioned in the gold impression (e.g., the human-written impression, "*No evidence of acute cardiopulmonary process*", while the system generates "*1. No evidence of acute cardiopulmonary process. 2. Hyperexpansion of the lungs compatible with patient's known COPD*").

6 CONCLUSION

Streamlining clinical note summarization through automation can be beneficial for healthcare providers, providing them with a suggested summary. We presented a novel graph-assisted summarizer that uses medical ontologies to enhance shared representations. Our model outperformed existing models on radiology datasets, suggesting its effectiveness. Additionally, we showed the influence of different types of graph relationships on performance and future improvement areas.

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*.

[2] Qianglong Chen, Feng-Lin Li, Guohai Xu, Ming Yan, Ji Zhang, and Yin Zhang. 2022. DictBERT: Dictionary Description Knowledge Enhanced Language Model Pre-training via Contrastive Learning. In *IJCAI*.

[3] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.* 22 (2004).

[4] Adam E. Flanders and Paras Lakhani. 2012. Radiology reporting and communications: a look forward. *Neuroimaging clinics of North America* 22 3 (2012).

[5] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-Up Abstractive Summarization. In *EMNLP*. ACL, Brussels, Belgium.

[6] Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word Graph Guided Summarization for Radiology Findings. *ArXiv* (2021).

[7] Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhuguo Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph Enhanced Contrastive Learning for Radiology Findings Summarization. In *ACL*.

[8] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *ArXiv* abs/1901.07042 (2019).

[9] Mark D. Kovacs, Maximilian Y Cho, Philip F. Burchett, and Michael A. Trambert. 2018. Benefits of Integrated RIS/PACS/Reporting Due to Automatic Population of Templated Reports. *Current problems in diagnostic radiology* 48 1 (2018), 37–39.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

[11] Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In *CLPsych@NAACL-HTL*.

[12] Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. Talati, and Ross W. Filice. 2019. Ontology-Aware Clinical Abstractive Summarization. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019).

[13] Diwakar Mahajan, Ching-Huei Tsou, and Jennifer J. Liang. 2021. IBMResearch at MEDIQA 2021: Toward Improving Factual Correctness of Radiology Report Abstractive Summarization. In *BioNLP*.

[14] Jon D. Patrick, Dung H. M. Nguyen, Yefeng Wang, and Min Li. 2011. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association : JAMIA* 18 5 (2011), 574–9.

[15] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*. Association for Computational Linguistics, Vancouver, Canada.

[16] Sajad Sotudeh, Nazli Goharian, and Ross Filice. 2020. Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization. In *ACL*. Association for Computational Linguistics, Online.

[17] Josef Steinberger and Karel Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *ISIM*.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*.

[19] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.

[20] Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2022. Dict-BERT: Enhancing Language Model Pre-training with Dictionary. Association for Computational Linguistics.

[21] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML*.

[22] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. GreaseLM: Graph REASoning Enhanced Language Models. In *ICLR*.

[23] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.

[24] Yuhao Zhang, Derek Merck, Emily B Tsai, Christopher D. Manning, and C. Langlotz. 2019. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *ACL*.

[25] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Boosting Factual Correctness of Abstractive Summarization with Knowledge Graph. *ArXiv* abs/2003.08612 (2020).