Using BART to Automatically Generate Discharge Summaries from Swedish Clinical Text

Nils Berg, Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University, Kista, Sweden
nilsjv.berg@gmail.com, hercules@dsv.su.se

Abstract

Documentation is a regular part of contemporary healthcare practices and one such documentation task is the creation of a discharge summary, which summarizes a care episode. However, to manually write discharge summaries is a time-consuming task, and research has shown that discharge summaries are often lacking quality in various respects. To alleviate this problem, text summarization methods could be applied on text from electronic health records, such as patient notes, to automatically create a discharge summary. Previous research has been conducted on this topic on text in various languages and with various methods, but no such research has been conducted on Swedish text. In this paper, four data sets extracted from a Swedish clinical corpora were used to fine-tune four BART language models to perform the task of summarizing Swedish patient notes into a discharge summary. Out of these models, the best performing model was manually evaluated by a senior, now retired, nurse and clinical coder. The evaluation results show that the best performing model produces discharge summaries of overall low quality. This is possibly due to issues in the data extracted from the Health Bank research infrastructure, which warrants further work on this topic.

Keywords: Patient Discharge Summaries, text summarization, clinical text, Natural Language Processing, Transformer, BART, synthetic text, negative results

1. Introduction

For clinicians in contemporary healthcare, documentation is a regular part of the daily tasks. One documentation task is the writing of discharge summaries. A discharge summary is a document created at the end of a care episode, such as a hospital admission, and documents that care episode (Scarfield et al., 2022). In this way, the discharge summary serves as one of the main tools of communication between secondary and primary care (Unnewehr et al., 2015).

Unfortunately, manually writing discharge summaries is a time-consuming process (Unnewehr et al., 2015), and as a consequence they are not always produced in time (Kripalani et al., 2007; Horwitz et al., 2013). Moreover, even when discharge summaries are made available in a timely manner, they are often of lacking quality in various respects (Kripalani et al., 2007; Unnewehr et al., 2015; Yemm et al., 2014; Callen et al., 2008; O'Leary et al., 2006; Braet et al., 2016).

Text summarization could potentially be applied to automatically summarize the text(s) which make up a hospital care episode, such as patient notes, into a discharge summary.

In recent years, state-of-the-art results have been achieved in text summarization with the use of Transformer-based language models (Alomari et al., 2022). One such model, and one which has achieved state-of-the-art results, is the Bidirectional Auto-Regressive Transformers (BART) model (Alomari et al., 2022), which is the model used in this paper.

2. Related Research

Previous research employing extractive text summarization (ETS) to summarize patient notes into discharge summaries has been conducted on Finnish data using various language independent methods such as distributional semantics and specifically the random indexing method (Moen et al., 2016). Using Chinese data, previous research has employed various neural network based methods for ETS (Xiong et al., 2019).

Previous research using abstractive text summarization (ATS) has been conducted more frequently. Here, the MIMIC-III data set (Johnson et al., 2016) has been frequently explored with various methods, such as using recurrent neural networks (Diaz et al., 2020) and different Transformer-based language models (Hartman and Campion, 2022; Zhu et al., 2023; Pal, 2022). Summarizing data from Japanese electronic health records (EHRs) has also been explored (Ando et al., 2022).

In addition to the research done on this topic with ETS or ATS, research with hybrid text summarization (HTS) has also been performed. Here, the MIMIC-III data set has also been explored using different combinations of recurrent neural networks, Bidirectional Encoder Representations from Transformers (BERT) models, and BART models (Shing

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 246–252

et al., 2021).

3. Knowledge Gap

As described in the *Related research* section, the task of summarizing patient notes into a discharge summary has been explored in various languages in previous research. However, to the best of the authors' knowledge, no research has previously been conducted on summarizing Swedish patient notes into discharge summaries.

4. Methods

4.1. Data

In this study, data from the research infrastructure Swedish Health Record Research Bank[1] (Health Bank), held by the Department of Computer and Systems Sciences (DSV) at Stockholm University (Dalianis et al., 2015), was used for fine-tuning a BART language model. Health Bank covers patient data from over two million patient, extracted from the Karolinska University Hospital, Sweden, between 2006 and 2014. All patient notes used in this research had been automatically de-identified and anonymized.

4.1.1. Data Set Structure

The data from Health Bank used in this paper is the so-called *Stockholm EPR Gastro ICD-10 Pseudo Corpus II* (hereinafter *Corpus II*) data set. [2] (Lamproudis et al., 2023), which consists of 351 730 patient notes, one row per note, in total 65 258 438 tokens, and encompasses 120 929 patients. Of these 351 730 notes, around 79 006 (22.4%) are discharge summaries. Each note is comprised of 6 columns, which are described below:

- `patientnr`, a unique serial number identifier for the patient which this patient note concerns. This identifier has no connection to any real-life identifier for the patient.
- `template_name`, a string identifier for the template used in the system which was used to create this patient note.
- `template_id`, an integer identifier for the template used in the system which was used to create this patient note.
- `recordnote_id`, a unique serial number identifier for this particular patient note.
- `codes`, the ICD-10 code(s) associated with this patient note, such as a diagnosis given to the patient at discharge.
- `full_note`, the free-text note written by the author of this patient note.

[1] Health Bank, http://www.dsv.su.se/healthbank

[2] This research has been approved by the Swedish Ethical Review Authority under permission, Dnr 2022-02386-02

Furthermore, the data set used in this paper did not have any internal structure or relationships, such as grouping of patient notes into care episodes, or any connection between patient notes and their corresponding discharge summary.

4.2. Model Used to Generate Discharge Summaries

As stated in the *Introduction* chapter, this paper makes use of a BART model to generate discharge summaries from patient notes. Specifically, a publicly available BART model[3] pre-trained on around 80 GB of Swedish text and developed by the Swedish National Library was fine-tuned on data from the Health Bank to perform the task of summarizing patient notes into discharge summaries. This model is referred to as *KB-BART* in this paper.

4.3. Data Pre-Processing

As described in section 4.1.1, there were no explicit relationships between patient notes, or between patient notes and discharge summaries. Thus, establishing what patient notes together form a care episode, and what discharge summary is related to that care episode, had to be done before the *KB-BART* model could be fine-tuned.

This task was performed first by sorting all patient notes first by the `patientnr` column, and then by the `recordnote_id` column in order to group patient notes belonging to one patient in chronological order. Then, all patient notes occurring chronologically between two discharge summaries where grouped as a care episode, and paired with the latter discharge summary.

After this pairing had been performed, the `full_note` column of all patient notes in a care episode associated with a discharge summary were concatenated together to form one text containing the entire care episode. Then, pairs where the text of the discharge summary was longer then the text of the care episode were discarded as this signified that the discharge summary did not actually summarize the care episode.

Once the pre-processing described above had been performed, the resulting care episode-discharge summary data set was used for fine-tuning the *KB-BART* model. Thus, the data set resulting from this pre-processing will be referred to as the *fine-tuning set* hereafter in this paper.

4.4. Fine-Tuning of the KB-BART Model

The process of fine-tuning the *KB-BART* model for the task of summarizing patient notes into discharge summaries was done in several steps. First, the fine-tuning set was split into four subsets via

[3] KB-BART, https://huggingface.co/KBLab/bart-base-swedish-cased

different methods of filtering out low quality samples. These subsets are in this paper referred to as *FULL*, *FILT1*, *FILT2* and *METR*, and were created in the following ways:

- The FULL subset was created by including all care episode-discharge summary pairs. Thus, it is identical to the full fine-tuning set.
- The FILT1 subset was created by picking the pairs in the full fine-tuning dataset where at least one term in the care episode also existed in the corresponding discharge summary.
- The FILT2 subset was created was created in the same way as FILT1, but instead of one term needing to be present in both the care episode and the discharge summary, 15% of all terms in the care episode needed to be present in the discharge summary in order to be included in the FILT2 subset.
- The METR (short for "metrics" subset) was created by applying the three metrics *Semantic coherence*, *Topic similarity*, and *Redundancy* (Bommasani and Cardie, 2020). Care episode-discharge summary pairs which fell under certain thresholds in these three metrics were filtered out to create the METR subset, as this indicated low quality samples in text summarization (Bommasani and Cardie, 2020).

Additionaly, for both FILT1 and FILT2, all patient must have been created from more than one patient note.

Each subset was split into a training set, a validation set, and a test set, consisting of 80%, 10%, and 10% of the data respectively. For each of these constructed subsets, a *KB-BART* model, identical for each subset, was fine-tuned based on the training set of that subset. In this way, four different fine-tuned *KB-BART* models were created, in order to see what subset produced the fine-tuned *KB-BART* model with the highest performance.

4.5. Evaluation of the Model Performance

In order to evaluate the performance of the four fine-tuned models, two types of evaluation were used: one automatic evaluation and one manual evaluation.

The automatic evaluation was based on Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), calculated on the test set of the subset which the model was trained on. The average ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-S scores of the discharge summaries generated from the test set by the models were the main measurement of this evaluation. Furthermore, the results of the model which achieved the highest ROUGE scores were then compared against two benchmarks: Oracle and Random, derived from a similar work (Moen et al., 2016).

The Oracle benchmark is based on extractive text summarization and generates summaries by picking the sentences which maximize the ROUGE-2 F-score in regards to an existing summary, until a given length threshold is reached (Moen et al., 2016). This benchmark is not usable in any real-life scenario as it relies on the presence of an existing summary (which defeats the purpose of generating a summary) but it has its' uses as a benchmark to beat as it likely generates summaries with high ROUGE scores.

The Random benchmark is also based on extractive summarization, and generates summaries by randomly combining sentences until a given length threshold is reached (Moen et al., 2016). This makes it a benchmark which any sensible summarization method should be able to outperform.

In this comparison, discharge summaries were generated by the benchmarks from the test set of the best performing model. Then, ROUGE scores were calculated based on these discharge summaries in order to compare them to the average ROUGE scores of the best performing model.

The manual evaluation was performed by a now retired senior nurse and clinical coder, and Swedish native speaker with several years of working experience (in this paper referred to as *the evaluator*). The evaluator manually reviewed ten randomly selected discharge summaries, generated by the model with the highest performance in the automatic evaluation, according to 12 criteria. Out of these 12 criteria, ten originated from a previous work where a similar manual evaluation was performed (Moen et al., 2016). Two criteria (criterion 9 and criterion 10) were added in order to evaluate potential hallucinations present in the generated discharge summaries. The criteria used for the manual evaluation are available in Table 1.

5. Results

5.1. Fine-tuning Set and Subsets

The pairing of care episodes to discharge summaries resulted in a fine-tuning set consisting of 20 345 care episode-discharge summary pairs. From this fine-tuning set, four subsets were derived, consisting of 20 345, 12 494, 2 575, and 7 722 care episode-discharge summary pairs, respectively.

5.2. Evaluation of the Model Performance

In the automatic evaluation, the mean ROUGE scores of the highest performing model was compared against two benchmarks. The results of this evaluation is available in Table 2. Based on these results, the model based on the *FILT2* subset performs best out of all the fine-tuned models. Thus, this model was further evaluated in the manual evaluation.

The implications of the results of the manual evaluation are described in section 6.2.

6. Discussion

6.1. Automatic Evaluation

The results from the automatic evaluation shows that model fine-tuned on the *FILT2* subset outperforms the other models in terms of all measured ROUGE score metrics.
Furthermore, the *FILT2* model outperformed the Random benchmark as well, which is positive as this implies that the *FILT2* model is better than randomness. However, since the margin with which the *FILT2* model outperformed the Random benchmark is not very significant, this implies that this model is not substantially better than randomness.
Moreover, the *FILT2* model was outperformed by the Oracle benchmark in terms of all measured ROUGE score metrics, but even here it should be noted that the margin with which Oracle outperforms the *FILT2* model is not very significant. This implies that the *FILT2* model approaches the upper bounds of what is achievable on the *FILT2* subset test set (Moen et al., 2016).

6.1.1. Comparison to Previous Research

In terms of the ROUGE scores, the results of the *FILT2* model is generally lower than the results reported in similar previous research in almost all cases for all metrics measured in this paper, see Table 3.

6.2. Manual Evaluation

Based on the results from the manual evaluation it can be stated that the *FILT2* model is prone to not include clinically important information in the discharge summaries that it generates when this information is available in the care episode that is being summarized. Based on the manual evaluation, the likelihood of including clinically important information differs depending on the type of information being summarized, with reason for admission and long-term diagnosis being least, and most, likely to be included in a generated discharge summary, respectively.
Furthermore, the discharge summaries generated by the *FILT2* model are also prone to include hallucinations of a severe nature in the discharge summaries that it generates. Moreover, based on the manual evaluation, discharge summaries generated by the *FILT2* model are lacking when it comes to readability, as the flow and overall content of the text was deemed to be very poor by the evaluator.

6.3. Aptitude of Data Set

As previously stated, the data set used in this paper consisted of only six columns, had no explicit grouping of patient notes into care episodes or connection between care patient notes and discharge summaries. Furthermore, basic information such as the when, where, and by whom the patient notes were written was not present in the data set.
As a result, the task of grouping the patient notes in the data set together into care episodes, and then pairing those care episodes together with the correct discharge summaries was largely performed on the basis of assumptions. Thus, there is no guarantee that all, or even a majority of, care episodes in the data set have been correctly established and/or paired with their respective discharge summary.

7. Conclusions

7.1. Performance of the Fine-Tuned Model

In this paper, four instances of a BART model pretrained on Swedish text were fine-tuned on four variations of a data set consisting of care episode-discharge summary pairs written in Swedish, extracted from the Health Bank research infrastructure, for the task of summarizing patient notes into discharge summaries. Based on an automatic evaluation, as well as a manual evaluation performed by a senior nurse and clinical coder, it can be concluded that the best performing fine-tuned BART model resulting from the work in this paper produces discharge summaries with severe shortcomings. Thus, this model is far from ready to be used in any real-life clinical setting.

7.2. Future Work

Since this is the first work to be conducted on the topic of automatically summarizing Swedish patient notes into discharge summaries, there are many possible directions for future work.
Firstly, while some efforts are made in this paper to correctly group patient notes together into care episodes, and then pair these with the correct discharge summary, future work should explore further efforts to more accurately perform this task. This task could either be performed with the current *Corpus II* data set, or aim to extract more data in the form of additional columns from Health Bank in order to alleviate this task.
Secondly, future work should explore the possibilities of using other text summarization methods with the *Corpus II* data set, such as Extractive Text Summarization, ETS or Hybrid Text Summarization, HTS. Perhaps of particular interest is HTS as it has shown good results in previous research on this topic (Shing et al., 2021).
Finally, future work should explore the possibility of replacing the *KB-BART* model used in this work with a similar model pre-trained on Swedish clinical text, rather than the "regular" Swedish text that *KB-BART* is pre-trained on. This is relevant as

ID	Question	Rating
1	Sender	Yes$= 1$, No$= 0$
2	Reason for admission	Yes$= 1$, No$= 0$
3	Long-term diagnosis	Yes$= 1$, No$= 0$
4	Procedures	Yes$= 1$, No$= 0$
5	Tests	Yes$= 1$, No$= 0$
6	Medication	Yes$= 1$, No$= 0$
7	Health status at discharge	Yes$= 1$, No$= 0$
8	Plans for the future	Yes$= 1$, No$= 0$
9	Does the summary contain information that cannot be traced back to the source notes?	Yes$= 1$, No$= 0$
10	If the question above is true, how serious is the incorrect information contained in the summary?	$0.0 - 1.0$ Trivial$= 0.0$, Severe$= 1.0$
11	Readability: how good is the flow of the text?	$0.0 - 1.0$ Bad$= 0.0$, Excellent$= 1.0$
12	Readability: how good is the content of the summary?	$0.0 - 1.0$ Bad$= 0.0$, Excellent$= 1.0$

Table 1: Evaluation criteria for manual evalation. Partially adopted from (Moen et al., 2016, p. 8).

Model	n	R1	R2	RL	RS
FULL	1 491	0.197	0.042	0.099	0.034
FILT1	1 170	0.202	0.041	0.099	0.036
FILT2	227	*0.280*	*0.057*	*0.122*	*0.068*
METR	554	0.195	0.043	0.096	0.033
Oracle	227	**0.300**	**0.090**	**0.128**	**0.074**
Random	227	0.260	0.045	0.110	0.058

Table 2: Performance of the models on respective test set, along with benchmark performance on *FILT2* test set. Best score per metric among fine-tuned models in italic. Best score per metric overall in bold. All values rounded to three decimals. *R1*=ROUGE-1, *R2*=ROUGE-2, *RL*=ROUGE-L, *RS*=ROUGE-S.

Work	R1	R2	RL
Abstractive			
(Diaz et al., 2020)	0.950	0.940	0.950
Hartman and Campion*	0.395	0.105	0.184
(Zhu et al., 2023)**	0.362	0.202	0.358
(Pal, 2022)***	0.383	0.238	0.349
(Ando et al., 2022)	0.153	0.196	0.121
This paper	0.280	0.057	0.122
Extractive			
(Moen et al., 2016)	0.382	0.184	0.367
(Xiong et al., 2019)	-	-	0.629
Hybrid			
(Shing et al., 2021)****	0.524	0.409	0.511

Table 3: Comparison with previous research. * (Hartman and Campion, 2022). Results from the so-called "truncation approach", as this is the most approach most comparable to the approach in this paper. ** Results from the so-called *DISCHARGE* set, as this is the most approach most comparable to the approach in this paper. *** Results from the so-called *Setup 1* set, as this is the most approach most comparable to the approach in this paper, as well as one of the highest performing. **** Averaged results across sections from *RNN+RL ext + BART* model, to give a comparison across all sections of the EHR, as this paper does, for the best performing model.

previous research has shown that this approach can increase performance in downstream tasks (Jerdhaf et al., 2022). In doing this, the potential increase in the fine-tuned model's performance on data from the Health Bank can be explored.

One interesting observation is the easiness to generate clinical language using the *KB-BART* model. This could prove to be an entrance point to generating large amounts of clinical text for use as training data, and should be explored further. However, the ethical issues in regards to the risk of generating text containing personal information in violation of the General Data Protection Regulation (GDPR) must be considered.

The data used for this article is available from Health Bank for academic use after registration by the user. The full work behind this paper is detailed in the first author's master's thesis (Berg, 2023).

8. Bibliographical References

Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. 2022. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech & Language*, 71:101276.

Kenichiro Ando, Mamoru Komachi, Takashi Okumura, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is In-hospital Meta-information Useful for Abstractive Discharge Summary Generation? In *2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 143–148, Tainan, Taiwan. IEEE.

Nils Berg. 2023. Fine-tuning and evaluating a Swedish language model for automatic discharge summary generation from Swedish clinical notes. Master's thesis, Karolinska Institutet, Stockholm University.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic Evaluation of Summarization Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Anja Braet, Caroline Weltens, Luk Bruyneel, and Walter Sermeus. 2016. The quality of transitions from hospital to home: A hospital-based cohort study of patient groups with high and low readmission rates. *International Journal of Care Coordination*, 19(1-2):29–41.

Joanne Callen, Melanie Alderton, and Jean McIntosh. 2008. Evaluation of electronic discharge summaries: A comparison of documentation in electronic and handwritten discharge summaries. *International Journal of Medical Informatics*, 77(9):613–620.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK – A Workbench for Data Science Applications in Healthcare. *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering*, Vol-1381:1–18.

Diana Diaz, Celia Cintas, William Ogallo, and Aisha Walcott-Bryant. 2020. Towards Automatic Generation of Context-Based Abstractive Discharge Summaries for Supporting Transition of Care.

Vince Hartman and Thomas R. Campion. 2022. A Day-to-Day Approach for Automating the Hospital Course Section of the Discharge Summary. *AMIA Annual Symposium proceedings. AMIA Symposium*, 2022:216–225.

Leora I. Horwitz, Grace Y. Jenq, Ursula C. Brewster, Christine Chen, Sandhya Kanade, Peter H. Van Ness, Katy L. B. Araujo, Boback Ziaeian, John P. Moriarty, Robert L. Fogerty, and Harlan M. Krumholz. 2013. Comprehensive quality of discharge summaries at an academic medical center. *Journal of Hospital Medicine*, 8(8):436–443.

Oskar Jerdhaf, Marina Santini, Peter Lundberg, Tomas Bjerner, Yosef Al-Abasse, Arne Jonsson, and Thomas Vakili. 2022. Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records. In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 30–32, Marseille, France. European Language Resources Association.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Sunil Kripalani, Frank LeFevre, Christopher O. Phillips, Mark V. Williams, Preetha Basaviah, and David W. Baker. 2007. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *JAMA*, 297(8):831–841.

Anastasios Lamproudis, Therese Olsen Svenning, Torbjørn Torsvik, Taridzo Chomutare, Andrius Budrionis, Phuong Dinh Ngo, Thomas Vakili, and Hercules Dalianis. 2023. Using a large open clinical corpus for improved ICD-10 diagnosis coding. In *To appear in AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, pages 74–81.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37.

Kevin J. O'Leary, David M. Liebovitz, Joseph Feinglass, David T. Liss, and David W. Baker. 2006. Outpatient physicians' satisfaction with discharge summaries and perceived need for an electronic discharge summary. *Journal of Hospital Medicine*, 1(5):317–320.

Koyena Pal. 2022. Summarization and Generation of Discharge Summary Medical Report, https://cs.brown.edu/media/filer_public/91/33/913389ac-49a0-4056-a886-424499c6e511/palkoyena.pdf, Visit date 2023-07-06.

Phoebe Scarfield, Thomas David Shepherd, Caitriona Stapleton, Alexandra Starks, Ellen Benn, Sara Khalid, Bryony Dayment, Alex Moate, Sandra Mohamed, and Jasmine Lee. 2022. Improving the quality and content of discharge summaries on acute medicine wards: a quality improvement project. *BMJ Open Quality*, 11(2):e001780.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes. ArXiv:2104.13498 [cs].

Markus Unnewehr, Bernhard Schaaf, Rusi Marev, Jason Fitch, and Hendrik Friederichs. 2015. Optimizing the quality of hospital discharge summaries – a systematic review and practical tools. *Postgraduate Medicine*, 127(6):630–639.

Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, and Jun Yan. 2019. A Study on Automatic Generation of Chinese Discharge Summary. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1681–1687, San Diego, CA, USA. IEEE.

Rowan Yemm, Debi Bhattacharya, and David Wright. 2014. What constitutes a high quality discharge summary? A comparison between the views of secondary and primary care doctors. *International Journal of Medical Education*, 5:125–131.

Yunqi Zhu, Xuebing Yang, Yuanyuan Wu, and Wensheng Zhang. 2023. Leveraging Summary Guidance on Medical Report Summarization, Visit date 2023-07-06. https://arxiv.org/abs/2302.04001.

9. Language Resource References

The text set *Stockholm EPR Gastro ICD-10 Pseudo Corpus II* is available from the research infrastructure Swedish Health Record Research Bank[4], at Stockholm University.

[4]Health Bank,http://www.dsv.su.se/healthbank