Automatic Discharge Summary Generation using Neural Network Models

TOKYO METROPOLITAN UNIVERSITY

Kenichiro Ando

Supervisor: Prof. Mamoru Komachi

Department of Computer Science
Graduate School of Systems Design
Tokyo Metropolitan University

This dissertation is submitted for the degree of
Doctor of Philosophy (Computer Science)

February 2023

Acknowledgements

The creation of this thesis was supported by many people. First, I am super grateful to my supervisor, Professor Mamoru Komachi, who teaches me how to conduct a research project from scratch. He kindly watched over my research life and I had the pleasure of learning a lot. I am also grateful to Professor Takashi Okumura for his significant contributions across all my studies. He invited me to this research topic and gave me important ideas. Especially, the medical analysis of my research would not have been possible without him. I also thank Dr. Yuji Matsumoto. He managed the pace of our studies excellently and without him, I would not be able to finish my Ph.D. program. Also, his team meetings were very inspirational and educational for me. I would like to thank Dr. Satoshi Sekine for his advice on another of my research topic, which was not included in this thesis. He continuously provided thoughtful comments and kept me in a comfortable research environment. He accepted me into RIKEN and gave me the opportunity to communicate with various people. In particular, I thank Mr. Koichiro Watanabe, who had the same interest in my research. It was very encouraging to me. My lab members always provided me with interesting research discussions. Thanks to them, my student life was much more enjoyable and painless, and I wish I could have stayed here longer.

Finally, I am especially grateful to all annotators, Ms. Mai Tagusari, Ms. Nobuko Nakagomi, Dr. Hiroko Miyamoto, Dr. Norihiko Inoue, and Mr. Takuaki Tani. Without their steady, high-quality annotations based on their medical knowledge, no study would have been completed. This project's success was due in large part to their patience in responding to my repeated requests. The annotation labels they created will be of great help to us in the future. Many thanks to Dr. Miyamoto and Dr. Inoue for taking the time to discuss the clinical aspects of my study, it was a great learning experience for me.

Abstract

During the patient's hospitalization, a physician must record daily observations of patients and summarizes them into a brief document called "discharge summary" when a patient is discharged. The discharge summaries play a crucial role in patient care, and are used to share information between hospitals and physicians. However, compilation of hospital discharge summaries is an onerous task for physicians, and such paperwork restricts physicians' time to spend with patients and causes burnout. Researchers have begun to apply automatic summarization techniques to address this problem. Because many high-performance summarization techniques have been developed in natural language processing, the generation of discharge summaries can be a promising application of the technology. In particular, automated generation of discharge summary using neural architecture can greatly relieve the physician's burden. Neural network based approaches have achieved remarkable performance in the summarization task, and are also employed in few previous works of discharge summary generation. The discharge summary generation task often has different characteristics than the general domain, such as containing noisy sentences, various meta-information, and many types of source documents. Thus, how to address these characteristics should be an important research topic. However, due to the limited number of studies on discharge summary generation using neural networks, the impact of such issues has not been well studied.

In addition, the generation of discharge summaries using neural networks has not been studied in Japanese because of the lack of a large data set of electronic health records. For the discharge summary generation in Japanese, it should be necessary to research in Japanese rather than English due to problems specific to the Japanese language. And the corpora and trained models created there whold be important assets for future work on

generating discharge summaries in Japanese. To this end, this thesis investigates three topics using a large multi-institutional health records archive in Japan.

First, a study was conducted to find the optimal linguistic granularity of input for extractive summarization using a neural architecture. Previous works in the general domain employed sentences and clauses as the self-contained linguistic unit for extractive summarization, but this study hypothesizes that they cannot cover medical meanings. Thus, the "clinical segment" was developed as a new self-contained linguistic unit for medical documents. It was defined with the support of physicians and annotated by medical professionals. For the verification of the effectiveness of the clinical segment, three types of input to a neural network-based classifier were experimented with: sentence, clause, and clinical segment. The results showed that using the clinical segment achieved the best performance and was the best input granularity for generating discharge summary. The parallel dataset of dummy discharge summaries and inpatient records created here, as well as the trained clinical segment splitter, are also publicly available.

Second, a study was conducted to investigate whether medical meta-information is useful for abstractive summarization of discharge summary. While medical data contain rich meta-information (e.g., disease, length of stay, etc.), it is not known how they affect the discharge summary generation task. Thus, four types of meta-information, hospital, physician, disease, and length of stay, were used in the abstractive summarization experiments. The results showed that adding the disease information into the summarization model increased summarization performance the most. In addition, the disease information also increased the precision of representations of disease and symptom in the outputs. This is the first time to generate discharge summaries in Japanese using an abstractive method with neural networks. The codes of the model used in this study are available.

Finally, a study was conducted to address the question of whether a discharge summary can be generated from only the inpatient records. Previous works of abstractive summarization in other domain have defined factual hallucination as information that cannot be generated from only source documents. Hence, this study is an analysis of factual hallucination in the generation of discharge summaries. The discharge summary generation task is assumed to require a variety of information sources (e.g., referral documents, prescriptions, and physician's memory) that are more than in the general domain. To this end, discharge

summaries were broken down into clinical segments and manually labeled their possible sources of information by medical professionals. The results showed that 39% of the information came from external sources. Compared to statics of dataset from general domain, this study found that the medical dataset was more externally dependent. The most common external source was the patient's past clinical records (17%), and the next most common external source was patient referral documents (7%). Remarkably, the results showed that 4% of the information in a discharge summary came from the physician's memory, suggesting that the information cannot be reconstructed from only the records. In addition, this study developed the "clinical role label" that represents the statement's medical role (e.g., description, action and evaluation) in the discharge summary for deeper analysis. It was defined with the support of physicians and annotated by medical professionals. By investigating the unsourced rate for each clinical role label, this study found that subjective descriptions (e.g., diagnosis and plan), with added the physician's interpretation, were especially externally dependent. The trained clinical role label classifier developed here is publicly available.

This thesis is organized as follows: Chapter 1 describes the motivation of this thesis and a brief explanation of each topic. Chapter 2 provides the related works and evaluation metrics. Chapter 3 details the datasets used across the studies and how they were created. Chapter 4 introduces the "clinical segment", a unit representing medical meaning to break down sentences, which is used in later chapters. Chapter 5 introduces the clinical role label, which assigns the role in medical records to a clinical segment for further in-depth analysis. Chapter 6 explores the optimal granularity of linguistic units to input into the neural extractive model for the discharge summary generation task. Chapter 7 examines the usefulness of medical meta-information for an abstractive summarization method. Chapter 8 analyzes whether discharge summary generation can be achieved using only the inpatient record, and what other types of information sources are needed. Chapter 9 concludes the thesis and describes future work.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Clinical notes are written daily by physicians from their consults and are used for their own decision-making or coordination of treatment. They contain a large amount of important data for machine learning, such as conditions, laboratory tests, diagnoses, procedures, and treatments. While invaluable to physicians and researchers, the paperwork is burdensome for physicians. A recent study found that family physicians spent 5.9 h of their 11.4 h workday on electronic health records (EHRs) [1]. In 2019, 74% of physicians spent more than 10 h per week on paperwork and administration [2]. Another study reported that physicians spent 26.6% of their daily working time on documentation [3]. Discharge summaries, a subset of these, also play a crucial role in patient care, and are used to share information between hospitals and physicians. It is created by the physician as a summary of notes during hospitalization at the time of the patient's discharge, which is known to be very time-consuming.

Artificial intelligence technology has been increasingly applied in various fields of medicine [4–10]. Its application in medical texts is expected to improve the efficiency of paperwork [11–13]. In natural language processing (NLP), various summarization techniques, especially neural network models, have demonstrated high accuracy in summarization benchmarks [14–19]. These technologies can be applied to summarizing inpatient records. Therefore, some studies have been conducted on the automated generation of the whole discharge summary using neural network models [20–25].

The discharge summary generation task often has different characteristics than the general domain, such as containing noisy sentences, various meta-information, and many types

of source documents. Thus, how to address these characteristics should be an important research topic. However, due to the limited number of studies on discharge summary generation using neural networks, the impact of such issues has not been well studied. To this end, this thesis addresses three topics on discharge summary generation using neural network models.

First, a study was conducted to find the optimal linguistic granularity of input for extractive summarization using a neural architecture. Some recent studies of extractive summarization investigated the best granularity units for neural model inputs [26, 27]. However, the granularity of inputs has not been explored for the summarization of medical documents. Thus, we attempted to identify the optimal granularity by defining three units with different granularities and comparing their summarization performance: whole sentences, *clinical segments*, and clauses. The *clinical segment* is our novel concept to express the smallest medically meaningful concepts. It was created based on our hypothesis that current linguistic units cannot cover medical meaning.

Second, a study was conducted to investigate whether medical meta-information is useful for abstractive summarization of discharge summary. previous studies used extractive or abstractive summarization methods, but most of them focused on only progress notes for inputs. Properly summarizing an admission of a patient is a quite complex task, and requires various meta-information such as the patient's age, gender, vital signs, laboratory values and background to specific diseases. Therefore, discharge summary generation needs more medical meta-information, than similar but narrower tasks such as radiology report generation. However, what kind of meta-information is important for summarization has not been investigated, even though it is critical not only for future research on medical summarization, but also for the policy of data collection infrastructure. Thus, four types of meta-information, hospital, physician, disease, and length of stay, were used in the abstractive summarization experiments. This is the first time to generate discharge summaries in Japanese using an abstractive method with neural networks.

Finaly, Most previous studies used only the inpatient record as the source for abstractive summarization task. Therefore, whether artificial intelligence can generate hospital discharge summaries from inpatient records remains an open question. The discharge summary generation task is assumed to require a variety of information sources (e.g., referral

documents, prescriptions, and physician's memory) that are more than in the general domain. Abstractive summarization tasks on other domain have defined factual hallucination as information that cannot be generated from only source documents. Hence, this study is an analysis of factual hallucination in the generation of discharge summaries. We broke down discharge summaries into clinical segments and manually labeled their possible sources of information by medical professionals. If physicians rely on their memory, it would be difficult to automatically generate a discharge summary solely from inpatient records, even with a top-performing summarization technique.

Our contributions are as follows:

- We have created the first publicly available parallel corpus of inpatient records and discharge summaries in Japanese. This will be useful for future research using electronic health records, such as the automatic generation of discharge summaries.
- We developed the clinical segment that breaks down healthcare texts into the smallest medical semantic units and an automatic splitter for this purpose. This is the first time, to the best of our knowledge, that a sentence is split into contextualized subsequences specific to the healthcare domain.
- We developed the clinical role label and its automatic assigner that assigns a clinical meaning to a clinical segment for the automated analysis of large healthcare texts.
- In comparing the sentence, clause, and clinical segment, we found that the best granularity for extractive automatic summarization of healthcare documents is the clinical segment. The results suggest that a unit reflecting medical semantics may be useful for model input.
- We found that medical meta-information is useful for the abstractive summarization of healthcare documents. In particular, a model encoding disease information can generate appropriate disease and symptom words consistent with the source.
- We investigated the origin of the information that appears in the discharge summaries to evaluate the possibility of the automated summarization of inpatient records alone. The analysis illustrates that only 61% of the total information is derived from inpatient records, and 39% of the information originated from sources other than records.

While past medical documents are the most common sources of external information, 11% of the information contained was not derived from any documents, which included speculation and post-discharge plans.

Chapter 2

Background and Related Works

2.1 Automatic summarization

Automatic summarization is a popularly studied topic in NLP [28, 29, 17–19, 16]. The purpose of the summarization task is to yield a short summary of a longer document for quick comprehension. Commonly targets of summarization task include research papers [30], news articles [31, 32], Q&A [33], and so on. An example of a news article summarization is shown in Table 2.1. In general, the summarization task includes a variety of components such as extracting and paraphrasing information, reasoning, and introducing world knowledge.

Automatic summarization has two main approaches: extractive and abstractive summarization. The former method extracts contents from the source text and combines them to generate a summary. The text of the generated summary is wholly derived from the source and does not contain new content. The latter method generates a summary by creating new content based on the source using some algorithms. The algorithms of abstractive method used in earlier works were sentence compression [34], sentence fusion [35, 36], and sentence revision [37], but at present encoder-decoder architectures are commonly used [17–19, 16]. It is a method that maps input text into linear space by the encoder and generates tokens autoregressively by the decoder [38], and is known for its high performance. In the medical field, extractive summarization method was commonly used for knowledge acquisition of clinical features such as diseases, prescriptions, examinations, etc.

Table 2.1 Example of a news summarization task. It is extracted from the CNN / Daily Mail dataset. The left column is the source text and the right is the target text.

Five Americans who were monitored for three weeks at an Omaha, Nebraska, hospital after being exposed to Ebola in West Africa have been released, a Nebraska Medicine spokesman said in an email Wednesday. One of the five had a heart-related issue on Saturday and has been discharged but hasn't left the area, Taylor Wilson wrote. The others have already gone home. They were exposed to Ebola in Sierra Leone in March, but none developed the deadly virus. They are clinicians for Partners in Health, a Boston-based aid group. They all had contact with a colleague who was diagnosed with the disease and is being treated at the National Institutes of Health in Bethesda, Maryland. As of Monday, that health care worker is in fair condition. The Centers for Disease Control and Prevention in Atlanta has said the last of 17 patients who were being monitored are expected to be released by Thursday. More than 10,000 people have died in a West African epidemic of Ebola that dates to December 2013, according to the World Health Organization. Almost all the deaths have been in Guinea, Liberia and Sierra Leone. Ebola is spread by direct contact with the bodily fluids of an infected person.	17 Americans were exposed to the Ebola virus while in Sierra Leone in March . Another person was diagnosed with the disease and taken to hospital in Maryland . National Institutes of Health says the patient is in fair condition after weeks of treatment.

2.2 Automatic summarization of medical text

Summarization on the medical domain can be roughly categorized by the type of input: structured and unstructured text. The tasks using unstructured text are generally noisier and more difficult than those using structured text. Most previous NLP for unstructured medical text has focused on normalization and prediction such as ICD codes, mortality rates, and readmission risks [39–45], which are also called visualization and summarization. They

target to extract or predict important information from the input, not the summarization we aim for, such as outputting a document. Also, some studies attempt to retrieve important information from EHRs [46–49], such as diseases, examination results, and medications, while they collect fragmented information and do not try to generate contextualized passages. Other studies generated several key sentences from the EHRs to give physicians a quick grasp of the main points [50–53].

Previous studies generating a whole discharge summary mostly used structured data as input [54–56]. Some recent studies try to generate a whole discharge summary from the input of free-form inpatient records, which is the same as ours [20–25]. While some studies employ extractive methods [22–25], in other studies, the encoder-decoder architecture of the neural model was used to generate sentences for abstractive summarization [20–22]. However, the amount of such research is not large.

2.3 Hallucination in abstractive summarization

As abstractive summarization can generate more flexible summaries, it has become a major approach in automatic summarization research [14–19]. However, abstractive summarization may sometimes unintentionally generate unfaithful descriptions, known as **hallucinations**. Summaries with hallucinations are fluent [57], but hallucinations degrade the summary quality. Therefore, they have drawn attention in the field [58–63].

Maynez et al. (2020) classified hallucinations into two types: intrinsic and extrinsic hallucinations [57]. Intrinsic hallucination is a phenomenon in which the concept or term itself is in source documents; its synthesis misrepresents the information in the source and the meaning becomes inconsistent. Extrinsic hallucination is a content that is neither supported nor contradicted by the source, and is caused by source documents with poor information. Therefore, the analyses of extrinsic hallucination in previous works almost equal our investigation of the information sources in discharge summaries. Please note that, in the discharge summaries, complementary statements may be inserted that are not explicitly stated in inpatient records; however, they can be easily inferred from the records by medical professionals. We do not consider this to be a hallucination if the information can be inferred, even though previous studies have defined hallucinations more rigorously.

2.4 Dataset of medical records

For advancing the research on summarization of clinical texts, appropriate language resources are indispensable. In English, public corpora of medical records are available, such as MIMIC-III [64], [65], and [66]. The number of resources available in Japanese is limited. The largest publicly available corpus is the one used for a shared task in an international conference, NTCIR [67]. A non-profit organization for language resources maintains another corpus, GSK2012-D [68]. However, their data volume is small, and their statistics exhibit significant difference from those of large-scale data, as illustrated in Table 3.2. This low-resource situation makes the processing of Japanese medical documents more challenging. Furthermore, that problem is critical for the methods using neural models, as they require a large amount of data.

In addition, the Japanese data set is important because Japanese language processing has the following characteristics. First, Japanese medical texts often contain excessive shortening of sentences and orthographical variants of terms originating from foreign languages. Besides, Japanese requires word segmentation. Most importantly, there is no Japanese parallel corpus of inpatient records and discharge summaries. Therefore, we built a new corpus in this thesis.

2.5 Evaluation metric

Measurement of the summarization quality must be automated to avoid costly manual evaluation. ROUGE [69] has been used as a standardized metric to measure the summarization quality in NLP tasks. Formally, ROUGE-N is an n-gram recall between a candidate summary and the reference summaries. When we have only one reference document, ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{gram_n \in Reference} Count_{match}(gram_n)}{\sum_{gram_n \in Reference} Count(gram_n)}, \tag{2.1}$$

where $Count_{match}(gram_n)$ is the maximum number of n-grams that co-occur in a candidate summary and a reference summary.

When we have several references, ROUGE-L is the longest common subsequence (LCS) score between a candidate summary and the reference summaries. As it can assess word relationships, it is generally considered a more context-aware evaluation measure than ROUGE-N. Specifically, ROUGE-L is computed as follows:

$$Recall_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{Reference_{tokens}}, \tag{2.2}$$

$$Precision_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{Summary_{tokens}}, \tag{2.3}$$

$$\text{ROUGE-L} = \frac{2Recall_{lcs}Precision_{lcs}}{Recall_{lcs} + Precision_{lcs}} \tag{2.4}$$

where u is the number of reference sentences, and $LCS_{\cup}(r_i, C)$ is the LCS score of the union of the longest common subsequences between the reference sentence r_i and C, where C is the sequence of candidate summary sentences. For example, if $r_i = (w_1, w_2, w_3, w_4)$, and C contains two sentences: $c_1 = (w_1, w_2, w_6, w_7)$ and $c_2 = (w_1, w_8, w_4, w_9)$, the longest common subsequence of r_i and c_1 is (w_1, w_2), and the longest common subsequence of r_i and c_2 is (w_1, w_4). The union of the longest common subsequences of r_i, c_1, and c_2 is (w_1, w_2, w_4), and $LCS_{\cup}(r_i, C) = 3/4$.

Chapter 3

Datasets

3.1 NHO data

Investigating issues of discharge summary generation using neural models needs a large-scale data set. In additoin, clinical records can be expressed in various dialects and jargons. Accordingly, a study on a single institution would lead to highly biased results in medical NLP tasks because of local and hospital-specific dialects. To study clinical document summarization, it is necessary to set up a multi-institutional environment to mitigate the potential bias caused by the medical records stored in a single EHR source. For this purpose, we designed an experiment using the largest multi-institutional health records archive in Japan, National Hospital Organization Clinical Data Archives (NCDA) [70]. NCDA is a data archive operated by the National Hospital Organization (NHO), which stores replicated EHR data for 66 national hospitals owned by this organization. Thus, the archive has become a valuable data source for multi-institutional studies that span across the country. The dataset we used for the study, referred to as **NHO data** hereafter, is the anonymized subset of the archive, which includes 24,641 cases collected from five hospitals that belong to the NHO. Each case includes inpatient records and a discharge summary for patients of internal medicine departments. The statistics of the target data are summarized in Table 3.2. As shown, the scale of the NHO data is much larger than that of GSK2012-D and MedNLP, which have been used in previous studies [67]. Accordingly, the results obtained using the NHO dataset are expected to be more general.

3.2 Dummy record

There is no publicly available Japanese parallel corpus of inpatient records and discharge summaries. Therefore, we built a publicly available small corpus of medical records. This corpus was built because annotation over the NHO data was restricted due to privacy concerns. An example of our corpus is shown in Table 3.1. This dataset is a parallel corpus containing 108 inpatient records and their discharge summaries, created by physicians who imagined hypothetical patients. In this paper, it is referred to as **Dummy record**. The statistics of the resulting corpus are given in Table 3.2. With respect to the inpatient records, the corpus is closer to real data than in previous studies, except for the number of sentences in a document. For the discharge summary, there are no publicly available Japanese corpora besides the one we built. Because of the summarization process, the sentences contain more words and characters than the source inpatient records.

3.3 Comparison of NHO data and dummy record

This section describes other characteristics of the two data sets. First, let us discuss the format of the inpatient record used in this thesis. Since inpatient records are created daily, they are inherently multi-document. Therefore, the discharge summary generation task could be seen as a multi-document summarization. However, they are all the same nature documents; thus, we address this task as a single-document summarization. For our experiments, we created the inpatient records by concatenating multiple inpatient records over the length of stay into a single document, which filters out descriptions that overlapped with earlier days' records. The inpatient record and the discharge summary always have a one-to-one correspondence, both in the NHO data and in the dummy record. This method may look strange, but it has been used in prior studies [20, 71], and although it is a naive approach, it fits as a verification tool for the later steps. The possibility of more advanced approaches are discussed in Chapter 9.

Next, we discuss copy-and-paste rates in the gold discharge summaries. We investigated the percentage of sentences in the discharge summary that were copy-and-pasted from the inpatient record. As a result, the copy-and-paste rate for the NHO data was 32% and 20% for the dummy record. The rates are similar, although the NHO data has a higher rate, and it is

Table 3.1 Example of a discharge summary. The left column shows the original Japanese texts, and the right column shows corresponding English translations.

#1 細菌性髄膜炎 4/20〜5/8 VCM 1250mg(q12h) 4/20 SBT/ABPC 1.5g 単回 4/20〜 MEPM 2g(q8h) 4/20〜4/23 デキサート 6.6mg(q6h) 4/20〜4/22 日赤ポログロビン 4/20 腰椎穿刺 1 回目 髄液 糖定量 30 mg/dl(血中糖 95mg/dl) 細胞数 2475/μl.	#1 Bacterial meningitis 4/20-5/8 VCM 1250mg (q12h) 4/20 SBT/ABPC 1.5g single dose 4/20- MEPM 2g (q8h) 4/20-4/23 Dexate 6.6mg (q6h) 4/20-4/22 Nisseki polyglobin 4/20 1st lumbar puncture, cerebrospinal fluid glucose level 30 mg/dl (blood glucose level 95 mg/dl), cell count 2475/μl.
グラム染色するも明らかな菌が見つからず、髄液培養でも優位な菌は培養されなかった。 細菌性髄膜炎に対するグラム染色の感度は 60%程度であり、培養に関しても感度は高くない。 また髄液中の糖はもう少し減るのではないだろうか。 確定診断はつかないものの、最も疑わしい疾患であった。 起因菌は MRSA, 腸内細菌等を広域にカバーするためバンコマイシン, メロペネム (髄膜炎 dose) とした。	Gram stain did not reveal any obvious bacteria, and cerebrospinal fluid culture also did not reveal any predominant bacteria. The sensitivity of the gram stain for bacterial meningitis is about 60%, and the sensitivity of the culture is not high either. Also, the glucose in the cerebrospinal fluid would have been slightly lower. Although no definitive diagnosis could be made, bacterial meningitis was the most suspicious disease. The causative organism was assumed to be MRSA, and vancomycin and meropenem (meningitis dose) were used to cover a wide range of enteric bacteria.

possible that the creation process is more efficient by copying and pasting. In addition, 87% of the summaries contained at least one copy-and-paste sentence. Another possible lead from these is that it may be possible to achieve good performance by mixing the extractive and abstractive methods, i.e., the same methods as in the present summarization task [72, 73], since there are some extractive representations.

Table 3.2 Statistics of the target data

Inpatient records

Dataset	Cases	Sentences/Document	Words/Sentence	Characters/Sentence
NHO data	24,641	192.0	9.0	18.1
GSK2012-D	45	97.4	7.5	15.1
MedNLP	278	22.6	12.7	22.4
Dummy record	108	274.1	9.1	18.5

Discharge summary

Dataset	Cases	Sentences/Document	Words/Sentence	Characters/Sentence
NHO data	24,641	35.0	12.4	23.3
Dummy record	108	17.4	18.6	34.4

Chapter 4

Clinical Segment

4.1 Introduction

In this chapter, we attempt to identify linguistic units that can properly cover medical meanings for later studies. This linguistic unit is used in the extractive summarization experiments in Chapter 6 and as the annotation target for the rest of this thesis. It is important to identify the medical domain-specific unit, rather than sentence and clause, for processing and analysis in medical documents.

In the past, several types of linguistic units have been defined, such as sentence, clause, or phrase, which are called self-contained linguistic unit [74]. However, they are in the general domain and would not always fit in the medical domain. In addition, the linguistic units in Japanese is a little different from that in English. In particular, clauses in Japanese have significantly different characteristics from clauses in English because they can be formed by simply adding a particle to a noun. Owing to this characteristics, Japanese clauses are often very short at the phrase level. Another similar concept is the elementary discourse unit [75], which breaks down sentences into smaller units for the analysis of relationships in discourse structure. This concept is also defined in Japanese [76]. However, the elementary discourse unit in Japanese is the clause itselff and the authors annotate discourse relations to clauses. Therefore, the boundary of separation is the same as the clause, so it was not tested in this study. Accordingly, they cannot constitute a meaningful unit that carries concepts of medical significance. Therefore, we need a new self-contained linguistic unit that has

Table 4.1 Examples of the three types of units. ⟨SEP⟩ indicates the boundary of either a segment or clause.

Units	Examples
Sentence	認知症が進んでおり自宅退院は困難であること、施設入居のためにはご家族の手続きが必要になることを説明 (We explained that it would be difficult to discharge her due to her advanced dementia, and that her family would need to make arrangements to move her into another facility.)
Segment	認知症が進んでおり⟨**SEP**⟩自宅退院は困難であること、⟨**SEP**⟩施設入居のためにはご家族の手続きが必要になることを⟨**SEP**⟩説明 (Due to her advanced dementia ⟨**SEP**⟩ it would be difficult to discharge ⟨**SEP**⟩ her family would need to make arrangements to move her into another facility ⟨**SEP**⟩ we explained)
Clause	認知症が進んでおり⟨**SEP**⟩自宅退院は⟨**SEP**⟩困難である⟨**SEP**⟩こと、⟨**SEP**⟩施設入居のためには⟨**SEP**⟩ご家族の手続きが必要になることを⟨**SEP**⟩説明 (Due to her advanced dementia ⟨**SEP**⟩ discharge ⟨**SEP**⟩ it would be difficult ⟨**SEP**⟩ (verb nominalizer) ⟨**SEP**⟩ to move her into another facility ⟨**SEP**⟩ her family would need to make arrangements ⟨**SEP**⟩ we explained)

a longer span than a clause in Japanese and expresses the smallest medically meaningful concept.

For this reason, we defined the *clinical segment* that spans several clauses but is shorter than a sentence. A comparison of the clause and sentence is shown in Table 4.1. As exemplified in the table, segments may comprise clauses connected by a conjunction to form a medically meaningful unit; alternatively, they may be identical to clauses. In addition, for the statistical analysis, the clinical segment must be defined formally so that a splitter can automatically divide sentences into segments. We investigated the performance of splitting clinical segments mechanically by building an automated splitter. The codes developed in this chapter are publicly available [1].

[1] https://github.com/ken-ando/Exploring-optimal-granularity-for-extractive-summarization-of-unstructured-health-records

4.2 Design and annotation of clinical segment

In designing the clinical segment, we attempted to distill the atomic events related to medical care as a single unit. For example, statements such as "jaundice was observed in the patient's conjunctiva," "the patient was diagnosed with hepatitis," and "a CT scan was performed" would lose their medical meaning if they are further split. In addition, medical events are the central statements in medical documents, whereas non-medical events play a relatively small role. Therefore, in this study, we considered only medical events as a component of self-contained units, and non-medical events were interpreted as noise. In other words, a clinical segment cannot be formed if the span consists only of non-medical events. A self-contained unit was defined with respect to semantics in previous studies. In our study, it was extended to a pragmatic unit based on domain knowledge. The details of the six segmentation rules are listed in Table 4.2.

Based on this definition, we annotated the clinical segment to the dummy record. The annotation was made by one author and medical professionals and labeled, resulting in two different labels. In the result, the total number of segments in the corpus was 3,816, the average number of segments per sentence was 2.18, and the average number of segment boundaries per sentence was 1.18. The agreement rate between the participants of the segmentation task and an author is 0.82, which is sufficiently high to be used for further study. The agreement rate is the accuracy of the workers' labels for the correct boundaries annotated by the author. Across this task, we adopted the labels annotated by one of the authors.

4.3 Automated segmentation

Table 3.1 shows a discharge summary—a type of medical record written by a Japanese physician. As illustrated, it is a noisy document: punctuation marks are missing, and line breaks appear in the middle of a sentence. Sentence boundaries may be denoted by spaces instead of punctuation marks. Therefore, for the further analysis of the three types of extraction units, we first need preprocessing for *sentence splitting* and *segment splitting*.

For sentence splitting, we adopt two naive rules below to define the boundaries of a sentence:

Table 4.2 Segmentation rules

Rule 1 *Split at the end position of a predicate, by a comma or a verbal noun.*
This is the base rule for segmentation, and others are exception rules.
(e.g., "絶食、⟨SEP⟩ 抗菌薬投与で ⟨SEP⟩ 肺炎は軽快。")
(e.g., "(After) fasting and ⟨SEP⟩ antibiotic use, ⟨SEP⟩ pneumonia was relieved.")

Rule 2 *If a segment is enclosed in parentheses, split a sentence at the positions of parentheses.*
To extract the clinical segment inside parentheses, parentheses sometimes become segment boundaries.
(e.g., "画像で「⟨SEP⟩ 両側肺門部に陰影あり、⟨SEP⟩ CT で両肺に多彩な浸潤影を認め ⟨SEP⟩ 重症肺炎」⟨SEP⟩ として 4 月 10 日に入院。")
(e.g., "On imaging, " ⟨SEP⟩ there are bilateral hilar shadows and ⟨SEP⟩ widespread consolidation in both lungs on CT scan, ⟨SEP⟩ (suspected of) severe pneumonia " ⟨SEP⟩ (the patient was) admitted to the hospital on April 10.")

Rule 3 *Split content that includes disease name.*
Disease names are often written as diagnoses and play an important role in EHRs. Therefore, even if rule 1 does not match, the content that includes disease names should be split.
(e.g., "肺炎疑いで ⟨SEP⟩ 当院紹介となった。")
(e.g., "Due to suspected pneumonia, ⟨SEP⟩ he was referred to our hospital.")

Rule 4 *Split examination results and their evaluation.*
Examination results and their evaluation are often written in a single sentence. Because the meaning of the examination results and their evaluation are clearly different, they should be divided even if rule 1 does not match.
(e.g., "血清クレアチニンキナーゼは 4512 U/L と ⟨SEP⟩ 高度に上昇していた。")
(e.g., "Serum creatinine kinase level was 4512 U/L, ⟨SEP⟩ which was highly elevated.")

Rule 5 *Do not split anything that is not related to the medical treatment.*
If the content is medically meaningless, its role is not important in its document, and it is not worthy of analysis. Therefore, the content with little relevance to medical treatment is not split, even if it matches rule 1.
(e.g., "ケアマネジャーに同伴されて来院した。")
(e.g., "She came to our hospital accompanied by her care manager.")

Rule 6 *Do not split content that does not add meaning.*
If the content that supplements the meaning of the previous description does not add meaning (e.g., "... schedule to [VP] ..." and "... continue the treatment ..."), it is not split even if it matches rule 1.
(e.g., "外来で抜糸を行う方針とした。")
(e.g., "It was planned to remove sutures as an outpatient.")
This includes contents where the semantic label does not change before and after the split.
(e.g., "発熱、盗汗、体重減少、喀痰、血痰は否定。")
(e.g., "Fever, sweating, weight loss, sputum, and bloody sputum were not observed.")
It also includes contents that represent the passage of time or assumptions.
(e.g., "抗菌薬開始後、発熱・腹痛は徐々に改善し")
(e.g., "After starting antibiotic use, fever and abdominal pain gradually improved.")

1. A statement that ends with a full-stop mark.

2. A statement that ends with a new line and has no full-stop mark.

There is oversimplification here, compared to sentence splitting tasks in medical NLP that have been studied [77, 78]. However, since it is not a focus of this study, we adopted this naive approach for its simplicity. In this process, we also used MeCab [79] as a tokenizer. The MeCab's dictionaries are mecab-ipadic-NEologd [80] and J-MeDic [81] (MANBYO 201905).

Next, sentences must be automatically split into clinical segments to efficiently analyze the huge dataset, NHO data. We compared several approaches to achieve the best splitting performance. In this study, we used 3,816 annotated segments in the corpus and applied six-fold cross-validation.

We used three rule-based splitters as baselines: a simple rule-based model for splitting by full-stop marks (**Full-stop**), another simple rule-based model for splitting by full-stop marks and verbs (**Full-stop & Verb**), and a complex rule-based model for splitting by clauses (**CBAP**) [82]. To be more precise, in the case of the Full-stop & Verb model, it starts with a verb and splits in front of the next occurring noun except for non-independents. The last model, which included 332 rules that were manually set up based on morphemes, was used to confirm that clinical segments have different boundaries than traditional clauses.

We used **SEGBOT** [83] as a machine learning method based on a pointer network architecture [84] for the splitting task. The method includes three phases: encoding, decoding, and pointing. An overview is shown in Fig 4.1. Medical records may include local dialects and technical terms that are not listed on public language resources. Accordingly, the splitter must handle even unknown words. In our approach, each input word is first represented by a distributed representation using fastText [85, 86]. FastText is a model that acquires vector representations of words considering the context. Notably, fastText can obtain vectors of unknown words by decomposing them into character n-grams. These vectors capture hidden information about a language, such as word analogies and semantics.

The performance of the splitter methods is summarized in Table 4.3. The machine-learning-based SEGBOT outperformed the others, with its F1 score being 0.257 points

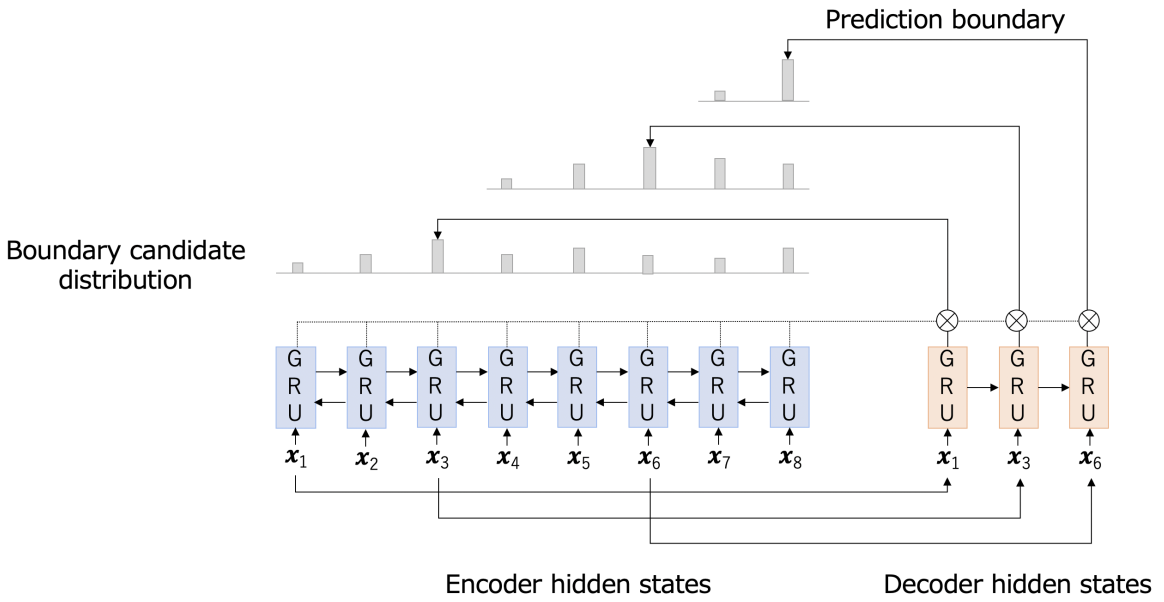


Fig. 4.1 Overview of SEGBOT.

Table 4.3 Results of the segmentation task.

	Precision	Recall	F1 score
Full-stop	0.521	0.187	0.275
Full-stop & Verb	0.569	0.610	0.589
CBAP [82]	0.368	0.464	0.411
SEGBOT [83]	**0.864**	**0.829**	**0.846**

The numbers in bold indicate the best performing methods.

higher than that of the Full-stop & Verb model, which was the second best. Since this precision of 0.864 is higher than the inter-annotator agreement, it is considered to be almost the upper bound. In addition, CBAP, which is a clause segmentation model, has a low F1 score of 0.411, suggesting that the definitions of the clause and the clinical segment are inherently different. The precision of the model with splitting at the full-stop marks (Full-stop) is only 0.521, indicating that the clinical segment is not always split at the full-stop marks, and that it is necessary to consider the context for splitting. Overall, the results suggest that machine learning is the best fit for the segmentation task. Thus, the data preprocessed by this method are used for the main experiment of this study.

4.4 Granularity comparison

Table 4.4 shows the statistical relation of the three types of units. The first column shows how many units are included in a sentence on average. The second and third columns show the average number of tokens and characters included in each type of units. The result suggests that segments are longer than clauses *on average*. Nevertheless, the difference of a clause and a segment is not significant, at least for the average number of characters. Accordingly, the relationship between clause and clinical segment granularity is worthy of a more detailed analysis.

We ensure the order of the three types of linguistic units, by an additional experiment on word-wise relation between clauses and clinical segments. For any two linguistic units in a sentence, there are four possible relationships (Fig 4.2): "Equal" is where the two match exactly, e.g., "認知症が進んでおり" of a clause and "認知症が進んでおり" of a segment. "Inclusive" is where a segment completely includes a clause, e.g., "自宅退院は困難である<SEP>こと、" of a clauses and "自宅退院は困難であること、" of a segment. "Included" is where a clause completely includes a segment. "Overlap" is where the two overlaps.

We obtained statistics of the four relationships, from all inpatient records and discharge summaries in the NHO data. The results are shown in Table 4.5. We found that 59.6% of them have the same boundaries. This is influenced by the many short sentences that have no boundaries. Then, "Inclusive" shared 20.0% of the relations. The sum of "Equal" and "Inclusive" turned out to be 79.6%, which is six times more than "Included" that shared only 13.1%. The figures gives the detailed dynamics of the relation between segments and clauses, shown just as 11.83 and 10.74 characters/unit in Table 4.4. Although the difference in the average length between segment and clause is small, there is a significant difference between segments and clauses in their relative sizes, when compared by each corresponding pair of the actual units.

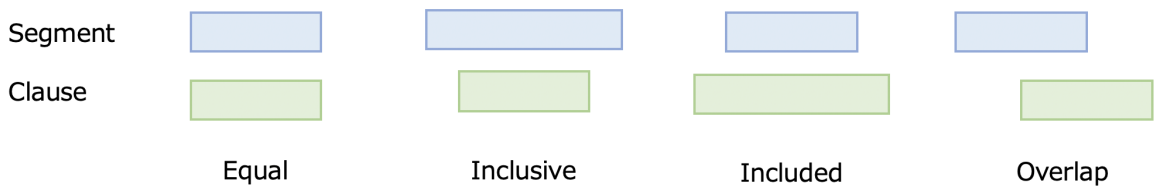


Fig. 4.2 The four types of relationship between clause and clinical segment.

Table 4.4 Granularity of three units. The numbers in bold indicate the smallest units.

Units	Units/Sentence	Tokens/Unit	Characters/Unit
Sentence	1	8.98	18.06
Segment	2.18	6.42	11.83
Clause	**2.75**	**5.74**	**10.74**

Table 4.5 The Relationships between clauses and clinical segments.

Relation types	Equal	Inclusive	Included	Overlap
Number of relationships	6,687,046 (59.6%)	2,239,839 (20.0%)	1,469,423 (13.1%)	821,663 (7.3%)

Chapter 5

Clinical Role Label

5.1 Introduction

For the analysis of the later summarization experiments in this thesis, we need to deeply examine expressions in NHO data. However, it is unrealistic to check expressions manually, for cost and privacy reasons. Therefore, this chapter defines labels of expression under physician supervision and develops a classification model for automation of the labeling. This is similar to medical term ontologies such as UMLS [87], but different because they are targeted only to entities. To the best of our knowledge, this is the first work that assigns medical roles to contextualized expressions. While previous studies have employed entity-based quantitative methods in the analysis of medical documents [88–91], our work shows the potential for extending that analysis over sentences. The codes developed in this chapter are publicly available [1].

5.2 Definition and annotation of clinical role label

Assuming the summarization phenomenon of physicians, the results of various processes are represented in the descriptions of discharge summaries. To analyze the patterns of summarization, we check expressions that appear in clinical documents and define the types. In this definition, it is assumed that the clinical facts are interpreted by physicians, and the

[1] https://github.com/ken-ando/Is-artificial-intelligence-capable-of-generating-hospital-discharge-summaries-from-inpatient-records

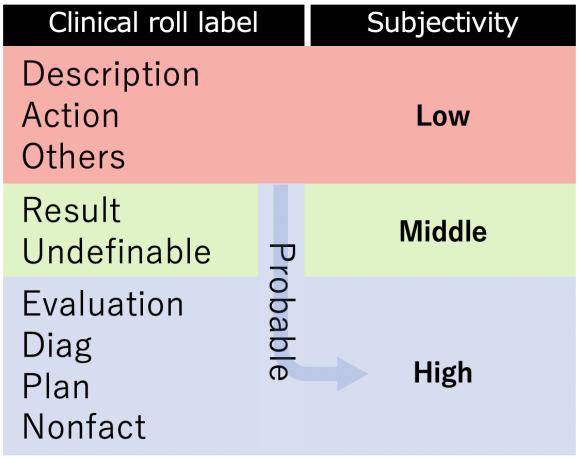


Fig. 5.1 Overview of the subjectivity and the clinical role label. The probable label is duplicated across six labels and they are classified as the high subjectivity.

processing progresses in this order in the summarization of inpatient records. For example, physicians may perform physical and laboratory examinations during the early stages of hospitalization. They recorded the results in the inpatient records as facts. Subsequently, evaluations of the test results, diagnoses, treatment plans, etc. would be performed by physicians based on their interpretations. Therefore, there must be a gradation in subjectivity in the descriptions that appear in inpatient records and summaries. Subjective descriptions may include interpretations of objective information in the source record. Based on this assumption, the clinical role labels are defined (Table 5.1). All definitions were designed under the supervision of a physician.

5.2.1 Low subectivity labels

First, *low subjectivity labels* are defined to include *Description*, *Action*, and *Others* labels. They consisted of objective facts and formed the basis of clinical records and discharge summaries.

Description labels comprise the content of past events and statuses. These are the fundamental contents of clinical records. For example, observations of patients, physical findings, test results, and paraphrasing of test results (e.g., “high blood pressure” instead of “Blood Pressure:180/90”), and past episodes fell under this category. The paraphrases included in this label are limited to the conversion of expressions without any interpretative comments.

Table 5.1 Details of the clinical role label. It shows label names, brief explanations, and examples in discharge summaries.

Label	Explanation	Example
	Low Subjectivity Labels	
Description	Past events and status.	Only pneumococcal urine antigen test results are positive.
Action	Past actions.	Discharged.
Others	Meaningless segments.	However“
	Middle Subjectivity Labels	
Result	Comments as seen, but can change slightly from person to person.	Infiltration shadow in the lower right lung field
Undefinable	Unclear whether descriptions are future plans caused by Japanese linguistic characteristics.	4月10日に入院。〈Hospitalized or will be hospitalized on April 10.〉
	High Subjectivity Labels	
Evaluation	Reasoning from facts.	Because it was considered to be an acute exacerbation of interstitial pneumonia
Diag	Clinical or definitive diagnosis.	Clinically diagnosed with small-cell lung cancer
Plan	Future treatment plans.	The patient was scheduled for long-term PCI.
Nonfact	Hearsay and assumptions, etc.	Considering his advanced age and limited life expectancy
Probable	Probabilistic expressions.	Suspected renal abscess or renal cell carcinoma.

The *history of present illness* section mostly consists of the *description* label because it is based on patients' past episodes.

Action label comprises the contents of someone's past actions (e.g., "hospitalized", "prescribed", and "discharged"). These were mostly medical treatment records. Here, action verbs can be active, passive, or other voices, and any form is acceptable for action content.

Others label comprises meaningless content from a medical perspective. Typical examples are dates, item names, parentheses before and after a segment, etc. (e.g., "【現病歴】" ⟨"[history of present illness]"⟩). Although these symbols are objective descriptions, they do not contain information about patients. Thus, this study reserved a class for such cases to simplify further processing.

5.2.2 Middle subectivity labels

Second, *middle subjectivity labels* are defined, which include *result* and *undefinable* labels. In clinical documents, determining the subjectivity of some descriptions is difficult; these categories are devised to accommodate such cases and maintain the quality of annotations for high and low labels.

The ***Results*** label comprises content that is slightly subjective to healthcare providers. For example, physicians record abnormalities and interpret images in radiological reports. However, they often comprise qualitative descriptions and objective expressions, which results in a combination of subjectivity and objectivity. Other examples include changes in test values (e.g., "improvement" and "worsening") that can also be influenced by physicians' subjectivity. Clinical documents may contain expressions that are difficult to categorize as factual or subjective. This label was intended as a buffer to cover borderline cases.

The ***Undefinable*** label comprises content that is unclear whether it is a reference for a future plan. In Japanese, to write a concise sentence, predicates are often transformed into nouns (e.g., "退院した" ⟨"Discharged. "⟩ → "退院" ⟨"Discharge."⟩ "検査する" ⟨"to examine."⟩ → "検査" ⟨"examination. "⟩). In such cases, whether the examples refer to past or future plans remains unclear.

5.2.3 High subectivity labels

Third, *high subjectivity labels* are defined, including *evaluation*, *diag*, *plan*, *nonfact*, and *probable* labels. This class comprises information that is a hypothetical or subjective statement by the writer. Such content is produced by accepting clinical findings as inputs, and then inferences, external knowledge, and personal insights are used to generate the outputs. This is the primary content of the *clinical course section*, which appears to be the most difficult part to summarize automatically.

The ***Evaluation*** label comprises content that is discussed and reasoned about, findings, test results, and events. This category is another core element of clinical records. A general example is the list of test results followed by discussion of the findings. In clinical texts, a description of *evaluation* may accompany a trailing *diag* description and can be inseparable if the descriptions are abbreviated (e.g., "diagnosed as COVID-19 based on the severe clinical course"). As our labeling framework allows multiple labels, a sentence may contain both *evaluation* and *diag* labels. However, this case rarely appeared in our annotation; thus, the *diag* label was prioritized over the *evaluation* label.

Diag labels were used for the clinical diagnosis. Although a definitive diagnosis can be performed objectively, the diagnosis relies on objective findings. Therefore, in this classification, diagnostic descriptions were considered low-objectivity classes. This label is similar to the *evaluation* label; it is also a core element of the clinical record. Note that there are medical concepts that can be both symptoms and diseases, depending on the context, such as "dyspnea." Such borderline cases are assigned to the *result* label to avoid contamination of the *diag* labels.

The ***Plan*** label was assigned to expressions that explicitly refer to future plans. In Japanese, such expressions often comprise certain terms, such as "予定 (schedule)" and "計画 (plan)." These are mainly written at the bottom of the inpatient records. They sometimes refer to the next scheduled visit and referral source that the patient will visit after discharge.

The ***Nonfact*** label comprises content written with hearsay or assumptions; however, it does not belong to any other label. Some characteristic words in Japanese, such as "if," "consider," and "say that," indicate that the content is not based on fact.

Finally, the ***probable*** label comprises clearly subjective content, such as "doubt" or "possibility." This label must have a multilabel structure as it can be added to any content In this

case, all contents labeled as *probable* are classified as *high* in the subjectivity label, regardless of the original label, because their information becomes subjective.

5.2.4 Annotation

We annotated dummy record with clinical role labels. We first split the sentences and decomposed them into 3,761 clinical segments, as described in Section 4.3. The annotation was conducted by two clinical workers, and the agreement rate was calculated as the accuracy, which was 0.790. The distribution of clinical role labels is shown in the left half of Table 5.2.

The most common label was *description*, followed by *action*. Both are past facts that appear to be appropriate considering the original purpose of the medical records, which was to record the medical treatment process. In addition, *description* was twice as common as *action*, suggesting that recording the past status plays a major role in clinical records. For high subjectivity, *evaluation*, *diag*, and *plan* were nearly the same in number, whereas *nonfact* and *probable* were relatively low. This suggests that the medical records consist of an equal amount of evaluation of findings and test results, clinical diagnosis, and plans for future treatment.

5.3 Automation of labeling

Using the annotated dummy records, this study trained a classification model that was used to classify the NHO data shown on the right-hand side of Table 5.2.

5.3.1 The classification model

An overview of the proposed model is shown in Fig 5.2. As the basic architecture for classification, this study adopted a pretrained neural model, BERT (Bidirectional Encoder Representations from Transformers) [92]. Because its parameters are learned from a large number of documents in advance, BERT is known to achieve good accuracy even with few training samples. In this study, UTH-BERT was used [93], an improved version of BERT

Table 5.2 Distribution of the clinical role and subjectivity labels. The labels in the dummy record were manually annotated by clinical workers and the NHO data were automatically annotated.

		Dummy record		NHO data	
Subjectivity	Clinical role	Number of segments (%)		Number of segments (%)	
Low	Description	1,463 (37%)	2,324 (61%)	484,385 (32%)	917,724 (60%)
	Action	797 (20%)		183,245 (12%)	
	Others	65 (2%)		241,729 (16%)	
Middle	Result	306 (8%)	646 (17%)	160,118 (11%)	401,049 (26%)
	Undefinable	340 (9%)		258,939 (17%)	
High	Evaluation	278 (7%)	844 (22%)	45,043 (3%)	205,671 (14%)
	Diag	255 (6%)		60,772 (4%)	
	Plan	264 (7%)		53,389 (4%)	
	Nonfact	82 (2%)		12,511 (1%)	
	Probable	133 (3%)		24,313 (2%)	

that was pre-trained on clinical records from the University of Tokyo Hospital. In contrast to previous Japanese BERT models [94–96], which were pre-trained mainly on web data such as Wikipedia, UTH-BERT was expected to perform better on documents in our target domain. (For more detailed architecture, training methods, and performance of UTH-BERT, see previous papers [92**?** , 93].)

This study also adopted a multitask learning framework. Multitask learning achieves improved performance by exploiting the relationship between labels and is considered to provide various benefits (e.g., regularization, eavesdropping, and data augmentation [97]). In our study, three labels (i.e., subjectivity, clinical roles, and probable labels) were assigned to a clinical segment, and multitask learning compensated for the small data volume of the dummy records by virtually multiplying the labels used for learning. Subjectivity prediction can also aid in a more complex clinical role prediction task.

The processing pipeline operates as follows: a clinical segment split from the target dataset is input into the BERT. The input segment was previously tokenized by WordPiece [**?**] and provided with tokens “[CLS]” for the head and “[SEP]” for the tail. Then, the $[\mathrm{CLS}]_{\mathrm{hidden}}$ vector from the final layer of BERT is obtained and inputted to a separate three-layer perceptron for each of the three labels. It calculates the cross-entropy loss value based

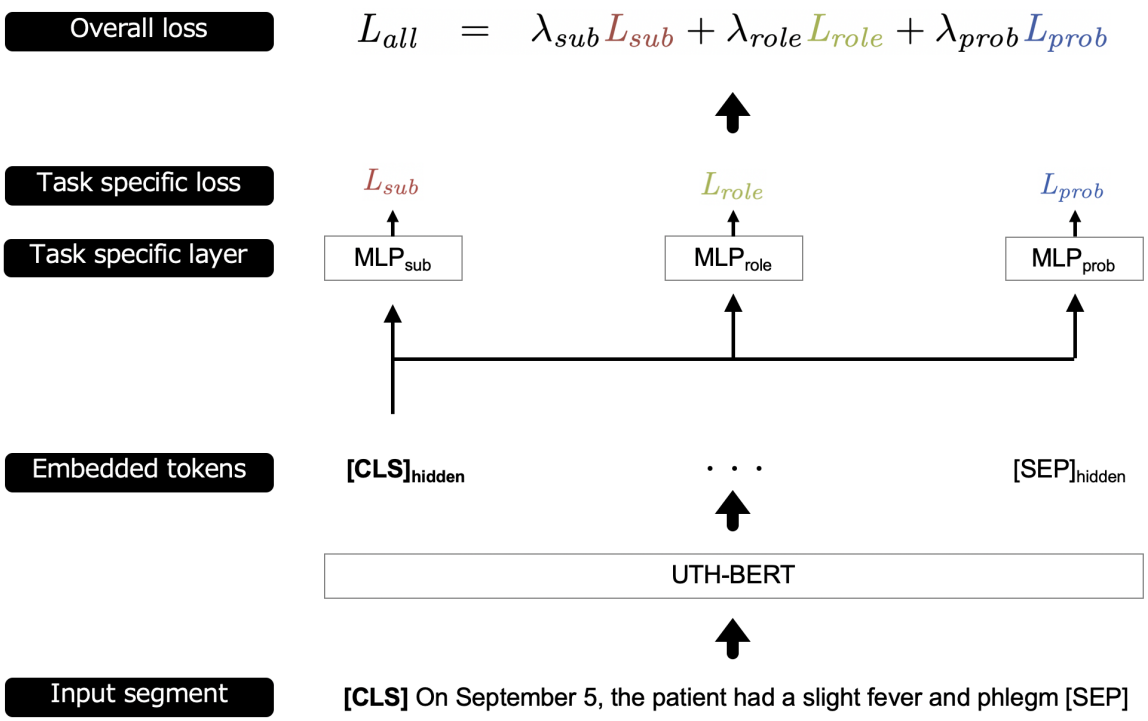


Fig. 5.2 Overview of the classification model for subjectivity, clinical role, and probable label. Each of the three labels is defined as three tasks. Input segments are fed to UTH-BERT, and then the outputs to the specific layers. Finally, the loss scores of three tasks are calculated and combined to obtain the overall loss score.

on the gold and predicted labels and obtains three loss values. The model was trained using the weighted sum of the three loss values as the overall loss. In this process, BERT is trained only on the parameters of the final layer. The weighted sum L_{all} is formulated as

$$L_{all} = \lambda_{sub} L_{sub} + \lambda_{role} L_{role} + \lambda_{prob} L_{prob}, \tag{5.1}$$

where L_{sub}, L_{role}, and L_{prob} are the loss values for subjectivity, clinical role, and probable, respectively, and λ_{sub}, λ_{role}, and λ_{prob} are the hyperparameters of the corresponding weights. λ_{sub}, λ_{role}, and λ_{prob} were normalized and summed to 1.

In the implementation, this study employed UTH-BERT, which was pre-trained using the method of whole-word masking. In addition, the Adam optimizer was used [98] for 20 epochs, and the learning rate $\eta = 0.00001$. Based on the available memory and training per-

formance, the batch size is set to 32. The setup of the other hyperparameters was the same as that in UTH-BERT. For the development and testing data, 300 samples were randomly selected from the dummy records, and the remaining examples were used as training data. The average results of three training runs with different seeds are reported.

5.3.2 Results of classification

The hyperparameter search and classification results are presented in Tables 5.3 and 5.4, respectively. The model performance was evaluated using the F1 score against the correct labels. The F1 score is the harmonic mean of recall and precision. Recall is formulated as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, where TP is the number of true positives and FN is the number of false negatives. In addition, precision was formulated as $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$, where FP is the number of false positives. Let recall be R and precision be P, $\mathrm{F1} = \frac{2\mathrm{RP}}{\mathrm{R+P}}$. In the hyperparameter search, each weight was changed by 0.25 and grid-searched to find the optimal value. The F1 scores with individual labels are found in the columns in which λ_{sub}, λ_{role}, and λ_{prob} are 1. This study found that multiple-label settings were always better than single-label settings.

In the detailed classification results for each label, this study found that the model could be classified with much higher accuracy for high and low subjectivities. The classification accuracy is lower for middle subjectivity, which is not surprising because this label includes ambiguous segments that improve annotation quality. This study did not use the middle label for further analysis. Furthermore, in the detailed labels for high and low subjectivities, this thesis found that *others* and *nonfact* are low. This was because of the small sample size of these labels. For the same reason, the *probable* label is less accurate.

Table 5.3 Results of a hyperparameter search. Three labeling tasks are conducted as independent tasks, and the weights of the tasks are slid by 0.25 to find the optimal value in multitask learning. Experiments are conducted using dummy records.

Hyperparameters																
	Subjectivity															
λ_{sub}	1	0.75		0.5			0.33	0.25				0				
	Clinical role															
λ_{role}	0	0.25	0	0.5	0.25	0	0.33	0.75	0.5	0.25	0	1	0.75	0.5	0.25	0
	Probable															
λ_{prob}	0	0	0.25	0	0.25	0.5	0.33	0	0.25	0.5	0.75	0	0.25	0.5	0.75	1
Accuracy																
$F1_{sub}$	0.830	0.834	0.825	0.846	**0.862**	0.856	0.856	0.859	0.844	0.840	0.823	0.574	0.534	0.854	0.406	0.598
$F1_{role}$	0.060	0.751	0.069	0.765	0.767	0.050	0.771	**0.774**	0.745	0.765	0.053	0.747	0.722	0.772	0.737	0.008
$F1_{prob}$	0.382	0.397	0.951	0.967	0.965	0.962	**0.969**	0.425	0.967	**0.969**	0.967	0.395	0.960	0.384	**0.969**	0.962

Table 5.4 Results of automatic labeling using dummy record. The hyperparameters of λ_{sub}, λ_{role}, and λ_{prob} are 0.5, 0.25, and 0.25 for subjectivity; 0.25, 0.75, and 0 for clinical role; and 0.33, 0.33, and 0.33 for probable label.

Roles	Precision	Recall	F1
Description	0.86	0.87	0.87
Action	0.88	0.81	0.84
Others	1.00	0.43	0.60
Result	0.65	0.65	0.65
Undefinable	0.55	0.77	0.64
Evaluation	0.61	0.83	0.70
Diag	0.79	0.88	0.83
Plan	0.81	0.65	0.72
Nonfact	0.00	0.00	0.00

Probable	Precision	Recall	F1
Positive	0.60	0.55	0.57
Negative	0.98	0.99	0.98

Subjectivity	Precision	Recall	F1
Low	0.92	0.90	0.91
Middle	0.66	0.73	0.70
High	0.85	0.84	0.84

The distribution of automatically assigned labels is shown in the right-hand half of Table 5.2, along with the distribution of the dummy records. This study found that low subjectivity was present in the same proportion as in the dummy record. Upon comparing middle and high subjectivity, the statistics show that middle subjectivity is more common. This is because many formatting expressions exist in NHO data, such as examination results, dates, and times. Overall, the distributions of the dummy records and NHO data were mostly consistent. This suggests the appropriateness of the automated labeling process.

Chapter 6

Exploring Optimal Granularity for Extractive Discharge Summary Generation

6.1 Introduction

Automated summarization of daily inpatient records involves various technical topics and challenges. For example, descriptions of important findings related to a patient's diagnosis require an extractive summary. Our preliminary experiments revealed that 20–31% of the sentences in discharge summaries were created by copying and pasting. This result proves that a certain amount of content can be automatically generated by extractive summarization. Meanwhile, when a patient is discharged from the hospital after surgery without any major problems, it is necessary to summarize the clinical record as the patient "recovered well after the surgery," even if more details of the postoperative process are described in the records. Therefore, such descriptions cannot be created by copy and paste, and needs to be abstracted. These observations suggest that the generation of discharge summaries is a complex process that is a mixture of extractive and abstractive summarization, and it remains unclear how to process the unstructured source texts, i.e., free-texts. To advance this research field, it is desirable to properly decompose these summarization processes and clarify their interactions.

To this end, this chapter focuses on the extractive summarization process by physicians. Some recent studies investigated the best granularity units in this type of summarization [26, 27]. However, the granularity of extraction has not been explored for the summarization of medical documents. Thus, we attempted to identify the optimal granularity in this context, by defining three units with different granularities and comparing their summarization performance: whole sentences, *clinical segments*, and clauses.

An overview of our study is shown in Figure 6.1. First, the inpatient records are split into sentences by some rules. Second, sentences are automatically split into clauses and clinical segments. Finally, we evaluate the performance of automatic summarization models with extractive architecture inputting the sentence, clause, and clinical segment units. The codes developed in this chapter are publicly available [1].

6.2 Related work

To identify the optimal granularity of extractive summarization, there are two approaches. One approach is a method that takes n word sequences of arbitrary lengths and compares them as the units for summarization. The other approach is a method that uses predefined linguistic units. Previous studies [26, 27] in this domain have used the latter approach and found that a sentence was a longer-than-optimal granularity unit for extractive summarization. A study adopted a clause as a shorter self-contained linguistic unit [74] instead of a sentence [26]. However, it remains unclear whether the clause performs the best in the summarization of clinical records or there could be further possibilities.

6.3 Summarization model

In an extractive summarization task, the goal is to automatically assign a binary label to each unit of the input to indicate whether this unit should be included in the summary. Therefore, we adopted a single classification model to cover the three types of units.

[1] https://github.com/ken-ando/Exploring-optimal-granularity-for-extractive-summarization-of-unstructured-health-records

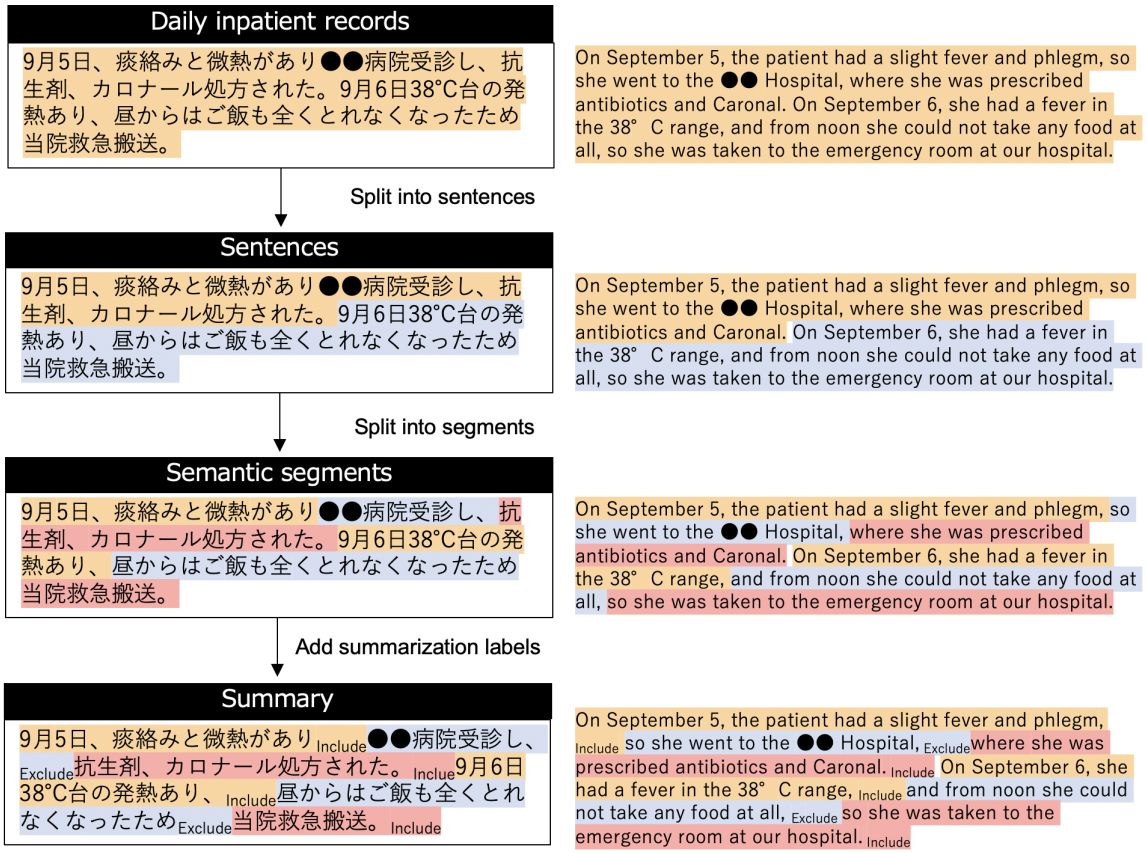


Fig. 6.1 Outline of our pipeline.

The top block is an example of the inpatient record, and the subsequent blocks indicate the chain of processes up to adding summarization labels.

Following Zhou et al. [26], we used a model based on BERT [92], as shown in Fig 6.2. Instead of the original work that adopted BERT as an encoder for extractive summarization, we adopted UTH-BERT [93].

Formally, let the i-th sentence contain l segments $S_i = (s_{i,1}, s_{i,2}, ..., s_{i,l})$. The j-th segment with k words in S_i is denoted by $s_{i,j} = (w_{i,j,1}, w_{i,j,2}, ..., w_{i,j,k})$. We add [CLS] and [SEP] tokens to the boundaries between sentences. After applying the UTH-BERT encoder, the vector of tokens is represented as $(w^{BT}_{i,j,1}, w^{BT}_{i,j,2}, ..., w^{BT}_{i,j,k})$. Next, we apply average pooling at the segment level. The pooled representation $s'_{i,j}$ is formulated as follows:

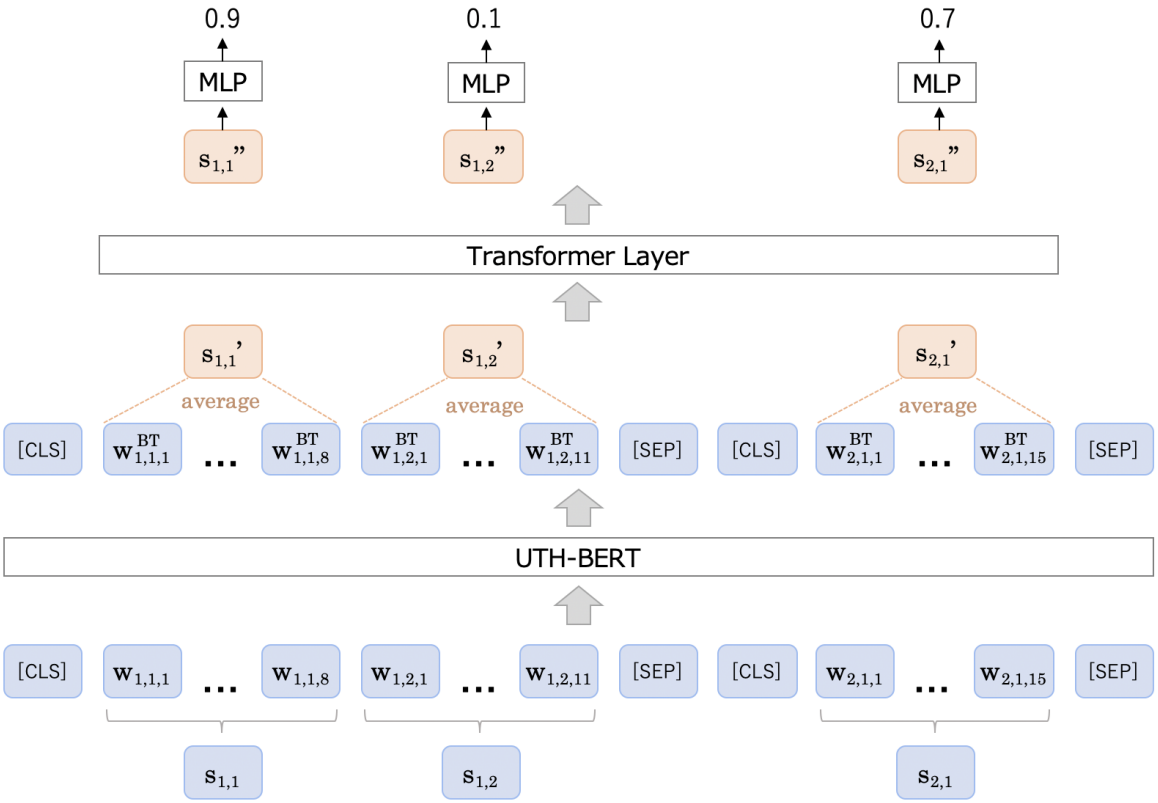


Fig. 6.2 Overview of classification model for clinical segments.

$$s'_{i,j} = \frac{1}{k}\sum_{1}^{k} w^{BT}_{i,j,k}. \tag{6.1}$$

Note that segments and clauses do not include the [CLS] and [SEP] tokens in average pooling. Subsequently, we apply a segment-level transformer [99] to capture their relationship for extracting summaries. The model predicts the probability of extracting for summary from those outputs as follows:

$$S'' = \text{Transformer}(S'), \tag{6.2}$$

$$p(s''_{i,j}) = \sigma(W_o s''_{i,j} + b_o), \tag{6.3}$$

where $S' = (s'_{1,1}, s'_{1,2}, ..., s'_{i,j})$ is a sequence of segments input to the transformer, and $S'' = (s''_{1,1}, s''_{1,2}, ..., s''_{i,j})$ is a sequence that is the output of the transformer. The training objective

of the model is the binary cross-entropy loss given the gold label $y_{i,j}$ and the predicted probability $p(s''_{i,j})$.

This model does not need to change its structure depending on the input units. For clauses, the span of the segments is replaced by that of the clauses. In the case of sentences, the average pooling is not performed; instead, we input the [CLS] token into the transformer.

6.4 Training data

Our model requires an entire document for training. However, our corpus could be too small to be used for the training of the model, and would compromise the robustness of the model. Accordingly, we used NHO data as training data by assigning pseudo labels. Following previous studies [27, 26], we used the ROUGE scores to automatically assign gold labels to the three units. We used the ROUGE score both to create the gold labels and to evaluate the model. This may seem unusual, but it is a commonly used approach in previous studies. As ROUGE is correlated with human scores [100], the best summary can be obtained by creating a system that maximizes this score during evaluation, regardless of whether this score was used during training. The labeling steps were as follows.

First, we applied the splitter created in Section 4.3 to the NHO dataset and split it into clauses and clinical segments. In this manner, we easily obtained a larger dataset. We used CBAP as a splitter for clauses and SEGBOT as a splitter for clinical segments.

Second, we measured ROUGE-2 F1 for each unit of the source documents (against the discharge summaries), which were then sorted in descending order of their scores. Thus, we obtained a list of units that were important for our summary.

Third, we selected the units from the topmost part of the list. At this stage, we stopped selecting units when the result exceeded 1,200 characters, which was the average length of the summaries in the NHO data.

Finally, we assigned positive labels to the selected units. The entire process yielded the gold standard for the training and evaluation without manual annotation. We randomly selected 1,000 documents each for the development and test sets, and we used the remaining 22,641 documents for the training data.

Table 6.1 Results of the summarization task.

Units	ROUGE-1	ROUGE-2	ROUGE-L
Sentence	31.91	2.50	7.93
Segment	**36.15**	**3.12**	**8.26**
Clause	25.18	1.30	6.62

The numbers in bold indicate the best performing methods.

6.5 Experiments and results

In this experiment, we used the three contextual units, instead of the n-gram units, and evaluated their impact on the summarization performance to determine which unit performs the best. The results of summarization, using the three types of units, are shown in Table 6.1. Comparing the three types of units in granularity, the model with clinical segments scored the highest in ROUGE-1, ROUGE-2, and ROUGE-L. The model with clinical segments outperformed sentences and clauses in summarizing inpatient records.

In summary, clinical segments exhibited the best performance in ROUGE and it lies between sentences and clauses in their size. Combining the results in this chapter, we can conclude that the segment units we introduced in this thesis are better and optimal units that lie between sentence and clause units.

6.6 Discussion

The result that extractive summarization with sentences is less effective than with other granularities is consistent with previous studies [26, 27]. Given the consistency of these results, this could be a universal property that must be exploited in further summarization tasks in NLP research.

In the summarization of medical documents, the experimental results of using linguistic units suggest that physicians create discharge summaries by capturing clinical concepts from the inpatient records. On the other hand, sentences and clauses performed poorly, probably because they were chunked only with syntactic information and did not deal with medical concepts. Accordingly, automatic summarization in the medical field requires not only syn-

tactic information but also high-level semantic and pragmatic information related to domain knowledge. Clinical segments are reasonable candidates as atomic units that carry medical information. Therefore, clinical segments can potentially be used to quantify the quality of medical documentation and to acquire more detailed medical knowledge expressed in texts.

Limitations in the current study and analysis are twofold: language and cultural dependency. Firstly, Japanese grammar and Japanese medical practices are very different from those of European languages, and there can be differences in the description, summarization, and evaluation processes. Accordingly, this pipeline using extractive method might be applicable only to Japanese clinical setting. In particular, the clinical segment was defined for Japanese, only labeled corpus for Japanese exists, so it is not naively applicable to other languages. However, the idea of capturing medical concepts may be useful for other languages. Also, more researches at various institutions would be preferable to confirm the generalizability of our results, although our study used the largest multi-institutional health records archive in Japan. Secondly, in some countries with different cultural background, *dictation* is used in clinical records and their summaries [101]. In this regard, Japanese hospitals do not use dictation to produce discharge summaries, which could result in frequent copying and pasting from sources to summaries. This custom could have contributed to using extractive texts in the discharge summaries in Japan. The analysis of the influence of this customary difference is left for future work.

6.7 Conclusion

In this study, we explored the best granularity for the automatic summarization of medical documents. The result indicated clinically motivated semantic units, larger than clauses, are the best granularity for the extractive summarization.

The results of this study suggest that the clinical segments that we have introduced are useful for automated summarization in the medical domain. This provides an important insight into how physicians write discharge summaries. Previous studies have used other entities to analyze medical documents [89, 102, 103]. Our results will help to provide more effective assistance in the writing process and automated acquisition of clinical knowledge.

Chapter 7

Is In-hospital Meta-information Useful for Abstractive Discharge Summary Generation?

7.1 Introduction

While there are some previous studies that discharge summary generation using abstractive summarization methods, most of them focused on only inpatient records for inputs. Properly summarizing an admission of a patient is a quite complex task, and requires various meta-information such as the patient's age, gender, vital signs, laboratory values and background to specific diseases. Therefore, discharge summary generation needs more medical meta-information, than similar but narrower tasks such as radiology report generation. However, what kind of meta-information is important for summarization has not been investigated, even though it is critical not only for future research on medical summarization, but also for the policy of data collection infrastructure.

In this chapter, we first reveal the effects of meta-information on neural abstractive summarization on admissions. Our model is based on an encoder-decoder transformer [99] with an additional feature embedding layer in the encoder (Figure 7.1). Hospital, physician, disease, and length of stay are used as meta-information, and each feature is embedded in the vector space. For experiments, we collect inpatient records, discharge summaries and

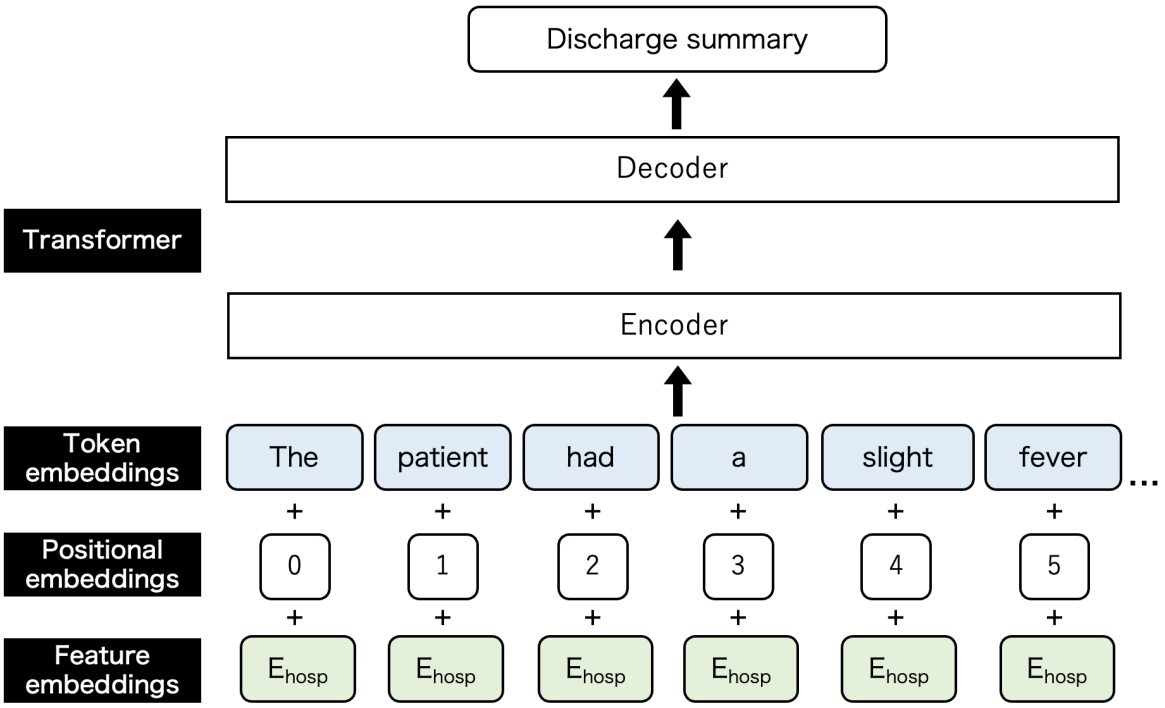


Fig. 7.1 Overview of our proposed method. A new feature embedding layer encoding hospital, physician, disease, and length of stay is added to the standard transformer architecture. The figure shows an example of hospital embedding.

coded information from the electronic health record system, which are managed by a largest multi-hospital organization in Japan. Our main contributions are as follows:

- We first apply the abstractive summarization method to generate Japanese discharge summaries.
- We found that a transformer encoding meta-information generates higher quality summaries than the vanilla one, and clarified the benefit of using meta-information for medical summarization tasks.
- We found that a model encoding disease information can produce proper disease and symptom words following the source. Also, we found that the model using physician and hospital information can generate symbols that are commonly written in the summary.

7.2 Related work

Studies using medical meta-information have long been conducted on a lot of tasks. Xu et al. [104] and Scheurwegs et al. [105] predicted diagnosis codes [106, 107] using medication codes [108, 109], procedure codes [110, 106], and lab test results. Choi et al. [111] predicted disease and medication information using medial metainformation of past admissions. Futoma et al. [112] and Zhang et al. [113] also used medication and physiologic meta-information to predict future admission and sepsis onset, respectively. In abstractive summarization on discharge summary, Diaz et al. [20] developed a model incorporating similarity of inpatient records and information of the record author. They presented an idea of integrating meta-information into the abstractive summarization model on medical documents, but did not reveal how meta-information would affect the quality of the summaries.

7.3 Methods

Our method is based on the encoder-decoder transformer model. The transformer model is known for its high performance, has been widely used in recent studies, thus it is suitable for our purpose. As shown in Figure 7.1, the standard input to a transformer's encoder is created by a token sequence $T = [t_0, t_1, ..., t_i]$ and position sequence $P = [p_0, p_1, ..., p_i]$, where i is the maximum input length. The token and position sequences are converted into token embeddings E_T and positional embeddings E_P by looking up vocabulary tables. The sum of E_T and E_P is input into the model.

In this chapter, we attempt to encode meta-information to feature embeddings. We follow the segment embeddings of BERT [92] and the language embeddings of XLM [114], which provide additional information to the model. Our method is formulated as follows: Let M be feature type, $M \in \{\text{Vanilla, Hospital, Physician, Disease, Length of stay}\}$, since we set five types of features. The vanilla feature is prepared for the baseline in our experiment and to equalize the total number of parameters with the other models. Feature embeddings E_M is created by looking up the feature table $Table_M = \{m_1, m_2, ..., m_j, ..., |M|\}$, where m_j is featue value (e.g., pysician ID, disease code, etc.) and $|M|$ is the maximum number of differences in a feature. In our study, $|M|$ is set to four different values depending on features. Specifically, they are as follows.

Hospital As shown in Table 7.1, the data includes five hospital records. They were obtained mechanically from the EHR system.

Physician Physicians are also managed by IDs in the EHR systems. The data contains 4,846 physicians, but setting $|M|$ to 4,846 caused our model's training to be unstable. Therefore, we hashed the physician IDs into 485 groups containing 10 people each. Specifically, as a naive strategy, we shuffled and listed the cases within each hospital, and hashed them into groups in the order of appearance of the physician IDs. So each group has the information about the relevance of the hospitals. The reason for employing a grouping strategy is described in Appendix A.

Disease Two types of disease information exist in our EHRs: disease names and disease codes called ICD-10[1]. We did not use any disease names in the inputs for our experiment. Instead, we encoded diseases with the first three letters of the ICD-10 code, because they represent well the higher level concept. The initial three letters of the ICD-10 codes are arranged in the order of an alphabetic letter, a digit, and a digit, so there are a total of 2,600 ways to encode a disease. In our data, some ICD-10 codes were missing, although all disease names were systematically obtained from the EHR system. For such cases, we converted the disease names into ICD-10 codes using MeCab with the J-MeDic [81] (MANBYO 201905) dictionary. Also, diseases can be divided into primary and secondary diseases, but we only deal with the primary diseases.

Length of stay The length of stay can be obtained mechanically from the EHR system and the maximum value was set to 1,000 days.

We set $|M|$ for vanilla, hospital, physician, disease, and length of stay to 1, 5, 485, 2,600, and 1,000, respectively[2]. The input to our model is the sum of E_T, E_P and E_M. We also prepare an extra model with all features for our experiments. This takes all four feature embeddings (hospital, physician, disease, and length of stay) added to the encoder.

[1]For example, botulism is A05.1 in the ICD-10 code and is connected to upper category A05, "Other bacterial foodborne intoxications, not elsewhere classified".

[2]Actualy, the types of diseases and length of stay were 835 and 286, respectively. And a padding id is added.

Number of cases	24,630
Average num of words in source	1,728
Average num of words in summary	434
Number of hospitals	5
Number of physicians	4,846
Number of diseases	1,677
Number of primary diseases	835
Length of stay	
Average	21
Median	9
STD	196

Table 7.1 Statistics of our data for experiment.

7.4 Experimental setup

7.4.1 Datasets and metrics

We evaluated our proposed method on a subset of the NHO data. The statistics of our data are shown in Table 7.1[3], which includes 24,630 cases collected from five hospitals. Each case includes a discharge summary and inpatient records for the days of stay. The data are randomly split into 22,630, 1,000, and 1,000 for train, validation, and test, respectively. Summarization performances are reported in ROUGE-1, ROUGE-2, ROUGE-L [69] and BERTScore [42] in terms of F1.

7.4.2 Architectures and hyperparameters

Due to our hardware constraints we need a model that is computationally efficient, so we employed the Longformer [115] instead of the conventional Transformer. Longformer can reduce memory usage by setting window size against calculating attention. Our implementation of Longformer[4] is based on the original author's codes[5].

[3]The standard deviation of the length of stay is much higher because the data set includes extremely long stays (about 26,000 days), but we found only 12 cases with length of stay above 1,000 days.

[4]https://github.com/ken-ando/Is-In-hospital-Meta-information-Useful-for-Abstractive-Discharge-Summary-Generation

[5]https://github.com/allenai/longformer

Model	R-1	R-2	R-L	BERTScore
Longformer	10.93	1.23	9.05	63.13
w/ Hospital	13.39	1.41	10.70	65.19
w/ Physician	14.57	1.02	10.60	62.30
w/ Disease	**15.38**	**1.96**	**12.17**	**66.80**
w/ Stay length	14.61	1.25	10.63	61.94
w/ All features	13.18	0.86	10.82	61.68

Table 7.2 Performance of summarization models with different meta-information. The best results are highlighted in bold. Each score is the average of three models with different seeds.

In our model, number of layers, window size, dilation, input sequence length, output sequence length, batch size, learning rate and number of warmup steps are 8, 256, 1, 1024, 256, 4, 3e-5 and 1K, respectively. Other hyperparameters are the same as in the original Longformer, except for the maximum number of epochs is not fixed and the best epoch. It is selected for each training using the validation data based on ROUGE-1. Also, the original Longformer imports pretrained-BART parameters to initial values, but we do not use pre-trained Japanese BART in this study. We used three GeForce RTX 2080 TI for our experiments.

Our vocabulary for preparing input to Longformer is taken from UTH-BERT [93], which is pre-trained on the Japanese clinical records. Since the vocabulary of UTH-BERT is trained by WordPiece [116], we also tokenize our data with WordPiece. However, the vocabulary does not include white space and line breaks, which cannot be handled, so we add those two tokens to the vocabulary, resulting in a total size of 25,002. The vocabulary has all tokens in full characters, so we normalized full-wdith characters by converting all alphanumeric and symbolic characters to half-width for byte fallback.

7.5 Experiments and results

As shown in Table 7.2, we found that all the models with encoded medical meta-information perform better in ROUGE-1 and ROUGE-L than the vanilla Longformer. However, in BERTScore, only hospital and disease models outperform the vanilla. Specifically, disease

information is most effective, improving ROUGE-1, ROUGE-2, ROUGE-L and BERTScore by 4.45, 0.73, 3.12 and 3.77 points over the vanilla model, respectively. This seems to be because disease information and the ICD-10 ontology efficiently cluster groups with similar representations. In contrast, in ROUGE-2 and ROUGE-L, the model with physician embedding is inferior to the vanilla model. This seems to be a negative effect of grouping physicians without any consideration of their relevance. It would be better to cluster them by department, physician attributes, similarity of inpatient records, etc. Regarding low ROUGE-2 scores in all models, a previous study [20] using the English data set also reported a low ROUGE-2 score of about 5%, which may indicate an inherent difficulty in discharge summary generation. In BERTScore, the models with the physician and the length of stay did not reach the performance of the vanilla model, suggesting that the system's outputs are unnatural for humans. The model with all features performed the lowest of all models in BERTScore. The reason for the low score of the model with all features seems to be that its number of parameters in feature embedding was four times larger than that of the model with the individual feature, and the amount of training data was insufficient.

7.6 Analyzing the precisions in generated words

To analyze the influence of encoded meta-information on the outputs, we evaluate the precisions of the generated text. Specifically, we measure the probability that the generated words are included in the gold summary to investigate if the proper words are generated. Some previous studies on faithfulness, which also analyze the output of summarization, have employed words or entities [58, 117, 118]. In this study, we focused on words, not entities, because we wanted to visualize expressions that are not only nouns. The words were segmented by MeCab with the J-MeDic. For each segmented word, the numeral and symbol labels were assigned as parts of speech by MeCab, the morphological analyzer, while the disease and symptom were assigned by the J-Medic dictionary.

The results, shown in Figure 7.2, indicate that the encoded disease information leads to generate more proper disease and symptom words. This indicates that the meta-information successfully learns disease-related expressions. The encoded hospital or physician information also improved the precision of symbols generation. This suggests that different hospi-

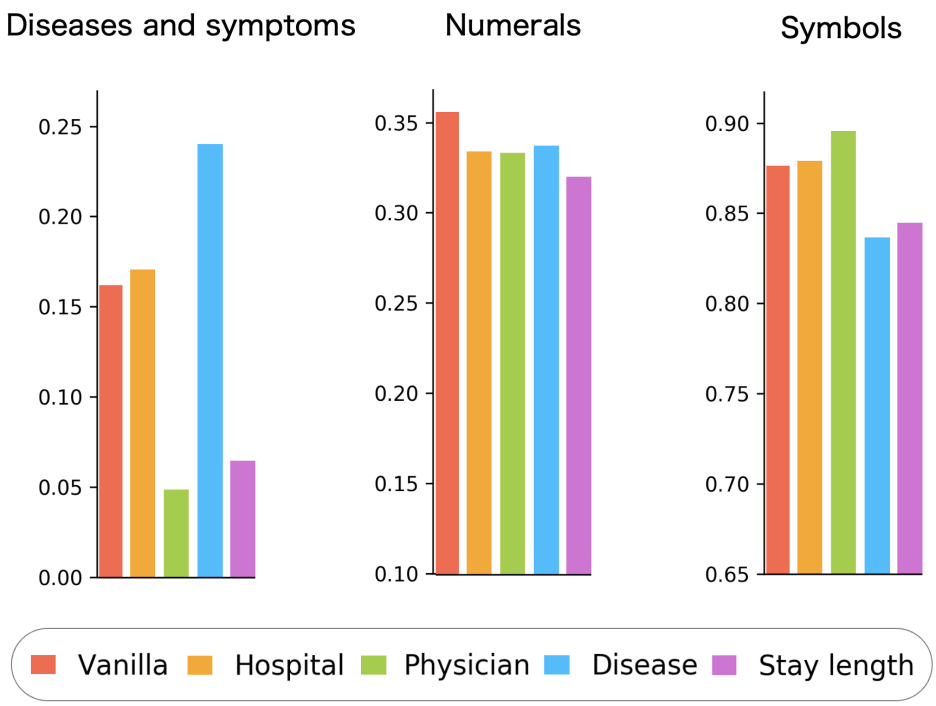


Fig. 7.2 The precisions of words in the generated summaries. The vertical axis shows the probability that the words exist in the gold summary.

tals and physicians have different description habits (e.g., bullet points such as "•", "*" and "-", punctuation such as "。" and ".", etc.), which can be grouped by meta-information.

7.7 Discussion

7.7.1 Limitations

Our limitations are that we used Japanese EHR, the limited number of tested features and not performing human evaluations. As for the efficacy of the meta-information, we believe that our results are applicable to non-Japanese, but it is left as a future work. Other meta-information may be worth verifying such as the patient's gender, age, race, religion and used EHR system, etc. It is hard to collect a large amount of medical information and process it into meta-information, so we may need to develop a robust and flexible research infrastructure to conduct a more large scale cross-sectional study in the future. In the discharge summary generation task, which demands a high level of expertise, the human evaluation

requires a lot of physicians' efforts and it is a very high cost which is unrealistic. This is a general issue in tasks dealing with medical documents, and this study also could not perform human evaluations.

7.8 Conclusion

In this chapter, we conducted a discharge summary generation experiment by adding four types of information to Longformer and verified the impact of the meta-information. The results showed that all four types of information exceeded the performance of the vanilla Longformer model, with the highest performance achieved by encoding disease information. We found that meta-information is useful for abstractive summarization on discharge summaries.

Table 7.3 Statistics on the number of cases handled by physicians. C/P denotes Cases/ Physician, which indicates how many cases an individual physician has.

Hospital	Median of C/P	Max of C/P
A	18	201
B	16	210
C	33	330
D	5	910
E	2	162

Appendix A

Method of Grouping Physician IDs

A most naive method of mapping physician IDs to features is without any grouping process. The data contains 4,846 physicians, so $|M|$ was set to 4,846. However it caused our model's training to be unstable. This might be due to the many physician IDs appearing for the first time in the test time. Table 7.3 shows the detailed number of cases handled by physicians. In all hospitals, there is a large difference between the median and the maximum of cases/physician. This indicates that a few physicians handle a large number of cases and many physicians handle fewer cases. It is impossible to avoid physician IDs first seen at test time without some process that averages the number of cases a physician holds. Due to this characteristic of our dataset, it was not suitable to use the physician IDs directly as features.

Chapter 8

Can Discharge Summaries Be Generated from Only Inpatient Records?

8.1 Introduction

There are some recent studies on the automated generation of the whole discharge summary [119, 20–25]. However, it remains an open question whether artificial intelligence can generate hospital discharge summaries from inpatient records. This is especially problematic for discharge summary generation using abstractive summarization method, since it often produces unfaithful outputs from limited sources [58–63]. In other words, in the case of discharge summary generation where information is derived from sources other than the inpatient records, if a neural model is trained only from the inpatient records, it will generate out-of-source information. To address this issue, it is important to find the source of the information expressed in the discharge summaries. If physicians rely on their memory, it would be difficult to automatically generate a discharge summary solely from inpatient records, even with a top-performing summarization technique.

Therefore, we designed the study to investigate the information sources of the discharge summaries (Fig 8.1). First, the discharge summaries were automatically split into *clinical segments* to break them down into medical semantic units. Second, medical professionals manually classified each description from discharge summaries to determine whether it originated from daily inpatient records. Using manual classification, expressions that are

completely different in appearance but semantically equivalent can be accurately identified. Finally, an in-depth analysis of the expressions in discharge summaries that could not be reconstructed from daily inpatient records was conducted. For this purpose, *clinical role labels* were used. To overcome the problems of large and strictly privacy-sensitive target data containing raw patient information, a small dataset of dummy health records was annotated, and an automatic classification model was built.

8.2 Related work

In a prior study, Maynez et al. [57] first suggested the possibility of drawing information from documents outside of sources contained in the dataset for the news domain. Also, in the simplification task, Devaraj et al. [120] explored factuality issues such as the insertion of extraneous information and the loss of key ideas in the simple text of the dataset. Mielke et al. [121] and Zhou et al. [117] investigated factuality issues in chatbots and machine translation, respectively. Seriously, prior study confirms that sources lacking key information will lead to more accelerated generating hallucinations of the model [59]. This issue can be resolved by providing additional external sources [57] or by using more stronger pre-training models to enable the integration of background knowledge [59]. Despite several studies, lack of source information is the most important factor for the hallucination. However, such issues have not been addressed in the medical domain. To the best of our knowledge, this study is the first attempt to address the hallucination problems in the summarization of clinical narratives. In the summarization task, we also first comprehensively explored potential source documents.

8.3 Datasets and preprocessing

Our target data is full NHO data, including 24,641 cases collected from five hospitals that belong to the NHO. For preprocessing, each sentence in the dataset was first split by end marks and line breaks. A primitive approach was adopted because the complex sentence-splitting model might introduce biases in subsequent analyses, as the clinical documents used in this study were noisy. Each sentence is then split into *clinical segments*. In this

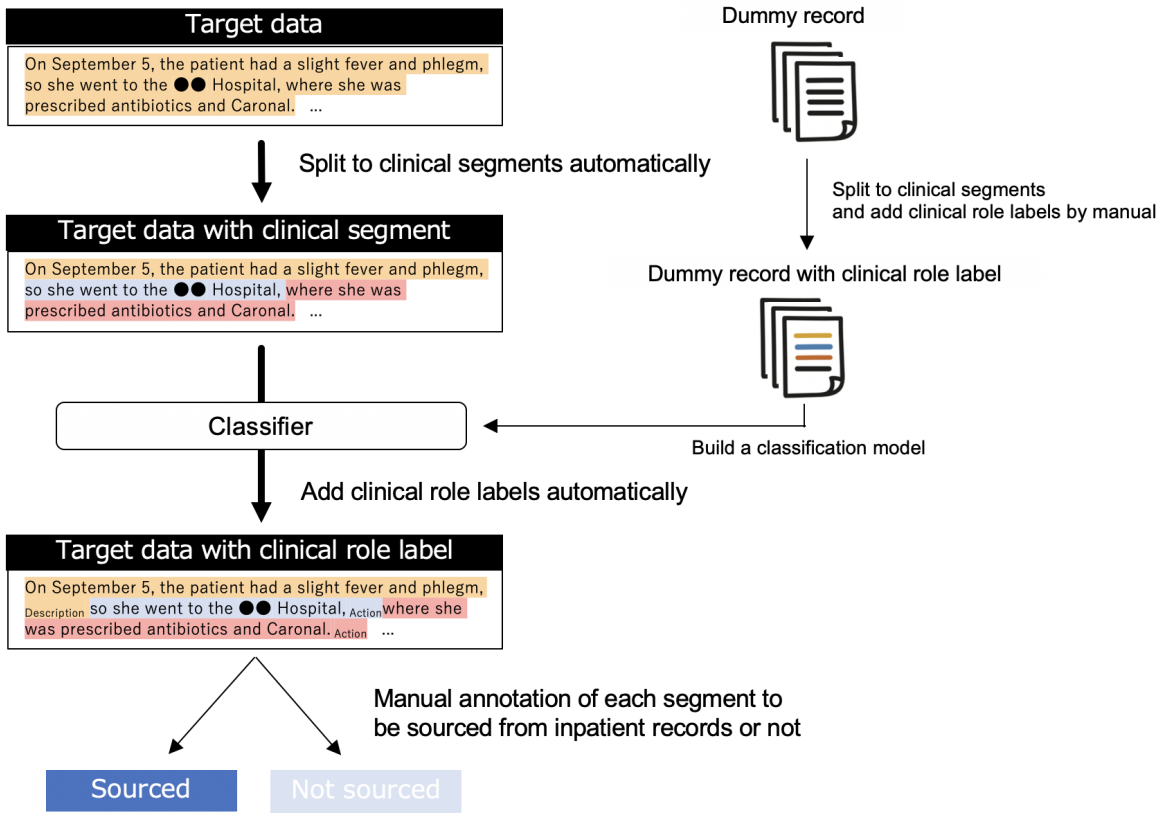


Fig. 8.1 Proposed framework of our study. The colored blocks in the dummy record represent the clinical segment developed in previous study, where the sentence is split by medical sense.

process, the model built by Chapter 4 was used for the automatic assignment of clinical segments.

8.4 Classification of unsourced segments

8.4.1 Methods

To measure the proportion of segments in the discharge summaries originating from inpatient records, a two-step approach was employed. A flowchart of the proposed process is shown in Fig 8.2. First, segments in the discharge summaries were automatically classified using a simple matching algorithm for inpatient records. If the exact segments were found

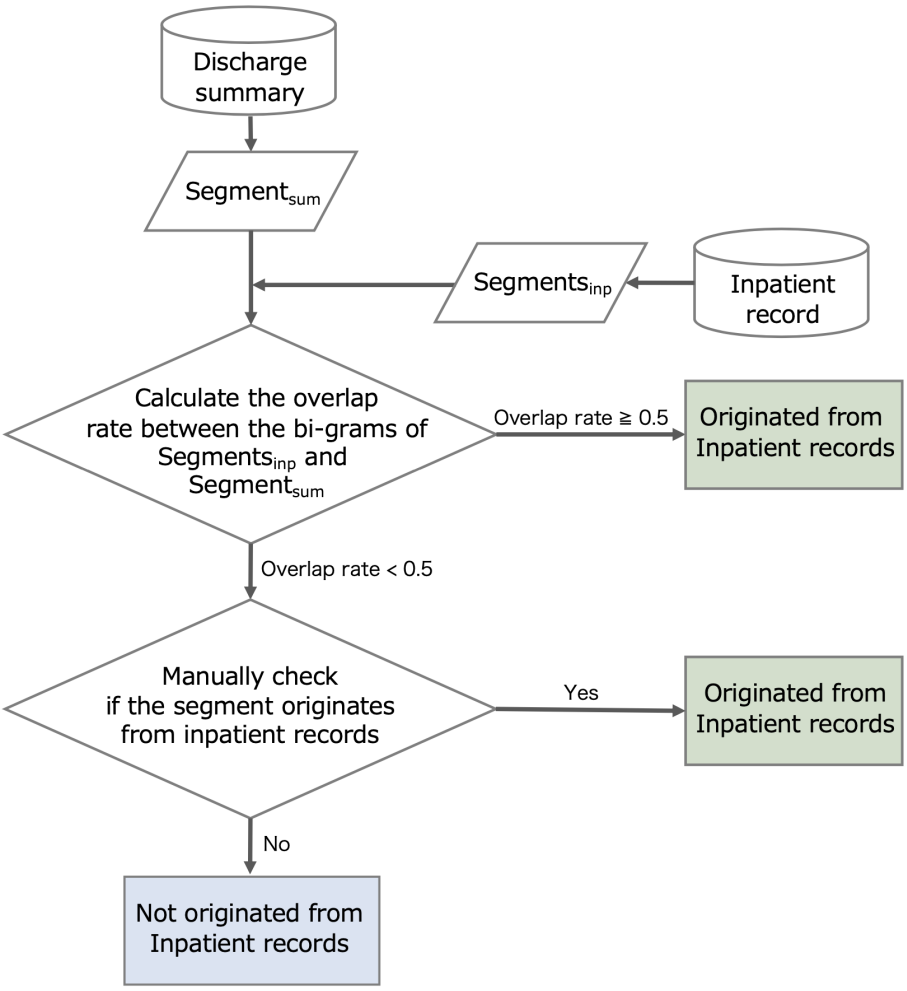


Fig. 8.2 Our annotation flowchart of the source origin. The source origin are manually decide in two steps using pre-filtering.

in the records, they were obtained from them. However, the naive algorithm cannot handle synonymous expressions, thus preventing a fully automated classification. Therefore, in the second step, this study employed the manual annotation of segments considered unsourced by automated classification. The target data comprised 772 segments extracted from 24 randomly selected documents. These documents were selected from the five hospitals in the NHO. Symbols from the system output, dates, and other symbols were excluded from this task because they were meaningless in the annotation.

Automatic filtering of unsourced segments

In the first step, word-based bi-grams were used to determine whether the segments in the summaries were sourced from inpatient records. To this end, a bi-gram set was created from all inpatient records, and a list of bi-grams from each discharge segment summary was created. Subsequently, coverage with the bi-gram set from inpatient records was measured. The bi-gram method was adopted because the distribution of the coverage ratio was closer to uniform across the entire value range (Fig 8.3a). For simplicity, the classification threshold was set to 0.5, which was validated through analysis.

Manually classification of unsourced segments

In the second step, segments with coverage ratios of less than 0.5 were manually annotated. A total of 408 segments were used for annotation. The task involved comparing each segment against inpatient records and labeling whether information in the segment was provided in the source. This task required both medical and clinical knowledge. Annotations were performed by an expert in NLP (Author K.A) and two medical professionals. To relieve the burden on annotators, the author first assigned temporary labels to all the data. Subsequently, a domain expert checked the labels and corrected them if they appeared wrong. Finally, another expert checked and fixed the labels. The inter-annotator agreement rate was 0.952, indicating the validity of the labels.

8.4.2 Classification results

Accuracy of automatic filter

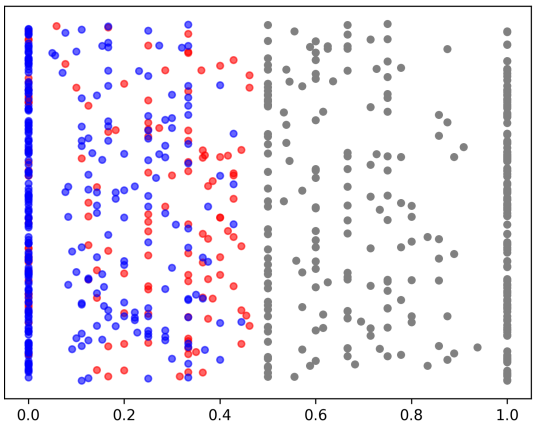
We validate the accuracy of the automatic filter we created by comparing it to manually classified results. The automatic and manual classification results for the annotations are shown in Fig 8.3a and 8.3b, respectively. The bi-gram match rate was divided into five intervals from 0 to 0.5 to confirm the validity of our threshold. The probability of the presence of unsourced segments in these intervals was measured (Fig 8.3b). This probability decreases as it approaches 0.5, with a low probability near 0.5, equal to 0.2. This indicates that our threshold of 0.5 is sufficient to cover the segments suspected to be unsourced.

Unsourced rate in clinical role and subjectivity labels

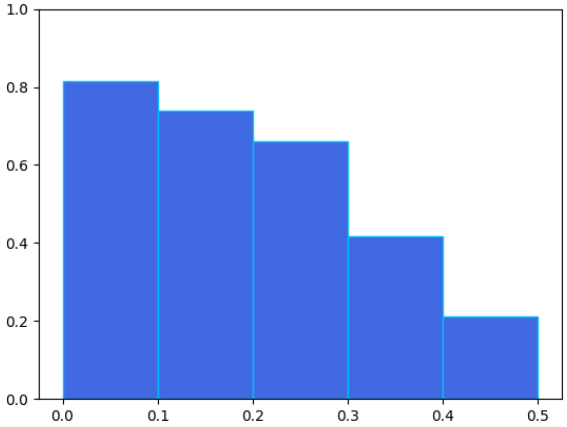
This chapter checked the amount of information in hospital discharge summaries that could not be reproduced from the inpatient records. Table 8.1 shows the detailed label results. The overall percentage of sourced segments was 61.3%, indicating that 38.7% of the information in the discharge summary was obtained from external documents other than inpatient records. In addition, the document-based unsourced rate, including at least one unsourced segment in a document, amounts to 87%. Considering the unsourced rate for each subjectivity, this study found that the unsourced probability is higher for high subjectivity than for low subjectivity. This suggests that statements involving subjectivity do not rely much on documents and are written by physicians themselves. For clinical role labels, *diag* and *probable* were relatively high. This indicates that the core of medical practice, such as diagnosis and prediction based on facts, is often generated in discharge summaries.

Unsourced rate in sections of discharge summary

Discharge summaries typically comprise descriptions of pre-hospital episodes and in-hospital information. The "pre-hospital" part consists of past medical history, a history of present illness, and results of examinations at the time of admission, whereas the "in-hospital" part comprises all patient descriptions obtained after admission. Table 8.2 summarizes the unsourced and high subjectivity rates in the pre-hospital and in-hospital settings. The unsourced rates for the "pre-hospital" and "in-hospital" parts are 0.434 and 0.318, respectively, illustrating the higher rate in the "pre-hospital" part. This is plausible because the hospitals that participated in this survey were central hospitals, and most patients visited them by referral. These hospitals had referral letters and past clinical records that could be used for summarizing inpatient records (more details are provided in Section 8.5). Additionally, the pre-hospital section had a lower percentage of high subjectivity segments. This reflects that the content of this section is mainly patient history. In contrast, the in-hospital section had a higher percentage of high-subjectivity segments, reflecting content such as speculation, planning, and diagnosis, which generally occur during hospitalization.



(a) Distribution of origin rates using bi-grams from the randomly sampled data. Red, blue, and gray dots are sourced, unsourced, and filtered out segments, respectively. Note that symbols and segments categorized as middle subjectivity are excluded. The y-axis values were randomly generated from a uniform distribution of visibility.



(b) Proportion of unsourced segments appearing in manually annotated data. The y-axis is the value averaged every 0.1 steps for segments with origin rates less than 0.5, as shown in Fig 8.3a.

Fig. 8.3 Origin rate of segments in discharge summaries against the inpatient records.

Table 8.1 Rate of unsourced segments in detailed labels. Because the clinical role and the subjectivity labels are automatically added as different tasks, the subjectivity label is not a weighted average of the clinical role labels. In contrast, "All" is a weighted average of low and high subjectivity.

Subjectivity	Clinical Role	Unsourced rate			
Low	Description	0.369	0.376	All	0.387
	Action	0.403			
	Others	0.304			
High	Evaluation	0.487	0.439		
	Diag	0.529			
	Plan	0.422			
	Nonfact	0.429			
	Probable	0.583			

8.5 Analyzing the origin of unsourced information

The results suggest that physicians refer to various documents and inpatient records when preparing discharge summaries. This section identifies the sources of information that appear in the discharge summaries in addition to the inpatient records. To this end, 14 labels

Table 8.2 Rate of unsourced and high subjectivity segments in two sections. The sections "Pre-hospital" and "In-hospital" include descriptions of patients before and after admission.

	Unsourced rate	High subjectivity rate
Pre-hospital	0.434	0.130
In-hospital	0.318	0.235

were developed to classify sources of information: *patient referral documents, outpatient clinical records, emergency room records* and *patient's past clinical records* (which cannot be categorized in other labels of past records and mainly include patient's past inpatient records) are descriptions of past history. *Prescriptions, nursing records, examination results, ECG reports, rehabilitation reports, surgical operation notes* and *anesthesia records* are descriptions of the current admission. *Other patients' clinical records, other documents*, and *information not derived from any documents* (i.e., a physician's memory or inference) are the descriptions of the others.

For example, drug information written in quantitative form was labeled as *prescriptions*. Events during rehabilitation were labeled as *rehabilitation reports*. The admission episodes of patients from the emergency department were labeled as *emergency room records*. Doctors' impressions and inferences are labeled *not derived from any documents*. These labels may appear lengthy; however, they facilitate further insight into the origin of the information written in the discharge summaries. Expressions labeled *not derived from any documents* included information that could not be recorded during the hospital stay, such as descriptions of the times of discharge and post-discharge schedules. They also included physicians' perspectives on diagnostic approaches and treatment options. They may also contain excessive abbreviations for the hospital stay such as "no significant change," descriptions of normal conditions such as "able to eat," and omission of details of standardized protocols such as "fluids and antibiotics." Annotation was performed by including two medical professionals, as described in Section 8.4.1. The inter-annotator agreement rate is 0.938. Such a high score indicates the objectivity of the designed annotation labels with a physician.

8.5.1 Statistics of external sources

The statistical results are listed in Table 8.3. Overall, this study found that 43.3% of the new information was derived from the patient's past clinical records. When patient referral documents are included, the coverage of the new information is 61.7%, which suggests that the availability of these two types of documents can complement 61.7% of the missing information.

As a general trend, there were no significant differences between the two groups in the low and high subjectivity columns. However, in the *not derived from any documents* row, high subjectivity segments indicate a higher proportion (18.2%) than low subjectivity segments (8.8%). This indicates that, when physicians write summaries, they often add information based on reasoning rather than memory.

A characteristic difference was observed in the prehospital and in-hospital periods. The top four documents in the prehospital section describe the history of patient admission. This is the natural result of this function. Among these, the patient's past clinical records showed a significantly high rate. This indicates that, in the hospitals studied in this chapter, a large number of admitted patients were former patients rather than referrals. However, information not derived from any document was the most common item in the in-hospital section. This is also a natural function of the section because it is a place to fill in doctors' perspectives.

Table 8.3 List of source documents that the annotators selected for each piece of information labeled as *not sourced from inpatient records*. The numbers indicate the percentage of external documents in each section. *Low subj*, *High subj*, *Pre-hosp*, and *In-hosp* in the table show the distribution of assumed external sources for segments classified as unsourced. Because it has a multi-label structure, each segment may have multiple source labels, and the percentile is calculated against the total number of assigned labels.

	Documents	All (%)	Low subj (%)	High subj (%)	Pre-hosp (%)	In-hosp (%)
Past history	Patient referral documents	18.4	19.5	14.5	21.4	12.3
	Outpatient clinical records	6.6	5.8	9.1	9.1	0.8
	Emergency room records	3.9	4.5	2.7	5.5	1.1
	Patient's past clinical records	43.3	43.6	42.7	56.0	16.1
Current admission	Prescriptions	2.0	2.2	0.9	0.0	6.3
	Nursing records	1.5	1.8	0.0	0.0	4.5
	Examination results	5.7	5.8	6.4	0.5	17.6
	ECG reports	0.0	0.0	0.0	0.0	0.0
	Rehabilitation reports	1.5	1.8	0.0	0.0	4.5
	Surgical operation notes	0.7	0.4	1.8	0.0	2.3
	Anesthesia records	0.0	0.0	0.0	0.0	0.0
Others	Other patients' clinical records	3.5	4.2	0.0	5.0	0.0
	Other documents	2.0	1.6	3.6	0.8	4.5
	Not derived from any documents	10.9	8.8	18.2	1.7	29.9

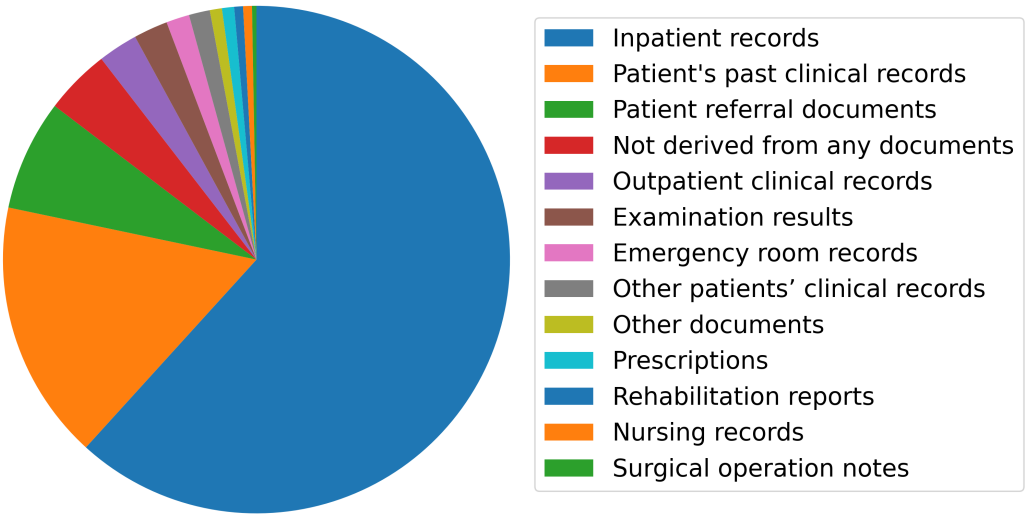


Fig. 8.4 Breakdown of the information source in discharge summaries.

8.5.2 Interpretation and generalizability of the results

The analysis indicates a breakdown of the origin of information that appears in the discharge summaries (Fig 8.4). Information derived from inpatient records constituted 61% of the discharge summaries. The next most common source was the patient's past clinical records (17%), and the third most common source was patient referral (7%) of the documents. To this point, 85% of the information in discharge summaries originates from documents associated with the patient. The fourth most common source was *not derived from any documents*, which explained 4% of the information sources.

As illustrated, physicians can refer to various documents, in addition to inpatient records, when they write discharge summaries. In this analysis, the number of target documents was limited because manual annotation was performed for accuracy. Although the analysis reveals that a substantial proportion of the contents in discharge summaries originate from sources other than inpatient records, the generalizability of the results should be verified. For this purpose, the variance between hospitals was analyzed and is listed in Table 8.4.

Focusing on the differences between hospitals, this study found that the unsourced rates differ greatly across hospitals. This difference can be ascribed to design differences in the documentation of electronic health record system vendors. These results suggest that hospital-specific bias must be considered when analyzing clinical narratives. Source avail-

Table 8.4 Rate of unsourced and high subjectivity segments in institutions. Roman numerals indicate the five surveyed hospitals.

	Hospital				
	I	II	III	IV	V
Unsourced rate	0.596	0.289	0.231	0.461	0.360
High subjectivity rate	0.112	0.084	0.259	0.203	0.292

ability may affect the way physicians write discharge summaries. Furthermore, the variation in the high subjectivity rates was limited, suggesting that the clinical reasoning process by physicians follows similar patterns across different types of facilities. In either case, the limited amount of data is a limitation of this study, and extending the studies to various types of institutions, probably with automated classification, would be valuable.

8.6 Discussion

This chapter investigated the origin of the information that appears in discharge summaries to evaluate the possibility of an automated summarization of inpatient records. The analysis results indicate that only 61% of the total information is derived from inpatient records, and 39% of the information originates from sources other than records. Manual evaluation by medical professionals identified past medical documents as the most common source of external information, such as patient referral documents and patient's past clinical records. These two types of source documents accounted for 62% of the missing information. This study also found that 11% of the information contained speculation and post-discharge plans that were not derived from documents.

Previous studies indicate that automated summarization using a trained model from inputs with incomplete information for the target summary leads to hallucinations [59]. A previous study on news summarization using a dataset with an unsourced rate of 73% in document-based counts yielded a high incidence of extrinsic hallucinations [57]. In news summarization, the content is created from the source and supplemented by other news articles or common sense, which explains extrinsic hallucinations [59]. Our study revealed that

the unsourced rate of expressions in the discharge summaries was 38.4% for the segment-oriented count and 87% for the document-based count. Therefore, if the dataset containing only inpatient records is used in the summarization of inpatient records, a higher incidence of hallucinations would be caused by the high unsourced rate. Considering the nature of healthcare, this result is unacceptable.

Clinical document summarization is inherently a multidocument summarization. Approximately 62% of the missing information could be generated if the patient's past clinical records (43.3%) and patient referral documents (18.4%) were available. However, 11% of the information depends on the physician's memory and clinical reasoning, and this portion is difficult to generate automatically. Therefore, automatic high-quality summarization using machine learning is considered infeasible, and machine summarization with a human post-editing process is the best solution for this problem.

A limitation of the present analysis lies in the volume of the target documents manually annotated and in the representativeness of the sampled target. A more thorough and detailed analysis might result in different statistics, and language differences must also be considered when applying the results to other languages. However, differences that may emerge in the additional analysis would be minor compared to the technical contributions of the present study. Extending the source material beyond inpatient records is necessary for the automated generation of discharge summaries. It is also necessary to improve the accuracy of abstractive summarization and present a draft that effectively elicits physicians' reasoning and memory.

8.7 Conclusion

This study investigated whether artificial intelligence and natural language processing can automatically generate discharge summaries. The results indicate that the majority of the discharge summaries originated from sources other than patient records. The patients ' past clinical records and patient referral documents were the most and second-most external sources, respectively. This study found that a certain amount of external information was generated by the physician's memory and clinical reasoning. The analysis suggests that the automated generation of discharge summaries is impossible using a naive collection

of inpatient records. The generation of discharge summaries involves multiple document summarizations and clinical reasoning with undocumented information by physicians in charge of hospitalized care.

Undoubtedly, the automatic generation of discharge summaries could reduce the heavy burden on medical practice; thus, development in this field is highly desirable. Our results suggest that research efforts must be made to establish an optimal interaction between humans and machines for the efficient authoring of discharge summaries by incorporating generated drafts and post-editing assistance.

Chapter 9

Conclusion

9.1 Conclusion

This thesis examined the following topics related to the generation of the discharge summary. First, since we did not find a parallel corpus of discharge summaries and hospitalization records in Japanese, we created a new dummy record by the physician. This is closer to the NHO data than the existing dummy medical record data set and is shown to be potentially more realistic.

Second, we designed a new linguistic unit to cover medical meaning. The units annotated by medical professionals are shorter than sentences and longer than clauses. We also developed an automatic splitter and showed that it could achieve high performance automatic splitting by using a neural model.

Third, we defined and annotated what clinical roles are represented by the clinical segment. We assigned a two-layered structure of subjectivity and detailed labels, and found that many segments represent objective facts. In addition, the classifier was trained using multi-task learning, which showed that it could automatically classify with good overall accuracy.

Fourth, we investigated what linguistic units are the most efficient inputs for generating discharge summaries using the extractive summarization method. Experimental results using the sentence, cause, and clinical segment showed that the clinical segment performed

best, suggesting that it may be better to use units that cover medical meaning for summarizing medical documents.

Fifth, we investigated whether medical meta-information is useful for generating a discharge summary using abstractive methods. We used four types of information: hospital, physician, disease, and length of stay, and found that inserting the disease information into the model was the best. It also yielded better accuracy in representing diseases and symptoms in the output than the vanilla model. These results indicate that we may better utilize the various meta-information obtained from outside of the source, depending on the purpose.

Finally, we investigated whether discharge summaries could be generated from only inpatient records. The results showed that 39% of the discharge summaries were obtained from outside of the inpatient record, suggesting that it is an unrealistic setting for the discharge generation task to use only the inpatient record as the source document, as many previous studies have done. In addition, few of the information was derived from physicians' memories, indicating the need for a variety of sources.

9.2 Limitations

The following is a summary of the limitations of this thesis.

Language dependent

Japanese grammar and Japanese medical practices are very different from those of European languages, and there can be differences in the description, summarization, and evaluation processes. Accordingly, this pipeline using extractive method might be applicable only to Japanese clinical setting.

In the clinical segment and clinical role label were defined for Japanese, only labeled corpus for Japanese exists, so they are not naively applicable to other languages. However, the idea of capturing medical concepts may be useful for other languages or cultures.

Culture dependent

In some countries with different cultural background, *dictation* is used in clinical records and their summaries, and it is reported that 62% of cases are created by dictation [101]. In this regard, Japanese hospitals do not use dictation to produce discharge summaries, which could result in frequent copying and pasting from sources to summaries, and in this study, 20-32% of discharge summaries were created by copying and pasting. This custom could have contributed to using extractive texts in the discharge summaries in Japan. Just for reference, compare the copy-paste rate for discharge summaries in the U.S. A study reported that 8% of dictated summaries [122] and another study reported that 54% of handwritten summaries were copy-pastes [123] (87% of the documents in our data). Note, this ratio depends on the hospital and the definition of copy-and-paste [124].

This cultural difference may have influenced the experiments conducted in Chapters 6 and 7. In particular, the analysis in Chapter 8 may yield different results in countries having different medical cultures, because the statistics are different even for different hospitals within a country. The analysis of the influence of this customary difference is left for future work.

Data-scale dependent

A limitation of the present results lies in the volume of the target documents we annotated mannually, and in the representativeness of the target we sampled. In particular, the annotations were difficult to conduct on a larger scale due to the high cost, which required the cooperation of healthcare professionals. Moreover, this study focused only on inpatient medical records, but other experiments using many different types of healthcare documents could be performed. Although our study used the largest multi-institutional health records archive in Japan, more researches at various institutions would be preferable to confirm the generalizability of our results. In addition, we may need to develop a robust and flexible research infrastructure to conduct a more large-scale cross-sectional study in the future.

The annotation of clinical segment and clinical role label are deeply influenced by the size of the data. Also, in the use of meta-information in Chapter 7, collecting more and varied meta-information such as race, gender, religion, etc. will lead to stronger conclusions.

In Chapter 8, more origins annotations and larger data sets with more variety of hospitals and documents will lead to greater contributions.

In addition, it is important to experiment with data that include a wider variety of diseases. Analyzing the differences between diseases is important, since some diseases are for examination purposes and others are for long-term stays.

Human evaluation

In the discharge summary generation task, which demands a high level of expertise, the human evaluation requires a lot of physicians' efforts and it is a very high cost which is unrealistic. This is a general issue in tasks dealing with medical documents, and this study also could not perform human evaluations.

This human evaluation issue can be raised for most all automation experiments, it is especially important to manually evaluate the discharge summaries generated in Chapters 6 and 7. However, this is still an understudied area, including policies on what discharge summaries are good, and further research is needed. In particular, the manual evaluations of gold discharge summaries are important for the future discharge summary generation task.

However, differences that may emerge in the additional analyses or experiments would be minor, compared with the contribution of our studies.

9.3 Future work

This section shows the main future works of this thesis. Regarding the clinical segment we developed, we suggested that it can capture medical meanings more effectively than existing units, but since we only evaluated it through the summarization task, it is necessary to investigate how well it captures medical meanings actually. For the clinical role label, we showed the possibility to analyze clinical documents automatically and in depth, but it needs to be evaluated whether it can be used for quantitative evaluation of various medical documents.

More advanced approaches can be applied for the summarization experiment we used the inpatient records as a single document. For single document summarization, we can

consider a pipeline that generates a summary based on important expressions, with extractive summarization as the first stage, which is often employed in very long summarization tasks [20, 125, 126]. For the inpatient records considered as multiple documents, a method to extract important information by extractive summarization [21] is also possible, as well as a method to capture the properties of documents in a single model [127].

Towards application to practical use

In the future, similar to other medical AI tasks such as radiology, it could be performed in collaboration with physicians. Actually, in cancer detection, AI and human collaboration is better performance than independent screening [128–130]. We can envision an idea where an initial draft is created by the AI and subsequently edited by a physician in discharge summary generation. Therefore, it is necessary to conduct research to establish the optimal interaction between humans and AI, which is still unexplored in the discharge summary generation. Studies in other medical domains have explored the conditions under which physicians trust AI prediction outputs and the importance of developing interfaces in concert with physicians, so we can draw on such studies to guide our work. In the case of a discharge summary generation system, we could display the source medical documents of information in representations of the output, or evidence of generation. This topic was an additional analysis we wanted to address in Chapter 8, but it was not possible due to the lack of data variety.

Other our insights for the future are as follows. Medical tasks require more careful output, i.e., more highly precision (ratio of true positive to false positive), because errors in the healthcare document have direct impacts on patient health. Therefore, it is necessary to build a highly reliable model with a focus on precision. To enhance reliability in the model's predictive output, it is expected to provide the evidence for its decisions, and this is already being actively studied as the explainable AI [131, 132].

In addition, the issue of physicians' trust for the model, this is also related to the hallucination. The hallucination that weird information is mixed into the generated summaries is a major clinical problem, and may lead to a loss of trust. This can be relieved by showing the evidence of generation. It is also important to evaluate separately the factual hallucinations that good to occur and the non-factual hallucinations that are not good to occur [57]. We

can also indicate the confidence level of the generated text. This has already been addressed in many studies as related to the calibration [133–135] and is a realistic solution.

Besides the performance aspect of the model, another challenge is the personalization for physicians. As pointed out in Chapter 7, it is necessary to pay attention to the habits of physicians. There are ways to address this challenge, such as requiring expressions that physicians must keep during a patient's stay or handling physician information like Chapter 7. These strategies will increase physicians' reliability of the model.

Finally, regarding the need for undocumented knowledge of physicians to generate discharge summaries, as shown in Chapter 8. This may be resolved in the future with very large data sets and models. In recent studies, a single giant model such as ChatGPT [1] or PaLM [136] can solve a variety of tasks, and they are suggested to contain a variety of abstracted knowledge such as world knowledge. Some giant models also exist for English medical purpose, such as GatorTron [137] or Med-PaLM [138], and they can also solve a variety of tasks. GatorTron was trained on billions of parameters using a corpus containing 90 billion words. In Japanese, a similar model could be constructed in terms of data size, but there are no computational resources available to train such a large model at the institution where the data exists (it probably cannot be moved outside due to data privacy issues). A proposed solution to the privacy issue is to use the federated learning [139, 140], in which a model is learned at a central facility in collaboration with many hospitals. This method does not risk personal information because it trains the model at each independent hospital and uses only the results of training. The next challenge seems to be these issues.

[1] https://openai.com/blog/chatgpt

Ethical Considerations

On this research infrastructure, informed consent and patient privacy are ensured in the following manner. At the national hospitals, notices about their policy and the EHR data usage are posted in their facilities. The patients who disagree with the policies are supposed to notify the hospital by an opt-out form, to be excluded from the archive. Likewise, minors and their parents can turn in the opt-out form, at will. To conduct a study on the archive, researchers must submit their research proposals to the institutional review board. Once the study is approved, the data are extracted from NCDA, and anonymized to construct a dataset for further analysis. The data are accessible only in a secured room at the NHO headquarters, and only statistics are allowed to be carried out of the secured room, for protection of patients' privacy.

The analysis of this research was conducted under the IRB approval (IRB Approval No.: Wako3 2019-22) of the Institute of Physical and Chemical Research (RIKEN), Japan, which has a collaboration agreement with the National Hospital Organization.

List of Publications

Journal Papers

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, Yuji Matsumoto. Exploring optimal granularity for extractive summarization of unstructured health records: Analysis of the largest multi-institutional archive of health records in Japan. PLOS Digital Health. 2022;1(9):1–19. doi:10.1371/journal.pdig.0000099.

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, Yuji Matsumoto. Is Artificial Intelligence Capable of Generating Hospital Discharge Summaries from Inpatient Records? PLOS Digital Health. 2022;1(12):1-21. doi:10.1371/journal.pdig.0000158.

Conference Paper

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, Yuji Matsumoto. Discharge Summary Generation with In-hospital Meta-information. Proceedings of the 27th International Conference on Technologies and Applications of Artificial Intelligence. 2022;.

References

[1] Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. The Annals of Family Medicine. 2017;15(5):419–426. doi:10.1370/afm.2121.

[2] Leslie Kane MA. Medscape Physician Compensation Report 2019; 2019 [cited 2021 Aug 6]. Available from: https://www.medscape.com/slideshow/2019-compensation-overview-6011286.

[3] Ammenwerth E, Spötl HP. The Time Needed for Clinical Documentation versus Direct Patient Care. A Work-sampling Analysis of Physicians' Activities. Methods of Information in Medicine. 2009;48(01):84–91. doi:10.3414/me0569.

[4] Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury. Nature. 2019;572(7767):116–119. doi:10.1038/s41586-019-1390-1.

[5] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for Beat-to-beat Assessment of Cardiac Function. Nature. 2020;580(7802):252–256. doi:10.1038/s41586-020-2145-8.

[6] Lu Q, Nguyen TH, Dou D. Predicting Patient Readmission Risk from Medical Text via Knowledge Graph Enhanced Multiview Graph Convolution. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021; p. 1990–1994.

[7] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease Variant Prediction with Deep Generative Models of Evolutionary Data. Nature. 2021;599(7883):91–95. doi:10.1038/s41586-021-04043-8.

[8] Bastani H, Drakopoulos K, Gupta V, Vlachogiannis I, Hadjicristodoulou C, Lagiou P, et al. Efficient and Targeted COVID-19 Border Testing via Reinforcement Learning. Nature. 2021;599(7883):108–113. doi:10.1038/s41586-021-04014-z.

[9] Divya S, Indumathi V, Ishwarya S, Priyasankari M, Devi SK. A Self-diagnosis Medical Chatbot using Artificial Intelligence. Journal of Web Development and Web Designing. 2018;3(1):1–7.

[10] Chen AI, Balter ML, Maguire TJ, Yarmush ML. Deep Learning Robotic Guidance for Autonomous Vascular Access. Nature Machine Intelligence. 2020;2(2):104–115. doi:10.1038/s42256-020-0148-7.

[11] Willyard C. Can AI Fix Medical Records? Nature. 2019;576(7787):S59–S59. doi:10.1038/d41586-019-03848-y.

[12] Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018. p. 1101–1111.

[13] Hodgson T, Magrabi F, Coiera E. Efficiency and Safety of Speech Recognition for Documentation in the Electronic Health Record. Journal of the American Medical Informatics Association. 2017;24(6):1127–1133. doi:10.1093/jamia/ocx073.

[14] Dong Y, Wang S, Gan Z, Cheng Y, Cheung JCK, Liu J. Multi-Fact Correction in Abstractive Text Summarization. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020; p. 9320–9331. doi:10.18653/v1/2020.emnlp-main.749.

[15] Cao M, Dong Y, Wu J, Cheung JCK. Factual Error Correction for Abstractive Summarization Models. Proceedings of the 2020 Conference on Empirical Methods in

Natural Language Processing. 2020; p. 6251–6258. doi:10.18653/v1/2020.emnlp-main.506.

[16] Haonan W, Yang G, Yu B, Lapata M, Heyan H. Exploring Explainable Selection to Control Abstractive Summarization. Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence. 2021;(15):13933–13941.

[17] Ma Y, Lan Z, Zong L, Huang K. Global-aware Beam Search for Neural Abstractive Summarization. Advances in Neural Information Processing Systems 34. 2021;34:16545–16557.

[18] Jing B, You Z, Yang T, Fan W, Tong H. Multiplex Graph Neural Network for Extractive Text Summarization. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021; p. 133–139. doi:10.18653/v1/2021.emnlp-main.11.

[19] Kwon J, Kobayashi N, Kamigaito H, Okumura M. Considering Nested Tree Structure in Sentence Extractive Summarization with Pre-trained Transformer. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021; p. 4039–4044. doi:10.18653/v1/2021.emnlp-main.330.

[20] Diaz D, Cintas C, Ogallo W, Walcott-Bryant A. Towards Automatic Generation of Context-Based Abstractive Discharge Summaries for Supporting Transition of Care. AAAI Fall Symposium 2020 on AI for Social Good. 2020;.

[21] Shing HC, Shivade C, Pourdamghani N, Nan F, Resnik P, Oard D, et al. Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes. ArXiv. 2021;abs/2104.13498. doi:10.48550/arXiv.2104.13498.

[22] Adams G, Alsentzer E, Ketenci M, Zucker J, Elhadad N. What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021; p. 4794–4811. doi:10.18653/v1/2021.naacl-main.382.

[23] Moen H, Heimonen J, Murtola LM, Airola A, Pahikkala T, Terävä V, et al. On Evaluation of Automatically Generated Clinical Discharge Summaries. Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics. 2014;1251:101–114.

[24] Moen H, Peltonen LM, Heimonen J, Airola A, Pahikkala T, Salakoski T, et al. Comparison of Automatic Summarisation Methods for Clinical Free Text Notes. Artificial Intelligence in Medicine. 2016;67:25–37. doi:10.1016/j.artmed.2016.01.003.

[25] Alsentzer E, Kim A. Extractive Summarization of EHR Discharge Notes. ArXiv. 2018;abs/1810.12085. doi:10.48550/arxiv.1810.12085.

[26] Zhou Q, Wei F, Zhou M. At Which Level Should We Extract? An Empirical Analysis on Extractive Document Summarization. Proceedings of the 28th International Conference on Computational Linguistics. 2020; p. 5617–5628. doi:10.18653/v1/2020.coling-main.492.

[27] Cho S, Song K, Li C, Yu D, Foroosh H, Liu F. Better Highlighting: Creating Sub-Sentence Summary Highlights. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020; p. 6282–6300. doi:10.18653/v1/2020.emnlp-main.509.

[28] Erkan G, Radev DR. LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research. 2004;22(1):457–479.

[29] Mihalcea R, Tarau P. TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004; p. 404–411.

[30] Cohan A, Dernoncourt F, Kim DS, Bui T, Kim S, Chang W, et al. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018; p. 615–621. doi:10.18653/v1/N18-2097.

[31] Nallapati R, Zhou B, dos Santos C, Gucehre C, Xiang B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016; p. 280–290. doi:10.18653/v1/K16-1028.

[32] Grusky M, Naaman M, Artzi Y. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018; p. 708–719. doi:10.18653/v1/N18-1065.

[33] Narayan S, Cohen SB, Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018; p. 1797–1807. doi:10.18653/v1/D18-1206.

[34] Cohn TA, Lapata M. Sentence compression as tree transduction. Journal of Artificial Intelligence Research. 2009;34:637–674. doi:10.1613/jair.2655.

[35] Barzilay R, McKeown KR. Sentence Fusion for Multidocument News Summarization. Computational Linguistics. 2005;31(3):297–328. doi:10.1162/089120105774321091.

[36] Filippova K, Strube M. Sentence Fusion via Dependency Graph Compression. 2008; p. 177–185.

[37] Tanaka H, Kinoshita A, Kobayakawa T, Kumano T, Katoh N. Syntax-Driven Sentence Revision for Broadcast News Summarization. 2009; p. 39–47.

[38] Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks; 2014.

[39] Sakishita M, Kano Y. Inference of ICD Codes from Japanese Medical Records by Searching Disease Names. Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP). 2016; p. 64–68.

[40] Luo J, Xiao C, Glass L, Sun J, Ma F. Fusion: Towards Automated ICD Coding via Feature Compression. Findings of the Association for Computational Linguistics. 2021; p. 2096–2101. doi:10.18653/v1/2021.findings-acl.184.

[41] Deznabi I, Iyyer M, Fiterau M. Predicting In-hospital Mortality by Combining Clinical Notes with Time-series Data. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021; p. 4026–4031. doi:10.18653/v1/2021.findings-acl.352.

[42] Zhang X, Dou D, Wu J. Learning Conceptual-Contextual Embeddings for Medical Text. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34(05):9579–9586. doi:10.1609/aaai.v34i05.6504.

[43] Komaki S, Muranaga F, Uto Y, Iwaanakuchi T, Kumamoto I. Supporting the Early Detection of Disease Onset and Change Using Document Vector Analysis of Nursing Observation Records. Evaluation & the Health Professions. 2021;44(4):436–442. doi:10.1177/01632787211014270.

[44] Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting Inpatient Falls Using Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study. JMIR Medical Informatics. 2020;8(4):e16970. doi:10.2196/16970.

[45] Katsuki M, Narita N, Matsumori Y, Ishida N, Watanabe O, Cai S, et al. Preliminary Development of a Deep Learning-based Automated Primary Headache Diagnosis Model Using Japanese Natural Language Processing of Medical Questionnaire. Surgical neurology international. 2020;11. doi:10.25259/SNI_827_2020.

[46] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Ohe K. TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification. Proceedings of the BioNLP 2009 Workshop. 2009; p. 185–192.

[47] Reeve LH, Han H, Brooks AD. The Use of Domain-Specific Concepts in Biomedical Text Summarization. Information Processing & Management. 2007;43(6):1765–1776. doi:10.1016/j.ipm.2007.01.026.

[48] Gurulingappa H, Mateen-Rajpu A, Toldo L. Extraction of Potential Adverse Drug Events from Medical Case Reports. Journal of biomedical semantics. 2012;3(1):1–10. doi:10.1186/2041-1480-3-15.

[49] Mashima Y, Tamura T, Kunikata J, Tada S, Yamada A, Tanigawa M, et al. Using Natural Language Processing Techniques to Detect Adverse Events from Progress Notes due to Chemotherapy. Cancer Informatics. 2022;21. doi:10.1177/11769351221085064.

[50] Liang J, Tsou CH, Poddar A. A Novel System for Extractive Clinical Note Summarization using EHR Data. Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019; p. 46–54. doi:10.18653/v1/W19-1906.

[51] Lee SH. Natural Language Generation for Electronic Health Records. NPJ digital medicine. 2018;1(1):1–7. doi:10.1038/s41746-018-0070-0.

[52] MacAvaney S, Sotudeh S, Cohan A, Goharian N, Talati I, Filice RW. Ontology-Aware Clinical Abstractive Summarization. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019; p. 1013–1016. doi:10.1145/3331184.3331319.

[53] Liu X, Xu K, Xie P, Xing E. Unsupervised Pseudo-labeling for Extractive Summarization on Electronic Health Records. Machine Learning for Health (ML4H) Workshop at NeurIPS 2018. 2018;.

[54] Hunter J, Freer Y, Gatt A, Logie R, McIntosh N, Van Der Meulen M, et al. Summarising Complex ICU Data in Natural Language. AMIA annual symposium proceedings. 2008;2008:323.

[55] Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, et al. Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. Artificial Intelligence. 2009;173(7):789–816. doi:10.1016/j.artint.2008.12.002.

[56] Goldstein A, Shahar Y. An Automated Knowledge-based Textual Summarization System for Longitudinal, Multivariate Clinical Data. Journal of Biomedical Informatics. 2016;61:159–175. doi:10.1016/j.jbi.2016.03.022.

[57] Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020; p. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

[58] Zhao Z, Cohen SB, Webber B. Reducing Quantity Hallucinations in Abstractive Summarization. Findings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020; p. 2237–2249. doi:10.18653/v1/2020.findings-emnlp.203.

[59] Xu X, Dušek O, Narayan S, Rieser V, Konstas I. MiRANews: Dataset and Benchmarks for Multi-Resource-Assisted News Summarization. Findings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021; p. 1541–1552.

[60] Chen S, Zhang F, Sone K, Roth D. Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021; p. 5935–5941. doi:10.18653/v1/2021.naacl-main.475.

[61] Aralikatte R, Narayan S, Maynez J, Rothe S, McDonald R. Focus Attention: Promoting Faithfulness and Diversity in Summarization. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021; p. 6078–6095. doi:10.18653/v1/2021.acl-long.474.

[62] Cao S, Wang L. CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021; p. 6633–6649. doi:10.18653/v1/2021.emnlp-main.532.

[63] Scialom T, Dray PA, Lamprier S, Piwowarski B, Staiano J, Wang A, et al. QuestEval: Summarization Asks for Fact-based Evaluation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021; p. 6594–6604. doi:10.18653/v1/2021.emnlp-main.529.

[64] Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a Freely Accessible Critical Care Database. Scientific data. 2016;3(1):1–9. doi:10.1038/sdata.2016.35.

[65] Voorhees EM, Hersh WR. Overview of the TREC 2012 Medical Records Track. Proceedings of the twentieth Text REtrieval Conference. 2012;.

[66] Özlem Uzuner, Goldstein I, Luo Y, Kohane I. Identifying Patient Smoking Status from Medical Discharge Records. Journal of the American Medical Informatics Association. 2008;15(1):14–24. doi:10.1197/jamia.M2408.

[67] Aramaki E, Morita M, Kano Y, Ohkuma T. Overview of the NTCIR-12 MedNLPDoc Task. In Proceedings of NTCIR-12. 2016;.

[68] Aramaki E. GSK2012-D Dummy Electronic Health Record Text Data [Internet]. Gengo-Shigen-Kyokai; 2013 Feb [cited 2021 Aug 6]. Available from: https://www.gsk.or.jp/catalog/gsk2012-d.

[69] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out. 2004; p. 74–81.

[70] National Hospital Organization [Internet]. 診療情報集積基盤 (In Japanese); 2015 Aug 5- [cited 2021 Aug 6]. Available from: https://nho.hosp.go.jp/cnt1-1_000070.html.

[71] Lebanoff L, Song K, Liu F. Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018; p. 4131–4141. doi:10.18653/v1/D18-1446.

[72] Xu S, Li H, Yuan P, Wu Y, He X, Zhou B. Self-Attention Guided Copy Mechanism for Abstractive Summarization; 2020. p. 1355–1362.

[73] Li H, Xu S, Yuan P, Wang Y, Wu Y, He X, et al. Learn to Copy from the Copying History: Correlational Copy Network for Abstractive Summarization. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021; p. 4091–4101. doi:10.18653/v1/2021.emnlp-main.336.

[74] Vladutz G. Natural Language Text Segmentation Techniques Applied to the Automatic Compilation of Printed Subject Indexes and for Online Database Access. Proceedings of the First Conference on Applied Natural Language Processing. 1983; p. 136–142. doi:10.3115/974194.974221.

[75] Mann WC, Thompson SA. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text-interdisciplinary Journal for the Study of Discourse. 1988;8(3):243–281.

[76] Kishimoto Y, Murawaki Y, Kawahara D, Kurohashi S. Japanese Discourse Relation Analysis: Task Definition, Connective Detection, and Corpus Annotation. Journal of Natural Language Processing. 2020;27(4):889–931. doi:10.5715/jnlp.27.889.

[77] Kreuzthaler M, Schulz S. Detection of Sentence Boundaries and Abbreviations in Clinical Narratives. BMC Medical Informatics and Decision Making. 2015;15(2):1–13. doi:10.1186/1472-6947-15-S2-S4.

[78] Griffis D, Shivade C, Fosler-Lussier E, Lai AM. A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain. AMIA Joint Summits on Translational Science Proceedings. 2016; p. 88–97.

[79] Kudo T. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Version 0.996 [software]; 2006 Mar 26 [cited 2021 Aug 6]. Available from: https://taku910.github.io/mecab.

[80] Sato T, Hashimoto T, Okumura M. Implementation of a Word Segmentation Dictionary Called Mecab-ipadic-NEologd and Study on How to Use It Effectively for

Information Retrieval. Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing. 2017; p. NLP2017–B6–1.

[81] Ito K, Nagai H, Okahisa T, Wakamiya S, Iwao T, Aramaki E. J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage. Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018;.

[82] Maruyama T, Kashioka H, Kumano T, Tanaka H. Development and Evaluation of Japanese Clause Boundaries Annotation Program. Journal of Natural Language Processing. 2004;11(3):39–68. doi:10.5715/jnlp.11.3_39.

[83] Li J, Sun A, Joty SR. SegBot: A Generic Neural Text Segmentation Model with Pointer Network. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018; p. 4166–4172. doi:10.24963/ijcai.2018/579.

[84] Vinyals O, Fortunato M, Jaitly N. Pointer Networks. Advances in Neural Information Processing Systems 28. 2015; p. 2692–2700.

[85] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics. 2017;5:135–146. doi:10.1162/tacl_a_00051.

[86] Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T. Learning Word Vectors for 157 Languages. Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018;.

[87] Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic acids research. 2004;32(suppl_1):D267–D270.

[88] Wu Y, Lei J, Wei WQ, Tang B, Denny JC, Rosenbloom ST, et al. Analyzing Differences Between Chinese and English Clinical Text: A Cross-Institution Comparison of Discharge Summaries in Two Languages. Studies in health technology and informatics. 2013;192:662.

[89] Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic Recognition of Disorders, Findings, Pharmaceuticals and Body Structures from Clinical Text: An Annota-

tion and Machine Learning Study. Journal of Biomedical Informatics. 2014;49:148–158. doi:10.1016/j.jbi.2014.01.012.

[90] Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, et al. Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative. Journal of the American Medical Informatics Association. 2015;22(1):143–154.

[91] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. Journal of the American Medical Informatics Association. 2011;18(5):552–556.

[92] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019; p. 4171–4186. doi:10.18653/v1/N19-1423.

[93] Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. PLOS ONE. 2021;16(11):1–11. doi:10.1371/journal.pone.0259763.

[94] Kurohashi-Kawahara Laboratory. ku_bert_japanese [software]; 2019 [cited 2021 Aug 6]. Available from: https://nlp.ist.i.kyoto-u.ac.jp/index.php?ku_bert_japanese.

[95] Inui Laboratory. BERT models for Japanese text [software]; 2019 [cited 2021 Aug 6]. Available from: https://github.com/cl-tohoku/bert-japanese.

[96] National Institute of Information and Communications Technology. NICT BERT 日本語 Pre-trained モデル [software]; 2020 [cited 2021 Aug 6]. Available from: https://alaginrc.nict.go.jp/nict-bert/index.html.

[97] Caruana R. Multitask Learning. Machine learning. 1997;28(1):41–75. doi:10.1023/A:1007379606734.

[98] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations. 2015;.

[99] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. Advances in Neural Information Processing Systems 31. 2017; p. 6000–6010.

[100] Liu F, Liu Y. Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2008; p. 201–204.

[101] Cannon J, Lucci S. Transcription and EHRs: Benefits of a Blended Approach. Journal of American Health Information Management Association. 2010;81(2):36–40.

[102] Wu Y, Lei J, Wei WQ, Tang B, Denny JC, Rosenbloom ST, et al. Analyzing Differences between Chinese and English Clinical Text: A Cross-Institution Comparison of Discharge Summaries in Two Languages. Studies in Health Technology and Informatics. 2013;192:662–666.

[103] Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, et al. Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative. Journal of the American Medical Informatics Association. 2015;22(1):143–154. doi:10.1136/amiajnl-2013-002544.

[104] Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, et al. Multimodal Machine Learning for Automated ICD Coding. Proceedings of the 4th Machine Learning for Healthcare Conference. 2019;106:197–215.

[105] Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data Integration of Structured and Unstructured Sources for Assigning Clinical Codes to Patient Stays. Journal of the American Medical Informatics Association. 2015;23(e1):e11–e19. doi:10.1093/jamia/ocv115.

[106] WHO. International Classification of Diseases, Clinical Modification (Ninth Revision); [cited 2022 Aug 30]. Available from: https://www.cdc.gov/nchs/icd/icd9cm.html.

[107] WHO. International Classification of Diseases, Clinical Modification (Tenth Revision); [cited 2022 Aug 30]. Available from: https://www.cdc.gov/nchs/icd/icd10cm_pcs.htm.

[108] U S Food and Drug Administration. National Drug Code Database Background Information; [cited 2022 Aug 30]. Available from: https://www.fda.gov/drugs/development-approval-process-drugs/national-drug-code-database-background-information.

[109] WHO. Anatomical Therapeutic Chemical (ATC) Classification System; [cited 2022 Aug 30]. Available from: http://www.whocc.no/atc/structure_and_principles.

[110] RIZIV. Rijksintituut Voor Ziekte- En Invaliditeitsuitkeringen Nomenclature; [cited 2022 Aug 30]. Available from: http://www.riziv.fgov.be/NL/nomenclatuur/Paginas/default.aspx#.VOX1TzU2x0x.

[111] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Proceedings of the 1st Machine Learning for Healthcare Conference. 2016;56:301–318.

[112] Futoma J, Hariharan S, Heller K. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. Proceedings of the 34th International Conference on Machine Learning. 2017;70:1174–1182.

[113] Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. IEEE Access. 2018;6:65333–65346.

[114] Lample G, Conneau A. Cross-lingual Language Model Pretraining. Advances in Neural Information Processing Systems. 2019;.

[115] Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. ArXiv. 2020;abs/arxiv.2004.05150. doi:10.48550/arxiv.2004.05150.

[116] Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv. 2016;abs/1609.08144. doi:10.48550/arXiv.1609.08144.

[117] Zhou C, Neubig G, Gu J, Diab M, Guzmán F, Zettlemoyer L, et al. Detecting Hallucinated Content in Conditional Neural Sequence Generation. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021; p. 1393–1404. doi:10.18653/v1/2021.findings-acl.120.

[118] Goodrich B, Rao V, Liu PJ, Saleh M. Assessing The Factual Accuracy of Generated Text. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019; p. 166–175. doi:10.1145/3292500.3330955.

[119] Ando K, Okumura T, Komachi M, Horiguchi H, Matsumoto Y. Exploring Optimal Granularity for Extractive Summarization of Unstructured Health Records: Analysis of the Largest Multi-Institutional Archive of Health Records in Japan. PLOS Digital Health. 2022;1(9):1–19. doi:10.1371/journal.pdig.0000099.

[120] Devaraj A, Sheffield W, Wallace B, Li JJ. Evaluating Factuality in Text Simplification. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022; p. 7331–7345. doi:10.18653/v1/2022.acl-long.506.

[121] Mielke S, Szlam A, Dinan E, Boureau YL. Reducing Conversational Agents' Overconfidence through Linguistic Calibration. Transactions of the Association for Computational Linguistics. 2022;10(0).

[122] Reinke CE, Kelz RR, Baillie CA, Norris A, Schmidt S, Wingate N, et al. Timeliness and Quality of Surgical Discharge Summaries After the Implementation of an Electronic Format. The American Journal of Surgery. 2014;207(1):7–16. doi:https://doi.org/10.1016/j.amjsurg.2013.04.003.

[123] Wang R, Carrington JM, Hammarlund N, Sanchez O, Revere L. An Evaluation of Copy and Paste Events in Electronic Notes of Patients With Hospital Acquired Conditions. International Journal of Medical Informatics. 2023;170:104934. doi:https://doi.org/10.1016/j.ijmedinf.2022.104934.

[124] Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe Practices for Copy and Paste in the EHR. Applied clinical informatics. 2017;26(01):12–34.

[125] Manakul P, Gales M. Long-Span Summarization via Local Attention and Content Selection. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021; p. 6026–6041. doi:10.18653/v1/2021.acl-long.470.

[126] Mao Z, Wu CH, Ni A, Zhang Y, Zhang R, Yu T, et al. DYLE: Dynamic Latent Extraction for Abstractive Long-Input Summarization. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022; p. 1687–1698. doi:10.18653/v1/2022.acl-long.118.

[127] Jin D, Jin Z, Zhou JT, Szolovits P. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. The Thirty-Fourth AAAI Conference on Artificial Intelligence. 2020;34(05):8018–8025. doi:10.1609/aaai.v34i05.6311.

[128] Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted Diagnosis for Knee Magnetic Resonance Imaging: Development and Retrospective Validation of MRNet. PLOS Medicine. 2018;15(11):1–19. doi:10.1371/journal.pmed.1002699.

[129] Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior Skin Cancer Classification by the Combination of Human and Artificial Intelligence. European Journal of Cancer. 2019;120:114–121. doi:https://doi.org/10.1016/j.ejca.2019.07.019.

[130] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer Collaboration for Skin Cancer Recognition. Nature Medicine. 2020;26(8):1229–1234.

[131] Pugoy RA, Kao HY. Unsupervised Extractive Summarization-Based Representations for Accurate and Explainable Collaborative Filtering. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021; p. 2981–2990. doi:10.18653/v1/2021.acl-long.232.

[132] Wang H, Gao Y, Bai Y, Lapata M, Huang H. Exploring Explainable Selection to Control Abstractive Summarization. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35(15):13933–13941. doi:10.1609/aaai.v35i15.17641.

[133] Nguyen K, O'Connor B. Posterior Calibration and Exploratory Analysis for Natural Language Processing Models. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015; p. 1587–1598. doi:10.18653/v1/D15-1182.

[134] Desai S, Durrett G. Calibration of Pre-trained Transformers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020; p. 295–302. doi:10.18653/v1/2020.emnlp-main.21.

[135] Ao S, Acharya X. Learning ULMFiT and Self-Distillation with Calibration for Medical Dialogue System. Proceedings of the Fourth International Conference on Natural Language and Speech Processing. 2021; p. 196–203.

[136] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways. arXiv. 2022;doi:10.48550/ARXIV.2204.02311.

[137] Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A Large Language Model for Electronic Health Records. npj Digital Medicine. 2022;5(1):1–9.

[138] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. arXiv. 2022;doi:10.48550/ARXIV.2212.13138.

[139] Dou Q, So TY, Jiang M, Liu Q, Vardhanabhuti V, Kaissis G, et al. Federated Deep Learning for Detecting COVID-19 Lung Abnormalities in CT: A Privacy-Preserving Multinational Validation Study. NPJ digital medicine. 2021;4(1):1–11.

[140] Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients with COVID-19: Machine Learning Approach. JMIR medical informatics. 2021;9(1):e24207.