

UETCorn at MEDIQA-Sum 2023: Template-based Summarization for Clinical Note Generation from Doctor-Patient Conversation

Duy-Cat Can¹, Quoc-An Nguyen^{1,†}, Binh-Nguyen Nguyen^{1,†}, Minh-Quang Nguyen^{1,†}, Khanh-Vinh Nguyen^{1,†}, Trung-Hieu Do² and Hoang-Quynh Le¹

¹University of Engineering and Technology, VNU Hanoi, 144 Xuan Thuy St, Cau Giay, Hanoi, Vietnam

²Hanoi Medical University, 1 Ton That Tung, Kim Lien, Dong Da, Hanoi, Vietnam

Abstract

Clinical note summarization is a challenging task that aims to extract the most relevant information from unstructured text documents and present it concisely and coherently. Large language models, such as OpenAI's GPT-3.5, have limitations in terms of trustworthiness and cost-effectiveness for generating complete clinical notes from patient-doctor dialogues. In this paper, we propose a novel template-based approach for clinical note summarization from dialogue. Our model is under the control of a specific template, which is constructed by our own team and relies on expert validation. The semantic-based partition module fills in the template with key facts extracted from the dialogue using a combination of rule-based and neural methods. The template-based summarization module fine-tunes state-of-the-art (SOTA) BART models and other strategies to generate fluent and informative summaries corresponding to parts in the template. We evaluated our approach on released datasets of MEDIQA-Sum 2023 Shared Tasks, which contain clinical notes from dialogue for various problems. Our approach achieves the best ROUGE-2 and ROUGE-L scores for full note summarization, outperforming several strong baselines.

Keywords

Template-based summarization, Generative Transformer, Semantic-based partition, Clinical Note Generation, Doctor-Patient Conversation

1. Introduction

Clinical summarization involves gathering, organizing, and presenting patient data to support clinical tasks. It can save time, improve clinical accuracy, and mitigate errors by assisting clinicians in collecting, distilling, and interpreting patient information. However, the challenge lies in dealing with the fragmentation and diversity of medical information from various sources and systems. To address this challenge, natural language processing (NLP) techniques

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

[†]These authors contributed equally.

✉ catcd@vnu.edu.vn (D. Can); annq@vnu.edu.vn (Q. Nguyen); 21020526@vnu.edu.vn (B. Nguyen);

19020405@vnu.edu.vn (M. Nguyen); 22025025@vnu.edu.vn (K. Nguyen);

dotrunghieu05220161@daihocyhanoi.edu.vn (T. Do); lhquynh@vnu.edu.vn (H. Le)


🌐 <http://uet.vnu.edu.vn/~catcd/> (D. Can); <https://uet.vnu.edu.vn/~lhquynh/> (H. Le)

🆔 0000-0002-6861-2893 (D. Can); 0000-0002-0605-5841 (Q. Nguyen); 0009-0004-9978-4593 (B. Nguyen);

0009-0006-5080-0702 (M. Nguyen); 0009-0003-0916-9208 (K. Nguyen); 0000-0002-1778-0600 (H. Le)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

can be used to automatically generate clinical summaries from electronic health records. Text summarization, which can be extractive or abstractive, is a key component of NLP for generating concise summaries [1].

In the context of clinical summarization, NLP can be applied to generate summaries of clinical encounters based on prior notes in EHRs [2]. These summaries provide an overview of a patient’s condition, diagnosis, treatment plan, and other relevant information for current or future care providers. Extractive methods select the most relevant sentences or phrases from prior notes, while abstractive methods generate new sentences or phrases that restate or paraphrase the original information [3]. Leveraging NLP techniques for clinical summarization has the potential to streamline the process and enhance the accessibility of critical patient information for healthcare professionals [4].

Moreover, as addressed by Mallen et al. [5], Large Language Models, though achieved SOTA results on multiple NLP tasks, are not completely trustworthy and efficient at all. They are uncontrollable and technically hardly able to be deployed offline. Instead, multi-component frameworks, which mostly follow a hybrid approach, are a more reasonable approach. In this work, we introduce a novel template-based approach designed to efficiently handle the problem of clinical note summarization. Our models have achieved promising results in the MEDIQA-Sum 2023 Shared Tasks [6].

2. Related Work

Abstractive summarization is a challenging task in natural language processing. Sequence-to-sequence learning combined is adopted as the backbone architecture for solving this task [7]. BART is an encoder-decoder model that is applicable to an immense range of end tasks including summarization, which combines a bidirectional encoder and an auto-regressive decoder [8]. With training T5, a portion of the input text is randomly masked, and the model is trained to predict the masked tokens based on the surrounding context [9]. PEGASUS is a pre-training large Transformer-based encoder-decoder model on massive text corpora with a new self-supervised objective [10]. SimCLS helps connect the learning objective and evaluation metrics in sequence-to-sequence learning by treating text generation as a reference-free evaluation problem and using contrastive learning [11]. Biomedical text summarization is a specific field in text summarization and has been spurred significantly [12]. Graph-based, word sense disambiguation (WSD) is used to tackle biomedical summarization. Moreover, the ontology method is applied to enrich the result of the summary [13].

Recent research has witnessed a surge of interest in the exploration of dialogue systems, especially in the biomedical domain. Recurrent neural networks (RNN) and Transformer-based sequence-to-sequence architectures are incorporated for summarizing medical conversations [14]. A variation of the Pointer Generator network is proposed to introduce a penalty on the generator distribution and capture important properties of medical conversations from standardized medical ontologies [15]. A multistage approach is leveraged to tackle the task by learning two fine-tuned models: one to summarize parts of the conversation and another to rewrite the collection of partial summaries into a complete summary [16].

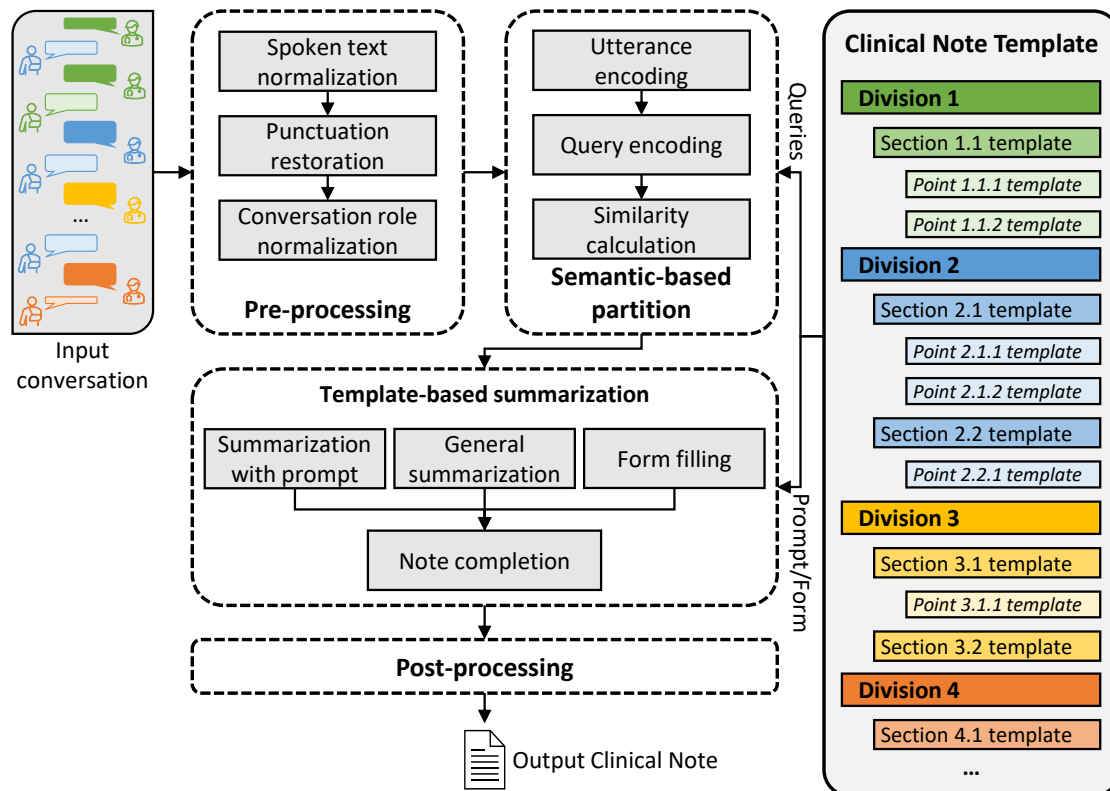


Figure 1: Overview of proposed model for Clinical Note Generation from Doctor-Patient Conversation.

3. Proposed Model

3.1. Overview of proposed model

The overall architecture of the proposed model is illustrated in Figure 1. Our proposed model contains four modules:

1. **Pre-processing:** This module does cleaning noises from the raw input dialogues, and outputs the cleaned ones.
2. **Semantic-based partition:** The cleaned dialogue is broken into smaller units and those which are supportive for a certain query set are gathered into a text. This text (namely extractive text) is significantly shorter than the whole dialogue and conveys only relevant information to queries.
3. **Template-based summarization:** We apply a specific strategy for individual part of a form. Each method receives corresponding extractive text and metadata (prompt, question) as input, then outputs a summary. The template decides how these summaries should be combined.
4. **Post-processing:** Finally, we design a module to fine-grain the output, which can output the format of a standard clinical note.

3.2. Template preparation

Table 1

Popular sections/points and their number of occurrences in the training set and validation set.

Division	Section/Point	Train	Validation	Total
Assessment and plan	ASSESSMENT AND PLAN	34	8	42
	PLAN	32	12	44
	INSTRUCTIONS	14	29	43
Objective results	RESULTS	52	18	70
Objective exam	PHYSICAL EXAM	44	14	58
	PHYSICAL EXAMINATION	16	3	19
Subjective exam	CHIEF COMPLAINT	59	17	76
	HISTORY OF PRESENT ILLNESS	45	13	58
	REVIEW OF SYSTEMS	50	15	65

By analyzing statistics from training and validation datasets and incorporating insights from medical professionals, we have developed a general clinical note template.

The template consists of four divisions, namely Subjective, Objective exam, Objective results, Assessment and plan. Each of these divisions include sections and their subpoints in clinical notes. We analyzed the number of occurrences of each section and point in the training set and validation set, part of which is shown in Table 1. We eliminated points with low occurrence frequency from the template. The template provides the following information for each point.

- **Semantic-based partition information.** We include queries, keywords, similarity score threshold and other partition configurations for each point.
- **Summarization information.** We include summarization strategy, output length threshold and model options. We include prefix and postfix templates that were derived by examining the format of clinical notes. Furthermore, we provide different postprocessing templates for each points.

3.3. Pre-processing

The input dialogue contains several errors from a standard speech-to-text process. We aim to address three following problems from the raw input:

- **Spoken Noise:** Different from written language input, dialogue contains more meaningless words and noises such as hesitation words (“um”, “hmm”, “uh”) and repeated words. These noises limit the capability of the large language model trained on cleaner data. We defined a set of replacements for such words and replace them in the raw dialogue.
- **Missing Punctuation:** 32 over 87 samples in the train and validation datasets do not contain punctuation. Punctuation plays an important role in downstream problems: chunking sentences, identifying semantic structures, understanding semantic representation (in Transformer-based models), etc. Following the work of Guhr et al. [17], we adopt a BERT model to restore the punctuation of the spoken-cleaned dialogues.

- **Wrong-Role Dialogue:** 4 over 87 dialogues in the train and validation datasets contain wrong-role utterances. This error leads to serious problems in the semantic capturing capability of the summarization models. To overcome this problem, we defined rules to detect the true role of speak in each utterance in the punctuation-restored dialogues and then align wrong-role dialogues.

3.4. Semantic-based partition

For each point (or section if it does not have any subpoint) in the template, we shrink the context volume to the meaningful threshold. A query set is pre-defined for each point in the template limiting the scope of meaningful information. The overall flow follows: firstly, dialogue is divided into smaller units (utterances and sentences); these units are then encoded into vector-based representations; and lastly, calculate the similarity score to drop the non-relevant units (lower than a certain threshold).

Utterance and Query Encoding Sentence-BERT (SBERT) [18] proposed a robust framework of trustworthy sentence-level and paragraph-level embedding based on BERT models. In our work, we adopt a State-Of-The-Art SBERT model trained on millions of English query-answer pairs and normal similar sentence pairs. This model embeds given queries and sentences in the dialogue into vectors of 384 dimensions.

Similarity Calculation Having embedding vectors set \mathbf{Q} for queries and \mathbf{U} for sentences in utterance, we calculate a matching score $sc(\mathbf{U}, \mathbf{Q})$ for each utterance by the following formula:

$$sc(\mathbf{U}, \mathbf{Q}) = \max_{(q,u) \in \mathbf{Q} \times \mathbf{U}} \left(\frac{\mathbf{u} \cdot \mathbf{q}}{\|\mathbf{u}\| \|\mathbf{q}\|} \right)$$

where $u, q \in \mathbb{R}^{384}$ are corresponding SBERT embedding vectors for each query in the predefined query set and meaningful sentences in an utterance, respectively. The returned value of $sc(\mathbf{U}, \mathbf{Q})$ is a scalar score, which determines the “support” of the current utterance to the given query set. We drop the non-relative utterances based on a certain threshold.

Alternatively, during analyzing the template, we find out some “signature words” in the dialogue that takes a strong effect on the appearance of valuable information. To improve the coverage, we implemented a strict matching method over sentence level. Given a query set, the output of this strict matching method is a text whose sentences contain any keyword in the set.

3.5. Template-based summarization

We propose three different strategies for this summarization method: General summarization, Summarization with prompt, and Form filling.

General summarization BART [8] proposed a combined scenario of two advanced architecture in Language Understanding task (BERT [19]) and Language Generation task (GPT [20]). We adopt large pre-finetuned BART models on several summarization datasets: SAMSUM [21],

CNN/DM [22], and XSum [23] dataset. During model inference, we combined these BART models with BEAM search [24] to generate the final sequence and used corresponding tokenizers to revert back to textual format.

Summarization with prompt GPT [20] showed promising performance on various text generation problems by controlling generation direction via a beginning prompt. Inspired by this idea, we also proposed the same scenario of controlling the BART decoder generation direction via a prompt, which is separated from the input. The other components are shared with the general summarization method.

Form filling Besides the need for abstractive summarization, certain points (or sections) require highly accurate information from the speakers. For such section, we define two other methods which rely almost on extractive approaches:

- **Answer Extracting:** We employ a DistilBERT model [25] trained on SQuAD dataset [26] for this task. The model receives input of a predefined question tailed with the context (the output from the semantic-partition module) and returns text spans indicating the answer for a given question in the context.
- **Shallow Abstractive:** We convert the output from the semantic-partition module into a more valued form following methods: cleaning non-written words, removing out-of-scope sentences via a similarity matrix and a relevance threshold, and aligning pronouns. These methods aim at a lossless information representation.

Note completion Having corresponding summaries for each point (or section), we format them in the order of the prepared form. The output is a raw clinical note.

3.6. Post-processing

To improve the generated output and mimic the writing style of real clinical notes, several adjustments are made. First, health indexes are modified by adding metric units and rewriting them in abbreviation form. Disease names are standardized using terminologies commonly seen in clinical notes. Verbal numbers are converted into numerical notation. The pronoun “you” is replaced with “the patient” in relevant sentences. Lastly, certain non-medical questions are removed from the output.

4. Experiments and Results

4.1. Dataset and metrics

We used a dataset from Subtask C named Full-Encounter Dialogue2Note Summarization in MEDIQA-Sum 2023 Shared Tasks [6]. The dataset includes 67 samples, 20 samples and 40 samples in the training set, validation set and test set respectively. Each sample contains a full encounter conversation and a clinical note (except for the test set) which summarises the related conversations. The encounter conversations are expressed in text, including the clinical roles

and contents. The clinical notes are structured as a template comprising four divisions, namely Subjective, Objective Exam, Objective Results, and Assessment and Plan.

Recall-Oriented Understudy for Gisting Evaluation score (ROUGE score) is used to evaluate our model [27]. The metric compares the generated note to a golden note, the higher scores show the closer relations between them. Some ROUGE scores such as ROUGE-1, ROUGE-2, and ROUGE-L are used to evaluate the automatic summarization model in this paper, where the ROUGE-1 F1 score are main evaluation scores.

4.2. Implementation

We implement the proposed model on Python 3.8.5 environment, torch 1.13.1 and transformers 4.26.0. We conduct our runs on a local machine of 8x Nvidia RTX 2080 Ti (11Gb CUDA each). In terms of mentioned pre-trained models, we adopt models under the public-sharing policy of Huggingface Model Hub, which are:

- SBERT model `sentence-transformers/all-MiniLM-L6-v2` for utterance and query encoding.
- DistilBERT model `distilbert-base-cased-distilled-squad` for answer extracting.
- BART models `philschmid/bart-large-cnn-samsum`, `lidiya/bart-large-xsum-samsum`, and `amagzari/bart-large-xsum-finetuned-samsum-v2` for summarization tasks (both general summarization and summarization with prompt)

4.3. Results

This section presents three different settings of our runs, namely three different versions of our template. We define a baseline template which conveys standard parts of a clinical note and relies majorly on the effectiveness of abstractive summarization models (BARTs). In the second version, we add more extractive parts in our template to handle the irrelevant information while texts are abstracted in neural models. Finally, we introduce the final fine-tuned version, which majorly inherits ideas of the two previous versions while having more complete prompts, queries, and questions.

Table 2 presents a summary of the performance results for our models and comparative models in Task C full note on the test dataset. Our template-based approaches achieves the highest scores on ROUGE-2 metric of **0.2331**, ROUGE-L metric of **0.2481**, and ROUGE-L-sum metric of **0.4653**. In comparison with the leading team on ROUGE-1 metric, our fine-tuning settings are slightly under **0.0022** while remaining outstandingly higher on all other metrics.

Table 3 summarizes the performance results of our models and the comparative models for each division of Task C on the test dataset in ROUGE-1 metric. Our approaches achieve the highest values in “*Objective exam*” and “*Assessment and plan*” divisions, while showing a promising result on average. Noticeably, “*Assessment and plan*” is the division that relies on only shallow abstractive methods in all our settings.

Table 2

The official results of Task C full note on test dataset.

Team		ROUGE 1	ROUGE 2	ROUGE L	ROUGE L sum
PULSAR	Run 1	0.2764	0.0979	0.1624	0.2362
	Run 2	0.2941	0.1160	0.1918	0.2608
HuskyScribe	Run 1	0.4697	0.1931	0.2228	0.4260
	Run 2	0.3184	0.1505	0.1903	0.2975
Tredence	Run 1	0.4863	0.1920	0.2361	0.4402
	Run 2	0.4998	0.2035	0.2430	0.4506
Baseline		0.4850	0.2096	0.2481	0.4545
+ fine-tuned template		0.4976	0.2331	0.2467	0.4653
+ extractive points		0.4971	0.2310	0.2434	0.4647

*The highest results for each metric are highlighted in bold.***Table 3**

The official results of Task C division on test dataset.

Team		Division				
		Subjective	Objective exam	Objective results	Assessment and plan	Average
PULSAR	Run 1	0.1788	0.1892	0.4393	0.1807	0.2470
	Run 2	0.4125	0.1892	0.4393	0.1807	0.3054
HuskyScribe	Run 1	0.4758	0.4177	0.3668	0.3932	0.4133
	Run 2	0.4683	0.4289	0.3633	0.3230	0.3959
Tredence	Run 1	0.5141	0.4045	0.4746	0.4285	0.4554
	Run 2	0.5107	0.3939	0.4765	0.4361	0.4543
Our baseline model		0.5031	0.4374	0.3323	0.4896	0.4406
+ fine-tuned template		0.4842	0.4346	0.3575	0.4970	0.4433
+ extractive points		0.4843	0.4384	0.3575	0.4970	0.4443

The highest results for each division are highlighted in bold.

5. Conclusion

In this paper, we introduce a novel template-based approach designed to efficiently handle the problem of clinical note summary from dialogue. Our approach relies on a multi-component framework: a pre-processing module, a semantic-based partition module, and a templates-based summarizing module. We evaluated our approach on released datasets of MEDIQA-Sum 2023 Shared Tasks, which contain clinical notes from dialogue for various problems. We propose a baseline template which conveys standard parts of a clinical note and relies majorly on the effectiveness of abstractive summarization models for pretraining seq2seq models.

Acknowledgments

Quoc-An Nguyen was funded by the Master, Ph.D. Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2022.ThS.001.

References

- [1] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al., Abstractive text summarization using sequence-to-sequence rnns and beyond, arXiv preprint arXiv:1602.06023 (2016).
- [2] K. Wilhelm, A. Finch, B. Kotze, K. Arnold, G. McDonald, P. Sternhell, B. Hudson, The green card clinic: overview of a brief patient-centred intervention following deliberate self-harm, *Australasian Psychiatry* 15 (2007) 35–41.
- [3] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, Z. Li, Towards unified surgical skill assessment, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9522–9531.
- [4] J. C. Feblowitz, A. Wright, H. Singh, L. Samal, D. F. Sittig, Summarization of clinical information: a conceptual model, *Journal of biomedical informatics* 44 (2011) 688–699.
- [5] A. Mallen, A. Asai, V. Zhong, R. Das, H. Hajishirzi, D. Khashabi, When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories, arXiv preprint arXiv:2212.10511 (2022).
- [6] W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, Overview of the mediqua-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations, in: *CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023*.
- [7] L. Hou, P. Hu, C. Bei, Abstractive document summarization via neural model with joint attention, in: *National CCF conference on natural language processing and Chinese computing*, Springer, 2017, pp. 329–338.
- [8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.
- [9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *The Journal of Machine Learning Research* 21 (2020) 5485–5551.
- [10] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 11328–11339.
- [11] Y. Liu, P. Liu, Simcls: A simple framework for contrastive learning of abstractive summarization, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 1065–1072.
- [12] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, et al., Review of automatic text summarization techniques & methods, *Journal of King Saud University-Computer and Information Sciences* 34 (2022) 1029–1046.
- [13] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, G. Del Fiol, Text summarization in the biomedical domain: a systematic review of recent research, *Journal of biomedical informatics* 52 (2014) 457–467.
- [14] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, et al., Generating medical reports from patient-doctor conversations using

- sequence-to-sequence models, in: Proceedings of the first workshop on natural language processing for medical conversations, 2020, pp. 22–30.
- [15] A. Joshi, N. Katariya, X. Amatriain, A. Kannan, Dr. summarize: Global summarization of medical dialogue by exploiting local structures., in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 3755–3763.
- [16] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, M. R. Gormley, Leveraging pretrained models for automatic summarization of doctor-patient conversations, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 3693–3712.
- [17] O. Guhr, A.-K. Schumann, F. Bahrmann, H. J. Böhme, Fullstop: Multilingual deep models for punctuation prediction (2021). URL: http://ceur-ws.org/Vol-2957/sepp_paper4.pdf.
- [18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: 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, pp. 3982–3992.
- [19] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
- [20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).
- [21] B. Gliwa, I. Mochol, M. Biesek, A. Wawer, SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 70–79. URL: <https://www.aclweb.org/anthology/D19-5409>. doi:10.18653/v1/D19-5409.
- [22] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: NIPS, 2015, pp. 1693–1701. URL: <http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend>.
- [23] S. Narayan, S. B. Cohen, M. Lapata, Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, ArXiv abs/1808.08745 (2018).
- [24] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strophe, R. Kurzweil, Generating high-quality and informative conversation responses with sequence-to-sequence models, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2210–2219.
- [25] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, in: NeurIPS EMC2 Workshop, 2019.
- [26] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv e-prints (2016) arXiv:1606.05250. arXiv:1606.05250.
- [27] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.