KYUTECH-LSSE-19899031

Doctoral Dissertation

The Development of Evaluation Metrics for Sentence Suggestion in Nursing and Elderly Care Record Application

Defry HAMDHANA

February 5, 2024

Department of Life Science and Systems Engineering
Graduate School of Life Science and Systems Engineering
Kyushu Institute of Technology

A Doctoral Dissertation
submitted to Graduate School of Life Science and Systems Engineering,
Kyushu Institute of Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Defry HAMDHANA

Thesis Committee:

Professor Sozo INOUE	(Supervisor)
Professor Keiichi HORIO	(Co-supervisor)
Professor Kaori YOSHIDA	(Co-supervisor)
Professor Kazukata SHIMADA	(Pre-defense Committee)

The Development of Evaluation Metrics for Sentence Suggestion in Nursing and Elderly Care Record Application*

Defry HAMDHANA

Abstract

This thesis proposes a novel metrics evaluation framework for assessing the quality of sentence suggestions generated by a model in nursing care record applications. The objective is to introduce a systematic approach for evaluating the quality of generated sentence suggestions, allowing for assessments comparable to caregiver evaluations. Our proposed framework aims to provide a comprehensive and standardized method for evaluating the efficacy of sentence suggestions. By establishing a systematic evaluation process, we seek to bridge the gap between automated assessments and human evaluation, contributing to the development of more reliable and accurate models in the field of nursing care record applications.

During the initial phase of our study, we used a Markov model to generate sentence suggestions within the context of nursing care record applications. These suggestions were then compared systematically against ground truth care records, serving as a reference for accuracy and relevance. Furthermore, we conducted a human evaluation to obtain caregivers' opinions and establish a ground truth for the assessment process. By comparing the model-generated suggestions with ground truth care records and expert evaluations, our study aims to assess the performance and applicability of the Markov model comprehensively.

Based on this foundation, our study evaluated the generated sentence suggestions using several existing metrics. The outcomes of these metrics were then

*Doctoral Dissertation, Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, KYUTECH-LSSE-19899031, February 5, 2024.

systematically compared against human evaluations, and the results were meticulously observed. Given the unique characteristics of care records, we found that the current evaluation metrics fell short of delivering satisfactory assessments. The intricacies of healthcare documentation necessitate a more nuanced approach to evaluation. Our findings underscore the need for customised metrics that can capture the specific intricacies and nuances of sentence suggestions within the context of care records.

In conclusion, our proposed evaluation metric outperforms several current evaluation methods in assessing sentence suggestion generation within care record applications. The meticulous comparison against existing metrics revealed the limitations of conventional approaches in adequately capturing the intricacies of healthcare documentation. By introducing a more tailored evaluation methodology, our study seeks to address these limitations and enhance the accuracy and relevance of assessments.

Keywords:

Sentence Suggestion, Nursing Care Record, Evaluation Metrics

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A nursing care record application collects electronic health records documenting elderly healthcare services and treatments. The information in elderly care includes diagnoses, examination results, care plans, prescribed medications, and performed medical procedures [51]. Nursing care records also include meticulous documentation of care provided, from caregivers to physicians, and encompass medical interventions. Nursing care records encapsulate data related to hospital visits or other nursing care facilities. Administrative details owned by the elderly are also integral components, making nursing care records repositories of critical health-related information.

In nursing care record applications, sentence suggestion emerges as a strategic solution to address caregivers' time constraints in reporting elderly conditions through nursing care record applications. To streamline the recording process, sentence suggestion involves completing sentences triggered by specific user-inputted words or phrases. The primary objective of sentence suggestions is to afford caregivers sufficient time to record patient conditions accurately. This method aims to facilitate the seamless and error-free documentation of patient information by automating the completion of sentences based on user input.

However, the use of sentence suggestion models in nursing care record applications is challenging because of the inherent structural complexity of care records. The challenges arise from diverse and non-standardised sentence structures, coupled with the use of specialised medical terminology. Care records typically encompass various linguistic patterns and medical-specific terminology, posing a significant hurdle for conventional sentence suggestion models. Addressing these challenges requires the development of more sophisticated models capable of accommodating the intricacies of diverse sentence structures

and medical terminology.

1.1 Problem Statement

The presence of evaluation metrics capable of precisely assessing sentence suggestions within care record applications, with indications closely aligning with human assessments, proves instrumental in advancing the development of specialized sentence suggestion models for care records. The inherent complexity of healthcare documentation underscores the critical need for accurate evaluations in this domain. A closer alignment between automated evaluations and human judgement is essential for refining and optimising sentence suggestion models in care record applications.

This imperative arises from the necessity for sentence suggestions to represent patients' conditions accurately within care record applications. The potential risks associated with inaccuracies in caregiver reports underscore the critical importance of ensuring the accuracy and reliability of sentence suggestions. Caregivers' information discrepancies can have serious consequences, emphasising the need for meticulous and error-free documentation. Therefore, the alignment between automated sentence suggestions and the actual patient conditions is crucial, as it directly affects the quality and safety of healthcare delivery.

The collective statement of the problem aims to enhance the accuracy of evaluation metrics in assessing the performance of sentence suggestion generation within care record applications.

- **Problem 1. The sentence suggestion does not accurately represent the original sentence.**
- **Problem 2. The current model exhibits a variance in evaluation compared to an expert, failing to align with expert opinions.**
- **Problem 3. A method can be used to assess the generation of sentence suggestions in care records.**

1.2 Research Questions

Based on the discussed problem statement, several issues need to be addressed to evaluate the quality of sentence suggestion generation in care record applications. The following research questions are formulated to guide the investigation:

1. **RQ1:** How effectively does the current sentence suggestion generation system capture the nuances of original sentences in care records?
2. **RQ2:** How does evaluating the sentence suggestion generation model differ from expert opinions in the context of care records?
3. **RQ3:** What methodologies can be employed to assess and enhance the overall quality of sentence suggestions generated in care record applications?

1.3 Key Contributions

This thesis aims to discover a more effective approach to evaluating sentence suggestion generation in care record applications. The following key contributions are delineated, each addressing a specific aspect of the identified problem statements:

1. This thesis proposes a novel evaluation metrics approach utilizing an HDP (Hierarchical Dirichlet Process) model embedded with word embeddings and customised pre-processing tailored to sentence structure.
2. This thesis conducts a comprehensive analysis and identification of the existing evaluation metrics employed in assessing sentence suggestions within care record applications.
3. This thesis presents a comparative analysis of the performance between the current evaluation metrics and the newly proposed metrics in evaluating sentence suggestion generation in care record applications.

1.4 Thesis Outline

This thesis comprises a total of 6 chapters. The following are subsequent sections related to each chapter:

- In Chapter 2, we introduce the background study by highlighting care records characteristics, the challenge of sentence suggestions and current evaluation metrics.
- In Chapter 3, we outline our data collection methods and present initial results from our experiments. This section provides an overview of the data-gathering process and offers insights into the preliminary outcomes, laying the groundwork for subsequent analysis and discussions.

- In Chapter 4, we describe our conducted experiments, showcasing the proposed methodology and its corresponding results.
- In Chapter 5, we analyze research findings, interpret their significance, and examine implications for specific sentence conditions in care records.
- In Chapter 6, we discuss the proposed EmbedHDP with a number of issues that were discovered during the research process.
- In Chapter 7, we summarize our contribution and possible future directions for research metrics evaluation for sentence suggestion in care record applications.

Chapter 2

Related Work

This chapter provides a comprehensive overview of the inception of our proposed evaluation metrics. It explores the contributions of previous research to the broader field of sentence generation, specifically within the domain of sentence suggestion. Emphasis is placed on a detailed discussion of the characteristics inherent in sentences found within nursing care records, forming a pivotal foundation for the current study.

2.1 Nursing Care Records

Nursing care records aim to record the elder's medical history, diagnosis, treatment, and actions by doctors or other health professionals [6]. These records serve as tangible evidence of assessments and interventions, promoting continuity of care by allowing other healthcare professionals to understand current care plans and treatments for patients easily. Efficiently maintaining records and documentation is a vital aspect of the responsibilities of healthcare professionals, encompassing nurses, and plays a crucial role in facilitating the delivery of safe and high-quality patient care. Regulatory standards governing the practice of nurses underscore the significance of upholding clear and accurate patient records. As the policies and procedures regarding the upkeep of patient records may differ among healthcare organizations, caregivers must verify and adhere to these guidelines [48]. The nursing process reflects the assessment, planning, implementation and evaluation principles shown in Figure 2.1.

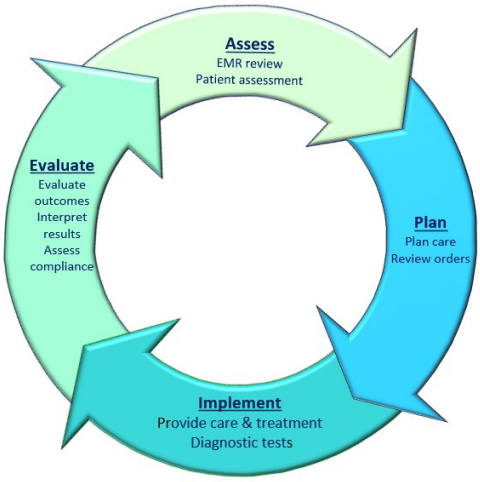

Fig. 2.1: Nursing Process

Commencing each shift involves conducting a primary assessment, following the Nursing Assessment Guideline. This assessment draws information from bedside handovers, patient introductions, essential documentation (covering safety checks, risk assessments, and clinical observations), and an electronic medical record (EMR) examination. The EMR review examines various components, including the patient storyboard[26], Informative Presence (IP) Summary[44], Victorian Children' s Tool for Observation and Response (ViCTOR) Graph, Notes, Results Review, Medication Administration Record (MAR)[50], Fluid Balance, Avatar, Orders, and Flowsheets. It is suggested to personalize tabs based on specific patient needs, emphasising standardizing layouts across wards.

Utilizing information obtained during the initial shift assessment, a collaborative care plan is formulated with the patient and family/carers to establish clear expectations for care. The Hub serves as a shift planning tool, offering a chronological representation of the care plan, encompassing ongoing assessments, diagnostic tests, appointments, scheduled medications, procedures, and tasks. Orders automatically populate the Hub, enabling nurses to document directly into Flowsheets in real-time. Effective order management is vital for the functionality of the Hub and must be addressed before handover to mitigate patient safety risks. Nurses can supplement the Hub with additional tasks for reminders, and all patient documentation, including observations and assessments, can be inputted into Flowsheets throughout the shift. Nurses can also use narrators or navigators for appropriate patient care documentation, and any clinical information not covered in these areas is recorded in real-time progress notes. This comprehensive documentation covers

abnormal assessments, changes in clinical state, adverse events, patient outcomes, family-centred care, and social issues.

Effective progress notes go beyond listing tasks or events, offering insightful information on occurrences, rationale, and implications for the patient and family. Accuracy and patient specificity are crucial; generic information like 'ongoing management' is discouraged. Duplication and vague statements, such as referencing information in the EMR, should be avoided. Professional nursing language, with minimal abbreviations following medicine and nursing standards, is preferred[46]. Real-time notes should be signed off after the initial entry, with subsequent entries added as addendums. An illustrative entry involves addressing a patient's increasing leg pain, urinary incontinence, and routine IV therapy blood, with comprehensive details and actions taken.

Elderly care can be applied in nursing because it addresses the unique healthcare needs and considerations associated with the ageing population, encompassing a holistic approach that involves physical, mental, and emotional well-being[3]. Integrating elderly care into nursing care applications ensures a tailored and comprehensive healthcare experience, considering the specific challenges and requirements of elderly individuals. This may involve features such as personalized health plans, medication management, mobility support, and monitoring of vital signs, all aimed at optimizing the quality of care provided to older adults. Using a time series approach, Caballero and Akella [7] developed a model to predict the elders' health conditions from nursing care applications. They underscore the importance of technology to increase understanding of elderly health status and enable more informed and effective decision-making in elderly care. In the development and role of nursing care records in the healthcare system, nursing care records refer to elderly medical records that are stored electronically and can be shared with authorized healthcare teams. The history of care records and technological developments have helped change how the healthcare system works [13]. Initially, nursing care records can be maintained in both paper and digital formats[19]. However, with the advent of digital technologies, electronic health records (EHRs) have become increasingly popular in healthcare settings. EHRs provide a safe and secure digital space to store patients' health information, including nursing care records[27]. Some benefits of using nursing care records are increasing the efficiency and quality of health care, improving patient safety, and facilitating research and development of drugs and treatments[53]. For a complete look at the sections that go into care records, see Figure 2.2. However, with technological developments, care records

are now more complex and capable of collecting, processing, and analyzing patient health data on a large scale. In addition, care records also have challenges related to their use, such as data security, interoperability, and proper use by health professionals.

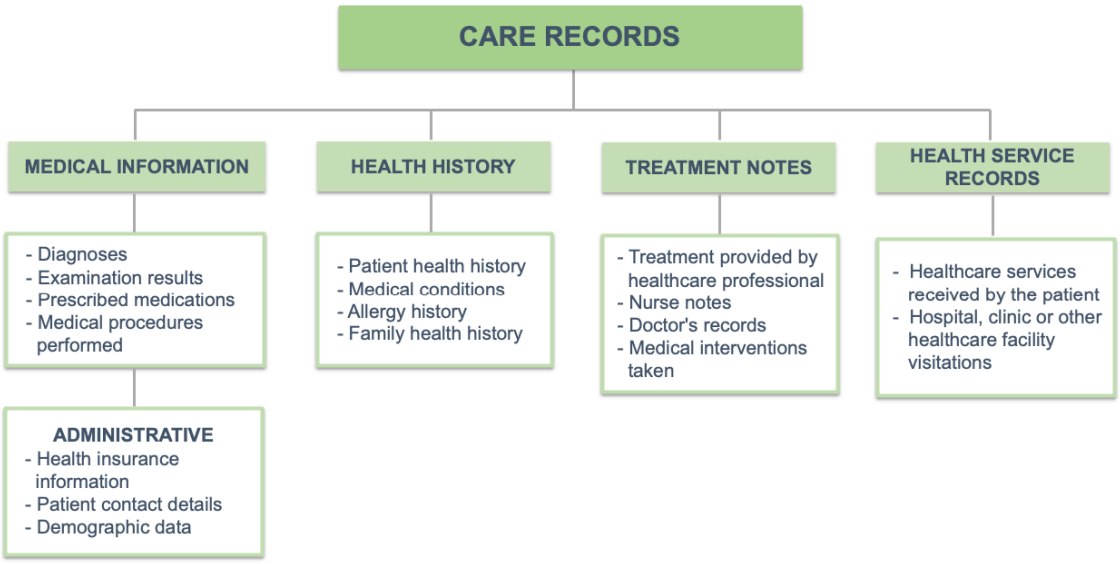

Fig. 2.2: The Objectives of Care Records

In this study, we used FonLog as a nursing care record application installed in more than 30 healthcare facilities in Japan. FonLog[32] is a mobile application designed as a data collection tool in human activity recognition for nursing services. Thus, caregivers easily identify and record patient activities using a mobile phone with key advantages such as recording the targeted patient, an easy-to-use interface, a recognition feedback interface, other customizable detail records, instant activity, and offline accessibility. As a default, FonLog has 88 activity types in Japanese. We focus on providing sentence suggestions on notices input in 31 activity types and containing free format record text, as shown in Table 2.1 and Table 2.2.

Table 2.1: Notices input in Activity type.

Activity type	**Record type**
1. バイタル (vitals), 2. リハビリ・レク (rehabilitationrecreation), 3. 往診・受診 (house callsvisit), 4. 処置 (treatment), 5. 入浴・清拭 (bathing/cleaning), 6. 外出対応 (going out), 7. 活力朝礼・ラジオ体操 (vitality morning/radio exercise), 8. 特記事項・連絡事項 (special notes/notifications), 9. 送迎 (transportation), 10. 事故等緊急対応 (emergency response), 11. 就寝前食事 (meal before bedtime), 12. モーニングケア (morning care), 13. ナイトケア (night care), 14. その他食事 (other meals), 15. 家族・来客対応 (family/visitor support), 16. 家族・医師連絡 (family/doctor contact), 17. 手書き記録 (handwritten records), 18. 入院 (hospitalization), 19. 離床・臥床介助 (assistance with getting out of bed and lying down), 20. 食事・服薬 (meals and medication), 21. おやつ (snacks), 22. 更衣介助 (assistance with changing clothes), 23. 口腔ケア (oral care), 24. 排泄 (excretion), 25. 日中利用者対応 (support for daytime user),	1. 特記事項 (spacial notes), 2. 状態・特記事項 (condition/special notes), 3. 連絡事項 (notifications), 4. 傷の状態・特記事項 (status/special notes)

Table 2.2: Notices input in Activity type (cont).

Activity type	**Record type**
26. 夜間利用者対応 (support for nighttime user), 27. 朝食 (breakfast), 28. 昼食 (lunch), 29. 夕食 (dinner), 30. 洗面介助 (washing assistance), 31 外泊 (overnight stay)	1. 特記事項 (spacial notes), 2. 状態・特記事項 (condition/special notes), 3. 連絡事項 (notifications), 4. 傷の状態・特記事項 (status/special notes)

One of the inherent challenges in dealing with nursing care records lies in their **diverse sentence structure**. These records exhibit a rich tapestry of non-standard sentence constructions, making them inherently more complex than the standardized language often encountered in general texts[47]. This diversity arises from the varied nature of elderly histories, medical observations, and treatment plans, which can manifest in different linguistic forms. Traditional language models, designed to focus on conventional grammatical structures, may encounter difficulties in accurately interpreting and generating content that mirrors the intricate sentence structures in care records.

Furthermore, the **specialized medical terminology** in nursing care records introduces additional complexity. These documents incorporate highly specialized medical terminology, ranging from specific drug names to detailed descriptions of medical conditions and treatment procedures. The intricate vocabulary employed in healthcare documentation is crucial for precision and clarity. Still, it poses a considerable challenge for language models and evaluation metrics that may not be well-versed in the nuances of medical discourse. Consequently, assessing the similarity and relevance of sentences containing these specialized terms becomes a formidable task.

Our work, which generates evaluation metrics for sentence suggestion in nursing care records, is considered crucial and essential. Nursing care records play a central role in documenting the medical history, diagnosis, treatment, and interventions of older adults. These records serve as tangible evidence of assessments and interventions, ensure continuity of care, and facilitate a comprehensive understanding of current care plans among healthcare professionals. The efficient maintenance of these records is a critical responsibility for healthcare professionals, particularly nurses, as it contributes significantly to the delivery of safe, quality patient care. Compliance with regulatory standards for nursing

practice underscores the importance of maintaining clear and accurate patient records. Given the differences in policies and procedures between healthcare organizations, caregivers must carefully view and adhere to these guidelines.

2.2 Sentence Suggestion

Several research studies related to sentence suggestion use the keyword sentence completion. Based on existing research, rule-based, n-gram, or language models were applied. Asnani et al. [1] explain that sentence completion utilizes techniques such as n-gram language models, neural network-based language models, and Markov Chain methods. N-gram and Markov language models are easy to understand and apply for short and simple texts. They also discussed the advantages of neural network-based language models, which can model words over long distances but require a lot of data to train and are expensive and time-consuming. In another study, Mirowski and Vlachos [33] researched to improve the performance of Recurrent Neural Network (RNN) language models by incorporating the syntactic dependencies of a sentence to have the effect of bringing in a context relevant to the word being predicted. In general, it can be concluded that this model is designed to learn word and grammar representations from text data and used to complete sentences automatically. The dependency Recurrent Neural Language Model (DRNLM) integrates word representation learning, grammar learning (dependency learning), and word order learning (recurrent learning) to produce accurate sentence representations. They evaluate DRNLM on three different datasets, namely TREC dataset, MCScript dataset, and CommonsenseQA dataset. As a result of the evaluation, DRNLM outperforms state-of-the-art methods on all datasets. In their research, Irie et al. [20] investigated the use of Recurrent Neural Networks (RNN) and bi-directional LSTM-RNN (Long Short-Term Memory) variations in estimating sentence probabilities. The research included two experiments: first, examining the effectiveness of using forward and backward RNNs in estimating sentence probabilities; second, testing the combined methods of forward and backward RNNs, as well as bi-directional LSTM-RNNs in estimating sentence probabilities. The results showed that using forward and backward RNNs separately resulted in relatively low accuracy in estimating sentence probabilities. However, when both methods are combined, the results are significantly better. In addition, the results of bi-directional LSTM-RNN are better than those of forward and backward RNN separately. However, bi-directional LSTM-RNN is more complex regarding neural network structure and com-

putation time. Therefore, this study concludes that combining forward, backwards, and bi-directional LSTM-RNNs is the most effective method for estimating sentence probabilities. Another study explores the crucial task of predicting the next word in a sequence, which is central to natural language processing applications such as speech recognition and machine translation. Introducing Recurrent Neural Networks (RNNs) as a pivotal advancement, the study highlights their capacity to process sequential data and retain information over time. Addressing the challenges of vanishing or exploding gradients, the study introduces Long Short-Term Memory (LSTM) networks, a specialized variant of RNNs capable of learning long-term dependencies. Furthermore, the study further delves into the Neural Language Model (NLM), a model combining RNNs and LSTMs, showcasing its superiority over traditional n-gram models in capturing semantic and syntactic information [14]. Comparative analyses underscore the practical implications of these advancements, signalling a shift towards more accurate and context-aware natural language processing in applications like speech recognition and machine translation.

In conclusion, featuring RNNs, LSTMs, and NLM unveils a transformative journey in natural language processing. The Neural Language Model, with its amalgamation of RNNs and LSTMs, emerges as a powerful paradigm, surpassing traditional n-gram models and promising enhanced accuracy and efficiency in applications such as speech recognition and machine translation. As the field continues to progress, these sophisticated models hold the potential to redefine language understanding and generation, shaping the future of natural language processing.

Rakib et al. [38] developed a Bangla word prediction model using GRU (Gated Recurrent Unit) based recurrent neural network (RNN) and Ngram language model. This research aims to improve word prediction accuracy and sentence completion in Bangla. The results show that the GRU model produces better accuracy in word prediction and sentence completion than the conventional RNN model. This research shows that combining the n-gram language model and the GRU model can significantly improve word prediction accuracy and sentence completion.

In the pursuit of determining the efficacy of the sentence suggestions produced by the model and their reflective application in care records sentences, using a robust evaluation metric becomes imperative. We have identified noteworthy assessment variations through a comprehensive analysis of existing metrics and a subsequent comparative examination against human evaluations. This observation underscores the need for an evaluation metric

that gauges the quality of sentence suggestions and aligns closely with expert opinions. The intricacies of care records demand a nuanced evaluation approach that goes beyond conventional metrics and encapsulates the domain-specific expertise inherent in the field. Human evaluation, while invaluable, may introduce subjectivity and variability. Thus, developing a specialized evaluation metric tailored to the unique intricacies of care records is crucial.

Language models, such as Markov models, have been widely acknowledged for their simplicity and effectiveness in handling short and straightforward texts[1]. This aligns seamlessly with the characteristic structure of nursing care records, which often comprise concise and to-the-point sentences. Thus, we hypothesize that the application of Markov language models can be particularly suitable for enhancing sentence suggestion generation within the context of nursing care records. Markov models, specifically Markov chain models, operate on the principle of sequential dependencies within a given sequence of elements, be it words or states. These models are built upon the assumption that the probability of the next element in the sequence depends solely on the current state, disregarding the conditions that preceded it. This inherent simplicity allows Markov models to be easily understood and implemented, making them attractive for various natural language processing tasks, including sentence suggestions.

In recent studies, Markov models have found application in predicting the next word in tweets based on the user's personality, derived from their previous tweets using the Big Five personality model and deep learning techniques[12]. This example underscores the versatility of Markov models in capturing sequential patterns, particularly in short-text formats such as those found on Twitter. This study explores the implementation of Markov chain models in this context, shedding light on their potential applications beyond personality-based tweet prediction. The success of Markov models in this study suggests their adaptability to diverse datasets characterized by short and sequential content.

Accordingly, based on the preceding information, empirical evidence strongly supports our hypothesis that the Markov model is an apt choice for implementation in sentence suggestions within nursing care record applications. The Markov chain model operates on the principle of sequential dependencies within a given sequence of elements, which, in this case, are words or phrases within nursing care records. The primary objective is to generate contextually relevant and coherent sentence suggestions based on the existing content in the record. Here is the following algorithm of a basic procedure for generating

sentence suggestions in nursing care records in Algorithm 1.

Algorithm 1: Algorithm of Markov Chain Model Algorithm for Sentence Suggestion in Nursing Care Records

```
Require:care_records
Ensure: initial_words, second_possible_words, transitions
Result: sentence suggestion
    foreach care_record in care_records do
        length ← len(care_record.split());
        foreach word in care_record.split() do
            if word.isInitial_words() = True then
                second_possible_words[word] ←
              second_possible_words.get(word);
            else
                prev_word ← word.previous();
                if word.isLast() = True then
                    Expand(transitions, (prev_word, word), "END");
                else
                    Expand(transitions, (prev_word, word), word);
    while expanding(dict, key, value) do
        if key not in dict then
            dict[key] = [];
        dict[key].add(value);
```

The concept of the Markov algorithm in generating sentence suggestions revolves around the definition of three key variables: initial_words, representing the initial set of words that serve as input from caregivers; second_possible_words, encompassing words that act as transitions from caregiver inputs with their inherent transition characteristics; and transitions, encapsulating all potential transition words between states of individual words. The algorithm commences by parsing each care record data, treating them as ground truth sentences. Through a meticulous word analysis, we store all unique first words in the initial_words variable along with their possible continuations. Subsequently, if a word is positioned at the end of a sentence, the program retains that word and suggests the

subsequent word as "END." Alternatively, if the word is not at the sentence's conclusion, we preserve all words that exhibit relationships with it and their potential transitions.

Markov models, specifically Markov chain models, operate on the principle of sequential dependencies within a given sequence of elements, be it words or states. These models are built upon the assumption that the probability of the next element in the sequence depends solely on the current state, disregarding the states that preceded it. This inherent simplicity allows Markov models to be easily understood and implemented, making them attractive for various natural language processing tasks. Rationale for Markov Models in Nursing Care Records:

- Short and Concise Sentences: Nursing care records typically contain succinct sentences conveying essential information. Markov models, designed to handle sequential dependencies in data, can effectively capture the structure of these short sentences, ensuring coherent and contextually relevant sentence suggestions.
- Sequential Dependency in Patient Data: Patient data in nursing care records often follow a chronological order, with each entry building upon the previous one. Markov models excel in capturing such sequential dependencies, providing a coherent framework for generating contextually appropriate sentence suggestions.
- Reduced Complexity: The simplicity of Markov models makes them well-suited for the sometimes hectic and time-sensitive nature of healthcare settings. Their straightforward nature allows for quick implementation without compromising the quality of generated suggestions.

Our research is important because we are working on generating evaluation metrics for sentence suggestions in nursing care records. Previous studies have shown that there are different methodologies for sentence suggestions, such as rule-based, n-gram, and language models. The use of advanced techniques, such as RNNs, LSTM networks, and NLMs has significantly advanced the processing fo sequential data. These techniques capture semantic and syntactic information, improving the accuracy of sentence suggestions.

As the field advances, it is important to understand the nuances and complexities of nursing care records. The use of specialized medical terminology and diverse sentence structures can be challenging for existing language models and evaluation metrics. As the field advances, it's important to understand the nuances and complexities of nursing care records. The use of specialized medical terminology and diverse sentence structures

can be challenging for existing language models and evaluation metrics. We understand the importance of a comprehensive evaluation approach that goes beyond current metrics. This is one of the reasons why our team is committed to developing an evaluation metric that takes into account the intricacies of nursing care records.

2.3 Current Metrics Evaluation

Natural Language Generation (NLG) stands as a crucial facet within the expansive field of Natural Language Processing (NLP), dedicated to the development of sophisticated software systems adept at producing coherent and intelligible text[40]. This dynamic subset of NLP is characterized by its versatility, encompassing diverse tasks to convert disparate inputs into human-readable text sequences. The inputs can take various forms, from structured datasets and tables to natural language prompts and visual elements like images. NLG's adaptability positions it as a fundamental component applicable to a myriad of NLP tasks, showcasing its utility in generating responses for chatbots, facilitating language translation between different languages, providing creative writing suggestions, and condensing intricate data analyses into succinct and comprehensible summaries.

Within NLG's expansive scope, its significance is underscored by its ability to bridge the gap between structured data and human-understandable language. It plays a pivotal role in enhancing human-computer interactions by enabling machines to communicate in a manner that resonates with human users. As technology advances, NLG continues to evolve, pushing the boundaries of what is achievable in natural language understanding and generation. Integrating NLG into various applications streamlines information dissemination and augments the user experience across diverse domains, from customer service interactions to content creation and data interpretation. Thus, NLG's multifaceted capabilities are a cornerstone in leveraging artificial intelligence to enhance language-related tasks and communication.

Evaluating NLG model outputs poses a substantial challenge, particularly due to the open-ended nature of many NLG tasks[9]. For instance, a dialogue system may generate multiple viable responses to the same user input, and a document can be summarized in various ways. As a result, human evaluation stands as the gold standard for assessing the quality of NLG outcomes. However, the expense associated with human evaluation prompts researchers to frequently turn to automatic metrics for routine progress quantification and system optimization.

General evaluation metrics in sentence generation refer to standardized criteria used to assess the quality and effectiveness of generated sentences. These metrics play a pivotal role in objectively measuring the performance of natural language generation systems. By employing predetermined benchmarks, evaluators can gauge the generated content's accuracy, coherence, and relevance[9]. The overarching goal of employing such metrics is to ensure that the generated sentences align with linguistic norms, providing a reliable means to assess the proficiency of automated sentence generation processes.

The primary function of general evaluation metrics in sentence generation is to provide a quantitative and unbiased assessment of the generated output. These metrics serve as a yardstick for researchers, developers, and practitioners to compare different sentence generation models, identify areas for improvement, and enhance the overall quality of natural language generation systems[4]. Additionally, they contribute to fostering transparency and accountability in the development of language models, facilitating a standardized approach to evaluating the diverse applications of sentence generation. Beyond mere performance measurement, these metrics also support the iterative refinement of algorithms, fostering advancements in natural language processing.

2.3.1 BERTScore

Evaluating text generation against gold standard references, a common practice in tasks like machine translation and image captioning, has been significantly enhanced by introducing BERTScore. This metric assesses the similarity between two sentences using contextual embeddings derived from pre-trained BERT models[55]. These embeddings capture semantic and syntactic information, resulting in a more nuanced evaluation. Notably, BERTScore has demonstrated a superior correlation with human judgments, outperforming existing metrics and proving more effective in model selection.

The provided algorithm outlines the mechanics of BERTScore, a metric designed for evaluating the similarity between a reference sentence (denoted as R) and a candidate sentence (denoted as C) in the context of natural language processing tasks. The algorithm employs a sequence of operations leveraging pre-trained BERT (Bidirectional Encoder Representations from Transformers) models to tokenize and embed the input sentences.

1. Tokenization: The reference and candidate sentences, R and C, undergo tokenization using a pre-trained BERT tokenizer. This process breaks down the sentences into constituent tokens.

Algorithm 2: BERTScore Algorithm

Input: Reference sentence R, Candidate sentence C

Output: BERTScore B

$R_{\text{tokens}} \leftarrow \text{BERT\_Tokenizer}(R)$

$C_{\text{tokens}} \leftarrow \text{BERT\_Tokenizer}(C)$

$R_{\text{embeddings}} \leftarrow \text{BERT\_Embeddings}(R_{\text{tokens}})$

$C_{\text{embeddings}} \leftarrow \text{BERT\_Embeddings}(C_{\text{tokens}})$

$S \leftarrow \text{Cosine\_Similarity}(R_{\text{embeddings}}, C_{\text{embeddings}})$

$P \leftarrow \text{Precision}(S)$

$R \leftarrow \text{Recall}(S)$

$F1 \leftarrow \text{F1\_Score}(P, R)$

$B \leftarrow 2 \times \frac{P \times R}{P + R}$

return B

2. Embedding: The BERT embeddings for the tokenized sequences, R_{tokens} and C_{tokens}, are obtained. These embeddings capture contextual information, including semantic and syntactic nuances.
3. Cosine Similarity Calculation: A cosine similarity matrix, denoted as S, is computed based on the embeddings of the reference and candidate sentences. The matrix reflects the cosine similarity values between corresponding tokens.
4. Precision, Recall, and F1 Score Calculation: Precision, recall, and F1 score are computed based on the cosine similarity matrix. Precision (P) measures the ratio of correctly matched candidate tokens, recall (R) measures the ratio of correctly matched reference tokens, and the F1 score is the harmonic mean of precision and recall.
5. BERTScore Calculation: BERTScore is calculated using the harmonic mean of precision and recall, represented as $B = 2 \times \frac{P \times R}{P + R}$. This metric provides a comprehensive measure of the similarity between the reference and candidate sentences, and it has demonstrated an improved correlation with human judgments compared to existing metrics.

This algorithmic representation encapsulates how BERTScore operates, utilizing BERT embeddings and cosine similarity to assess the quality of generated text against reference sentences in a manner that aligns with human evaluation judgments.

A key strength of BERTScore lies in its adept handling of meaning-preserving lexical and compositional diversity, including paraphrases and word order changes. Unlike traditional methods reliant on exact or heuristic matching of n-grams, BERTScore allows for soft matching of tokens based on their embeddings. This feature enables the metric to capture long-range dependencies and ordering information, enhancing its capacity to evaluate the intricacies of text generation.

Furthermore, BERTScore is characterized by its simplicity, task-agnostic nature, and user-friendly implementation. Unlike some evaluation methods, BERTScore operates without needing external resources such as stemmers, synonym lexicons, or paraphrase tables. It is not specifically trained or optimized for any particular evaluation task, rendering it a versatile tool applicable to various languages and domains. Its computational efficiency further adds to its appeal, making BERTScore a convenient choice for integration into existing evaluation pipelines and contributing to the advancement and standardization of text-generation assessment practices.

2.3.2 Cosine Similarity

Cosine similarity stands as a crucial metric employed to quantify the similarity between two vectors within a multi-dimensional space[54][34]. Its widespread application is notably prominent in fields such as natural language processing, information retrieval, and machine learning, where assessing the likeness between documents or text-based data is essential. In the realm of natural language processing, cosine similarity proves invaluable for tasks such as document clustering, sentiment analysis, and text summarization, where the understanding of semantic relationships is paramount. This metric has also found substantial utility in information retrieval systems, playing a pivotal role in ranking documents based on their relevance to user queries. Furthermore, in the domain of machine learning, cosine similarity is frequently harnessed for document classification, collaborative filtering, and similarity-based recommendation systems, showcasing its versatility in various applications that demand a nuanced evaluation of vector-based data. Overall, cosine similarity emerges as a fundamental tool, providing a robust and efficient means to gauge the degree of similarity between vectors, thereby enhancing the capabilities of systems and algorithms across diverse domains. The cosine similarity is calculated based on the cosine of the angle between two vectors, represented as A and B. The formula for cosine similarity is as follows:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}{\mathbf{A} \cdot \mathbf{B}} \tag{2.1}$$

Here:

- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ represent the magnitudes (or lengths) of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively.
- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of vectors $\mathbf{A}$ and $\mathbf{B}$.

The resulting value ranges from -1 to 1, where a cosine similarity of 1 indicates that the vectors are identical, 0 means that the vectors are orthogonal (no similarity), and -1 indicates complete dissimilarity.

Cosine similarity is advantageous because it is independent of the vector's magnitude and only depends on the direction[5]. This property makes it useful for comparing documents or text regardless of their length. In text analysis, each document is represented as a vector in a high-dimensional space, where each dimension corresponds to a term in the document. By comparing the cosine similarity between these vectors, one can assess the degree of similarity or dissimilarity between the corresponding documents[18].

In information retrieval, cosine similarity ranks documents based on their relevance to a given query[17]. Machine learning is useful for tasks like clustering, where documents with similar content are grouped based on their vector representations. Overall, cosine similarity is a versatile and widely used metric for assessing similarity between vectors in various applications.

2.3.3 ROUGE

ROUGE, an acronym for Recall-Oriented Understudy for Gisting Evaluation, plays a pivotal role in natural language processing and information retrieval. It constitutes a set of metrics specifically designed to evaluate machine-generated text automatically. This set of metrics has garnered widespread adoption, particularly in assessing automatic summarization and machine translation outputs. Its fundamental purpose is to evaluate the quality and effectiveness of machine-generated summaries or translations by systematically comparing them to reference summaries or translations crafted by human experts. By applying ROUGE metrics, we can quantitatively measure the degree of overlap and similarity between the automatically generated text and the ideal human-crafted reference, providing valuable insights into the performance of natural language processing systems and machine translation algorithms.

ROUGE's significance is especially evident in scenarios where the precision of machine-generated content in capturing key information from the reference text is of utmost importance. By focusing on recall-oriented metrics, ROUGE emphasizes the ability of the generated text to recall and replicate essential information present in the human-created reference. This approach aligns with the overarching goal of ensuring that machine-generated content is not only relevant but also comprehensive and faithful to the information encapsulated in the reference material. As a versatile and widely utilized evaluation tool, ROUGE continues to be instrumental in advancing the field of natural language processing and bolstering the development of effective machine-generated content across various applications.

The primary focus of ROUGE metrics is on the content overlap and overlap of n-grams (contiguous sequences of n items, typically words) between the generated text and the reference text[30][31]. The underlying assumption is that a good summary or translation should contain essential information present in the reference text.

Some key components and metrics within the ROUGE framework include:

1. ROUGE-N (N-gram Overlap): This metric evaluates the overlap of n-grams between the system-generated output and the reference text. ROUGE-1 measures unigrams (single words), ROUGE-2 measures bigrams, and so on.
2. ROUGE-L (Longest Common Subsequence): Instead of looking at exact n-gram matches, ROUGE-L measures the longest common subsequence (LCS) between the system-generated output and the reference. This is particularly useful in capturing the information flow and sentence structure.
3. ROUGE-W (Weighted N-gram Overlap): This metric assigns different weights to different n-grams based on their lengths. It aims to give more importance to longer common sequences.
4. ROUGE-S (Skip-bigram Overlap): ROUGE-S measures the overlap of skip-bigrams, which are pairs of words that have a maximum gap size in between. This is useful for capturing semantic similarity.

ROUGE scores are typically reported as precision, recall, and F1 scores, providing a comprehensive evaluation of the system-generated text compared to the reference text. Higher ROUGE scores indicate better agreement between the machine-generated output and human-created reference summaries or translations.

2.3.4 BLEU

BLEU, which stands for Bilingual Evaluation Understudy, is an innovative method designed for the automatic evaluation of machine translation, offering a solution that is not only expeditious and cost-effective but also language-independent, and demonstrates a high correlation with human evaluation processes. BLEU utilizes a refined n-gram precision measure, which involves counting matching words and phrases between a candidate translation and one or more reference translations[35]. To prevent the accumulation of plausible words, these counts are then clipped by the maximum reference counts. Additionally, BLEU incorporates a brevity penalty factor to discourage excessively short candidate translations in comparison to reference translations. Notably, this penalty is computed across the entire corpus rather than on a sentence-by-sentence basis, providing a comprehensive assessment.

The culmination of BLEU's evaluation is a unified score ranging from 0 to 1, where a score of 1 signifies a flawless match with a reference translation[8]. This consolidated score encapsulates both the modified n-gram precisions and the brevity penalty, providing a succinct yet comprehensive measure of the translation quality.

The operation of BLEU involves several critical steps used to evaluate the quality of machine translation against human reference texts. The following are the key steps in the mechanism of BLEU:

1. Tokenization:
    - Candidate translation (machine translation) and reference translation (human translation) are segmented into smaller units, such as words or n-grams (sequences of n consecutive words).
2. Modified N-gram Precision:
    - Counting the number of matching words and phrases between the candidate translation and one or more reference translations for each n-gram.
    - Employing a "clipping" step to limit the count of matching words to not exceed the maximum reference count. This prevents the overgeneration of reasonable words.
3. Brevity Penalty:
    - Calculating the length of the candidate translation and the lengths of reference translation(s).

- Computing a brevity penalty factor to discourage excessively short candidate translations compared to the reference translations.
- The brevity penalty is calculated over the entire corpus rather than per sentence for a more comprehensive evaluation.

4. Geometric Mean:
   - Calculating the geometric mean of the previously computed precision for each n-gram.
   - Applying the brevity penalty to the geometric mean to generate the final value.
5. BLEU Score:
   - The BLEU score is the final value that encompasses both aspects, providing a score that reflects how closely the machine translation aligns with the human reference translation.
   - The BLEU score ranges from 0 to 1, where 1 indicates a perfect match with the reference translation.

This mechanism provides an overview of the extent to which machine translation achieves similarity with human reference texts. BLEU incorporates modified n-gram precision, and brevity penalty, and produces a score that offers a quantitative understanding of the machine translation quality in the context of automatic evaluation.

Numerous evaluation metrics are commonly applied for various general tasks, including assessing semantic or syntactic similarity, conducting evaluations in summary tasks, and appraising machine translation quality. Each metric offers unique evaluation mechanisms for assessing sentence suggestions in the context of medical records, accompanied by inherent limitations. In Table 2.3 below, each current evaluation metric is described along with its limitations.

Table 2.3: Comparison of Evaluation Metrics for Sentence Suggestion in Care Records

Evaluation Metrics	**Mechanism**	**Limitation**
BERTScore[55]	Comparing contextual embeddings of reference and candidate sentences using pre-trained BERT models	It relies on pre-trained BERT models, which may not capture domain-specific nuances effectively
Cosine Similarity[37]	Calculates the cosine angle between two vectors to determine their similarity, frequently used to compare text documents in vector space.	Fails to account for word order, and mistakenly rate semantical difference with similar word sets.
ROUGE[41]	Measures the overlap of n-grams and the longest matching sequence between a generated summary and reference texts.	Might overlook semantic accuracy as it is based on lexical overlap, not considering the context or meaning of the words.
BLEU[39]	Scores machine translations by matching n-grams to reference texts and adjusting for translation length.	Can miss the adequacy and fluency of translation as it primarily relies on n-gram overlap, ignoring semantic coherence.

From Table 2.3, valuable insights can be gleaned for adaptation and improvement by delving into the intricacies of the evaluation mechanisms of each method and understanding their limitations. Reviewing these metrics provides a robust foundation for identifying crucial points that can be leveraged to refine and enhance the evaluation process. For instance, the strength of word embeddings lies in their ability to recognize proximity between words based on their vector representations. On the other hand, the weakness of n-grams overlap is its tendency to ignore semantic coherence, failing to consider the context and meaning of words within a sentence.

The applicability of EmbedHDP to domains besides nursing care records needs careful consideration, primarily in evaluating its adaptability to diverse data characteristics.

EmbedHDP has effectively evaluated sentences with short and seemingly incomplete structures, where sentences may lack subjects or objects. That means domains with similar characteristics can adapt. The model's proficiency in handling incomplete sentences aligns with scenarios where linguistic structures may vary.

Moreover, the relevance of the task should be reviewed, as EmbedHDP was initially designed to assess sentence suggestions within nursing care records. Specific domain-related words in nursing care records are acceptable for the model's performance, particularly when evaluating the similarity between sentence suggestions and ground truth within the scope of nursing care records. Incorporating word embeddings and dictionaries proves instrumental in capturing the nuances of domain-specific language. One key consideration lies in the involvement of domain experts who can contribute invaluable insights to enhance the model's relevance and effectiveness. Consequently, before extending the application of EmbedHDP to different domains, a thorough understanding of data characteristics, task relevance, and expert involvement is indispensable to its successful adaptation.

Our work involves generating evaluation metrics for sentence suggestions in nursing care records. Evaluating natural language generation (NLG) model outputs in the healthcare domain, particularly in generating sentence suggestions for care records, is important. It presents unique challenges due to the open-ended nature of many NLG tasks, which can make assessing the quality of generated content complex. While human evaluation is considered the gold standard, it can be resource-intensive. That's why the development of automated metrics is crucial for routine progress quantification and system optimization.

In the field of nursing care records, evaluating text can be challenging due to diverse sentence structures and specialized medical terminology. However, we are excited to share customized evaluation metrics tailored to the intricacies of nursing care records. This metric will help address the challenges of evaluating this type of text and ensure that we are providing the best possible care for our patients. We understand that current metrics, such as BERTScore, cosine similarity, ROUGE, and BLEU, have limitations.

Our work aims to enhance the evaluation process by exploring each evaluation metric in detail, as outlined in Table 2.3. For example, EmbedHDP demonstrates the effectiveness of applying word embeddings and dictionaries in capturing the nuances of domain-specific language. We believe that a thorough understanding, task relevance, and the involvement of domain experts are crucial for ensuring the relevance and effectiveness of the model. Our work is important because we evaluate sentence suggestions within nursing care records.

We take a domain-specific approach to ensure the quality and relevance of generated information in healthcare documentation, which is crucial for improving elderly care.

Our undertaking, focused on the generation of evaluation metrics for sentence suggestions in nursing care records, is characterized by its absence in the current research landscape, its inherent importance, and the associated challenges that underscore its significance. The prevailing gap in existing literature regarding specialized evaluation metrics for the unique intricacies of nursing care records signifies the absence of a dedicated framework to systematically assess the quality and relevance of generated sentence suggestions in healthcare documentation. The importance of our work is underscored by the critical role played by nursing care records in documenting and communicating essential healthcare information. These records, laden with diverse sentence structures and specialized medical terminology, require a nuanced evaluation approach that extends beyond conventional metrics.

Moreover, our work is inherently challenging due to the diverse and complex nature of sentence structures within nursing care records. These records exhibit a rich tapestry of non-standard sentence constructions, further complicated by the presence of specialized medical terminology. Conventional language models designed for standardized language often struggle with accurately interpreting and generating sentence that mirror the intricate sentence structures found in nursing care records. The challenge is compounded by the need for evaluation metrics to navigate the varied linguistic forms that arise from the unique healthcare histories, medical observations, and treatment plans associated with the ageing population.

In summary, our work on the generating evaluation metrics for sentence suggestions in nursing care records is not only missing but also important and challenging. It aims to bridge the existing gap, recognizing the crucial role of tailored evaluation metrics in ensuring the accuracy and relevance of generated content within the complicated landscape of healthcare documentation.

Chapter 3

EmbedHDP Method to Improved Evaluation Metrics

This chapter will delve into the proposed evaluation model, EmbedHDP. The functional mechanisms of EmbedHDP have demonstrated their effectiveness in addressing challenges inherent in elderly care records.

3.1 Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a robust topic modelling technique to extract themes or topics from sentences. In general, topic modelling is a method used to extract the primary topics or themes from a large corpus of documents or text [24]. The essence of topic modelling is to identify hidden patterns in the text and discover interconnected topics based on words that frequently co-occur in documents. The HDP procedure represents an enhancement of Latent Dirichlet Allocation (LDA), a method derived from the certainty theorem [25] that aims to extract statistical structures of documents from various topics based on vocabulary distribution. HDP introduces a hierarchical structure that enhances its ability to capture latent topics within a corpus. Unlike LDA, HDP demonstrates superiority in automatically determining the number of topics, eliminating users' need to specify this parameter in advance. This adaptive capability makes HDP highly suitable for scenarios where the underlying topic structure is unknown. Here, we present the equation of HDP model used to calculate the similarity between a sentence generation and its ground truth:

Algorithm 3: Hierarchical Dirichlet Process Algorithm

Input: sentence1 ← sentence similarity, sentence2 ← ground truth

Output: similarity score between sentence1 and sentence2

S_1 be the set of unique tokens in sentence1,

S_2 be the set of unique tokens in sentence2,

D be the dictionary formed by combining S_1 and S_2,

C_1 be the Bag of Words (BoW) vector representing sentence1 in the corpus,

C_2 be the BoW vector representing sentence2 in the corpus,

HDP$(D, [C_1, C_2])$ be the trained Hierarchical Dirichlet Process model with dictionary D and corpus $[C_1, C_2]$,

T_1 be the topic distribution vector for sentence1 obtained from the trained HDP model,

T_2 be the topic distribution vector for sentence2 obtained from the trained HDP model.

The similarity between sentence1 and sentence2 can be calculated using a similarity metric, for example:

$$\text{Similarity}(T_1, T_2) = \text{Cosine Similarity}(T_1, T_2)$$

3.2 Word Embedding

Word embedding, a pivotal component of natural language processing, has garnered considerable attention for its capacity to represent words in a continuous vector space. This computational technique, exemplified by models such as Word2Vec[15], GloVe[36], and FastText[22], transforms words into dense numerical vectors, capturing intricate semantic relationships and nuanced contextual information. Word embedding involves harnessing neural networks to learn from vast corpora, enabling the models to discern subtle linguistic patterns and relationships. By considering the co-occurrence of words in sentences, these models create embeddings that encapsulate both semantic similarities and syntactic structures. For instance, in the context of sentiment analysis, word embedding allows algorithms to understand the sentiment behind words and phrases by recognizing their proximity in the vector space.

The adoption of word embeddings in natural language processing tasks offers a myriad of advantages. Unlike traditional one-hot encoding, word embedding provides a dense

representation that preserves semantic nuances, facilitating more effective language understanding. For instance, the words "ナース (nurse)" and "介護者 (caregiver)" might be located closer in the embedding space, reflecting their semantic similarity. Moreover, the ability of word embedding models to generalize well enhances their performance on unseen data, making them robust across diverse applications. In machine translation, for example, word embedding assists in capturing cross-language semantic relationships, improving translation accuracy for words with similar meanings but different linguistic expressions[16]. Additionally, in information retrieval, word embedding enables more accurate matching of user queries with relevant documents by understanding the contextual similarities between words. As a result, word embedding stands as a pivotal technique, advancing the capabilities of computational linguistics and bolstering the efficiency of diverse natural language processing applications.

Several studies have assessed the efficacy of various word embedding models, including diverse linguistic contexts and applications. Investigations have ranged from comparing pre-trained word embedding vectors for word-level semantic text similarity in Turkish[49] to evaluating Neural Machine Translation (NMT) for languages such as English and Hindi[45]. Additionally, the accuracy of three prominent word embedding models within the context of Convolutional Neural Network (CNN) text classification[10] has been explored. The culmination of these studies suggests that the fastText word embedding model consistently outperforms its counterparts. In the specific domain of my study, focusing on care records composed of Japanese sentences employed by caregivers to report on the development and conditions of elderly patients, fastText emerges as the optimal choice. Its ability to handle infrequent or uncommon words by generating vectors for subwords makes fastText particularly adept in this scenario. The versatility and robustness exhibited by the fastText model underscore its effectiveness across a spectrum of linguistic tasks, making it a preferable choice in applications involving diverse and specialized vocabularies.

3.3 EmbedHDP

Our original motivation was to address the distinctive characteristics of our nursing care records' sentences, which often exhibit short, incompleteness, and a prevalence of medical terminology. The EmbedHDP model has proven to be highly effective in handling these challenges inherent in nursing care record sentences, as shown in Figure 3.2 and Figure 3.3. Consequently, our proposed evaluation metric has successfully achieved our intended

motivation or goal. The method has demonstrated enhanced accuracy compared to other evaluation metrics, signifying a notable advancement in the accurate evaluation of sentence suggestions within the unique context of nursing care records. Graphically, the architecture of our model can be visualized in Figure 3.1:

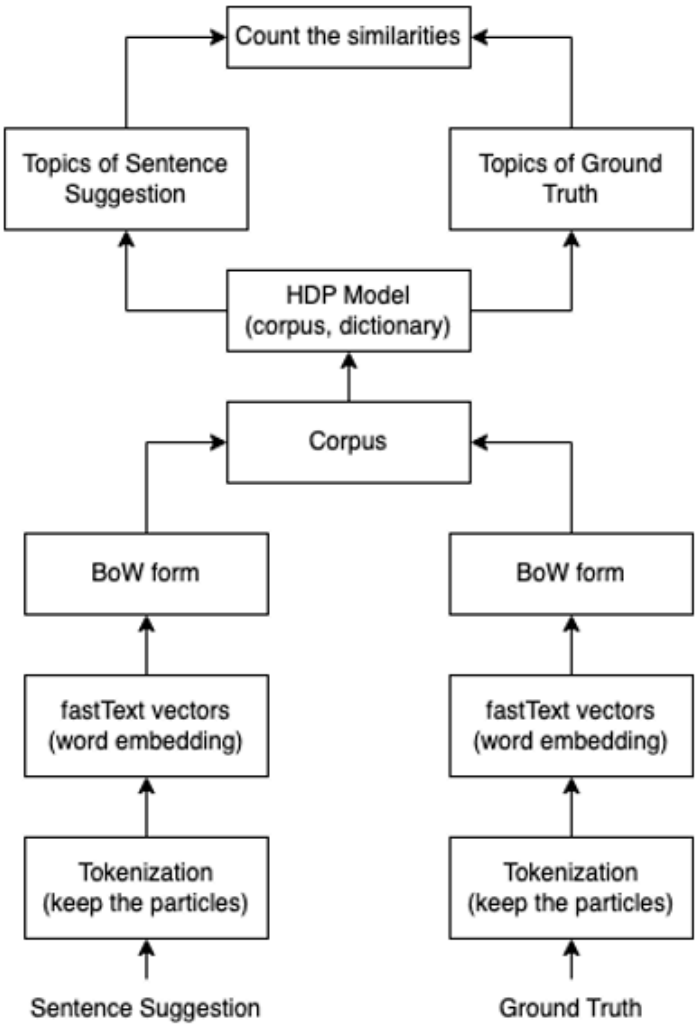

Fig. 3.1: EmbedHDP architecture

As illustrated in Figure 3.1, pre-processing sentences before being trained by HDP involves several key steps. Firstly, the sentence undergoes tokenization, a process where it is separated into individual tokens. We skip stemming and lemmatization at this stage while retaining particles attached to words. Once tokenized, these tokens are converted into vectors using fastText. Before processing them by the HDP model, we transform the vectors for each token into a Bag of Words (BoW) format. Afterwards, we merge the vectors from sentence suggestions with ground truth, treating them collectively as

the corpus. This comprehensive pre-processing workflow aims to prepare the data for optimal training and analysis by HDP, ensuring that the model is fed with comprehensively represented and formatted input. The EmbedHDP pre-processing can be explained in more detail as follows:

3.3.1 Tokenization

The Japanese language captivates attention with its unique linguistic structure, where verbs occupy the final position in sentences [28]. Additionally, the Japanese language employs special particles to indicate subjects, objects, or other additional information. In the tokenization process of Japanese, I utilize the Mecab library (-Owakati)[43]. I omit the lemmatizing and steaming processes. Another step taken in the tokenization process is to preserve the particles attached to each word. Linguistic particles in Japan refer to a distinctive feature of the Japanese language where small words or particles are used to convey grammatical relationships and nuances in a sentence[42]. These particles play a crucial role in indicating the subject, object, direction, or emphasis of a statement, and their presence significantly influences the overall meaning of a sentence. Here are the particles retained to remain attached to words during the tokenization process, as shown in the following table. This decision is made to ensure that the additional information encapsulated in these particles remains intact, avoiding loss during the analysis process and enabling optimal utilization.

Table 3.1: Functions of Several Particles and Verbs in Japanese

Japanese Particles	**Explanation and Example**
は (wa)	The topic particle that indicates the topic or subject of a sentence. For example, ”わたし は がくせい です” (Watashi wa gakusei desu) means ”I am a student.”
へ (e)	Indicates direction or destination. For instance, ”ともだち へ いきます” (Tomodachi e ikimasu) means ”I am going to a friend.”
で (de)	Indicates the place or method in which an action takes place. For example, ”レストラン で たべます” (Resutoran de tabemasu) means ”I eat at the restaurant.”
を (wo)	The object particle that indicates the object of an action. For example, ”りんご を たべます” (Ringo o tabemasu) means ”I eat an apple.”
の (no)	The possessive particle or connector between two nouns. For example, ”わたし の くるま” (Watashi no kuruma) means ”My car.”
ある (aru)	A verb indicating existence or possession. For example, ”ほん が あります” (Hon ga arimasu) means ”There is a book.”
あり (ari)	The past or formal form of the verb ”ある” (aru) indicating existence.
する (suru)	A common verb meaning ”to do.” For example, ”しゅくだい を する” (Shukudai o suru) means ”To do homework.”
なる (naru)	A verb meaning ”to become.” For example, ”せんせい に なりたい” (Sensei ni naritai) means ”I want to become a teacher.”
し (shi)	A conjunction used to express two related actions or qualities. For example, ”りんご し いちご” (Ringo shi Ichigo) means ”Apples and strawberries.”

Table 3.2: Functions of Several Particles and Verbs in Japanese (cont.)

Japanese Particles	**Explanation and Example**
て (te)	The te-form of a verb, indicating an ongoing action. For example, "たべ て います" (Tabete imasu) means "I am eating."
ます (masu)	A polite form of verbs indicating present actions. For example, "たべ ます" (Tabemasu) means "I eat" or "I will eat."

3.3.2 Creating Corpus

The corpus is the most crucial element in training HDP to derive topics. By default, corpus generation involves converting tokens within sentences using the BoW model. To achieve optimal results in HDP and facilitate the comparison of similarity between two sentences, the corpus is generated with the assistance of the fastText model. Specifically, we use the cc.ja.300.bin, which encompasses a 7 GB vector in the Japanese language. An advantage of this model is its capability to generate vectors even for less familiar words, such as "介護者". This attribute enhances the model's versatility and ensures comprehensive coverage in vector representation.

The additional challenge is that the HDP model exclusively accepts the BoW format. This implies a direct processing barrier for vectors generated by fastText into the HDP model. The subsequent steps to overcome this hurdle involve converting the vectors into BoW format with the following stipulations:

1. **Set Vector Length:** Assign a fixed vector length in the BoW format, specifically 10. This decision is grounded in the consideration that our sentences are not excessively long, thereby mitigating potential biases arising from vector length discrepancies.
2. **Highest Frequency Elements:** Select elements based on their highest frequencies under the assumption that the highest frequency serves as a representative token for each element.
3. **Scaling Factor:** Due to the considerable length of vectors produced by fastText, the resultant BoW-formatted vectors become exceedingly small (0.000x). This phenomenon leads to nearly identical topics when trained in the HDP model. To

counteract this issue, each vector is multiplied by 100, ensuring positive values throughout the vector and resolving the disparity.

4. **BoW Representation:** The outcome of these steps is the acquisition of BoW-formatted vector representations for each token in both sentences. This transformation facilitates seamless compatibility with the HDP model during the training process.

3.3.3 Dictionary

The Hierarchical Dirichlet Process (HDP) is a statistical model addressing resource allocation challenges in data clustering. In the context of HDP, the term "dictionary" refers to a stochastic distribution concept employed to represent the distribution of topics within a dataset. Specifically, within the EmbedHDP framework, the dictionary encapsulates the local topic distribution specific to the nursing care record dataset. The dictionary fundamentally plays a role in determining the extent of the topics present across the entire dataset and the proportion of topics applicable to specific data clusters.

In this study, we comprehensively compiled a set of 268 unique words that profoundly represent nursing care record applications. These words, such as リドメックスローション (redomex lotion), 病院 (hospital), リハビリテーション (rehabilitation), 不安 (anxiety), 感染 (infection), and others, were strategically chosen to represent diverse aspects of nursing care records. Incorporating such a dictionary in the model contributes to a nuanced understanding of the topics prevalent within nursing care record applications, fostering insights into the unique linguistic characteristics inherent in this domain.

Due to the challenges posed by incomplete or fragmented sentences, hierarchy in EmbedHDP provides flexibility because it does not require prior specifications regarding the number of groups or topics available. This is also an important point in non-standard sentence structures. The hierarchy provides flexibility and adaptability, making it an effective tool for modeling data with complex and uncertain structures. Here is an example of using EmbedHDP that yields useful assessments in incomplete sentences.

Table 3.3: Sample 1 illustrates how HDP can effectively address incomplete or fragmented sentences.

Sentence Suggestion	**Ground Truth**
コルセット作ることを報告する (report making a corset)	コルセットを作ることを勧められる (advised making a corset)

Here are the respective scores assigned by humans as a benchmark, EmbedHDP, and several other current evaluation metrics.

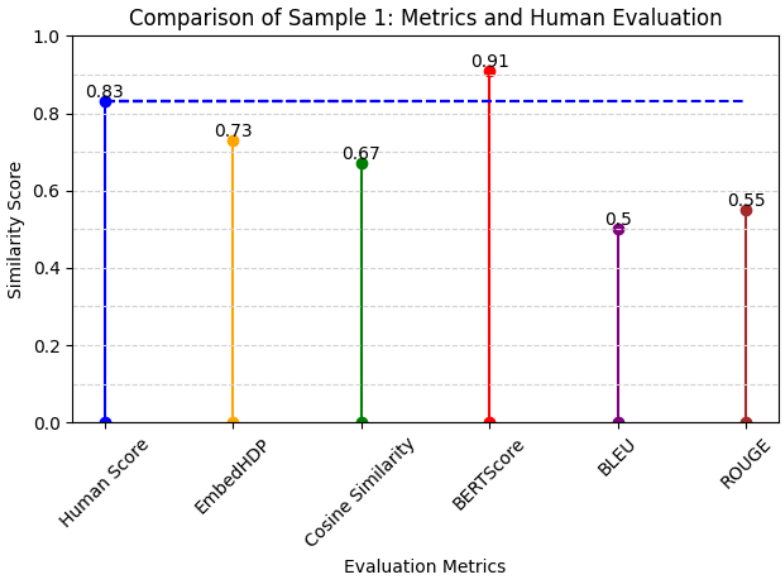

Fig. 3.2: Metrics and human evaluation assessment of Sample 1

The human score is a benchmark, reflecting the expert opinion and nuanced understanding required in the context of care records. EmbedHDP introduces a unique approach by leveraging hierarchical Dirichlet processes and domain-specific dictionaries, showcasing its potential to align closely with human assessments. Incorporating additional metrics like BERTScore, cosine similarity, ROUGE, and BLEU further enriches the evaluation process, enabling a more nuanced and comprehensive analysis of the model's performance.

Another challenge in care records involves the presence of medical terminology within sentences. EmbedHDP can address this challenge using a dictionary containing words relevant to care records. The dictionary is utilized during model training to compute potential topic distributions for each word within the sentences. Additionally, word embedding models play a crucial role, as their vector representation strength enables the

capture of semantic meaning in individual words. Essentially, these vector representations lie in their ability to bring vectors of words with similar meanings closer together in vector space. Here is an example of how word embeddings help provide a significant assessment based on expert opinion. This is a sample of data derived from word embedding-based scoring optimization.

Table 3.4: Sample 2 illustrates how word embedding can effectively address the similarity of words in both sentences.

Sentence Suggestion	**Ground Truth**
熱があったので、看護師に報告して中止しました。 (I had a fever, so I informed the nurse and cancelled the session.)	熱発の為、ナースに報告し中止。 (Due to fever, we informed the nurse and discontinued the treatment.)

Here are the respective scores assigned by humans as a benchmark, EmbedHDP, and several other current evaluation metrics for Sample 2.

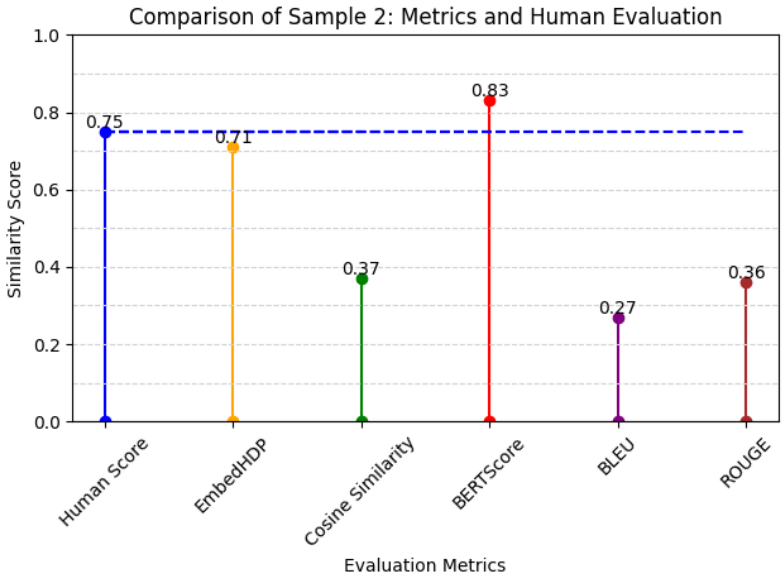

Fig. 3.3: Metrics and human evaluation assessment of Sample 2

From the above example, we can analyze that EmbedHDP can provide relevant assessments closely resembling human evaluations.

However, EmbedHDP model may not be as effective when dealing with relatively long sentences or sentences that consist of more than 14 words. This limitation may arise

from the substantial amount of information contained within lengthy sentences, making it challenging to capture the semantic nuances comprehensively across the entire sentence.

Table 3.5: Sample 3 illustrates how sentence length affects the assessment quality of the model.

Sentence Suggestion	**Ground Truth**
両眼内障であること、右眼は緑内障疑いで眼圧が高くなっては弱い痛み止めを屯用で出しておくので飲んで心臓の状態が良いとの連絡あり (I was informed that I have bilateral eye disorders, that my right eye is suspected of having glaucoma, and that my intraocular pressure is high, so they give me a weak painkiller to take, and that my heart is in good condition.)	両眼白内障であること、右眼は緑内障疑いで眼圧が高くなっていること、だから目が見えにくくなっている、と説明を受けられ、眼圧を下げる点眼薬を処方されたこと (He explained to me that he had cataracts in both eyes, that his right eye had high intraocular pressure due to suspected glaucoma, and that he was having difficulty seeing and was prescribed eye drops to lower the intraocular pressure.)

Here are the respective scores assigned by humans as a benchmark: EmbedHDP and several other current evaluation metrics for Sample 3.

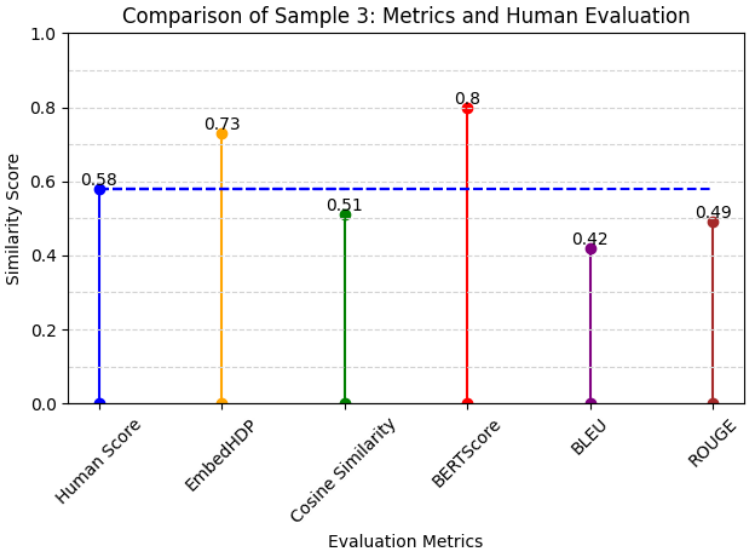

Fig. 3.4: Metrics and human evaluation assessment of Sample 3

Chapter 4

Data Collection

This chapter will delve into the proposed evaluation model, EmbedHDP. The functional mechanisms of EmbedHDP have demonstrated their effectiveness in addressing challenges inherent in elderly care records.

4.1 Overview of Proposed Framework for Sentence Suggestion in Nursing Care Record Applications

The proposed framework aims to enhance the efficiency and accuracy of nursing care record applications by integrating a sophisticated model for sentence suggestion. At its core, this framework leverages the input of caregivers who contribute vital notices, initiating a comprehensive data manipulation and analysis process. Firstly, the collected textual data undergoes pre-processing, wherein it is segmented into tokens. These tokens are then systematically incorporated into dictionaries encompassing initial words, second words, and transition words, each with its corresponding percentage, a fundamental aspect of our sentence suggestion model. Subsequently, as caregivers input initial words, the model dynamically generates sentences by calculating the highest probability words following the input and iteratively building a coherent sentence. This approach streamlines the recording process and adapts to individual caregiver preferences.

Furthermore, the generated sentence suggestions are systematically compared with those originating text from the application, utilizing a topic similarity calculation. The ensuing similarity scores are then juxtaposed against caregiver-assigned scores, employing evaluation metrics to discern the correlation coefficients. This systematic framework aims to establish a robust and user-oriented approach to sentence suggestion within nursing care

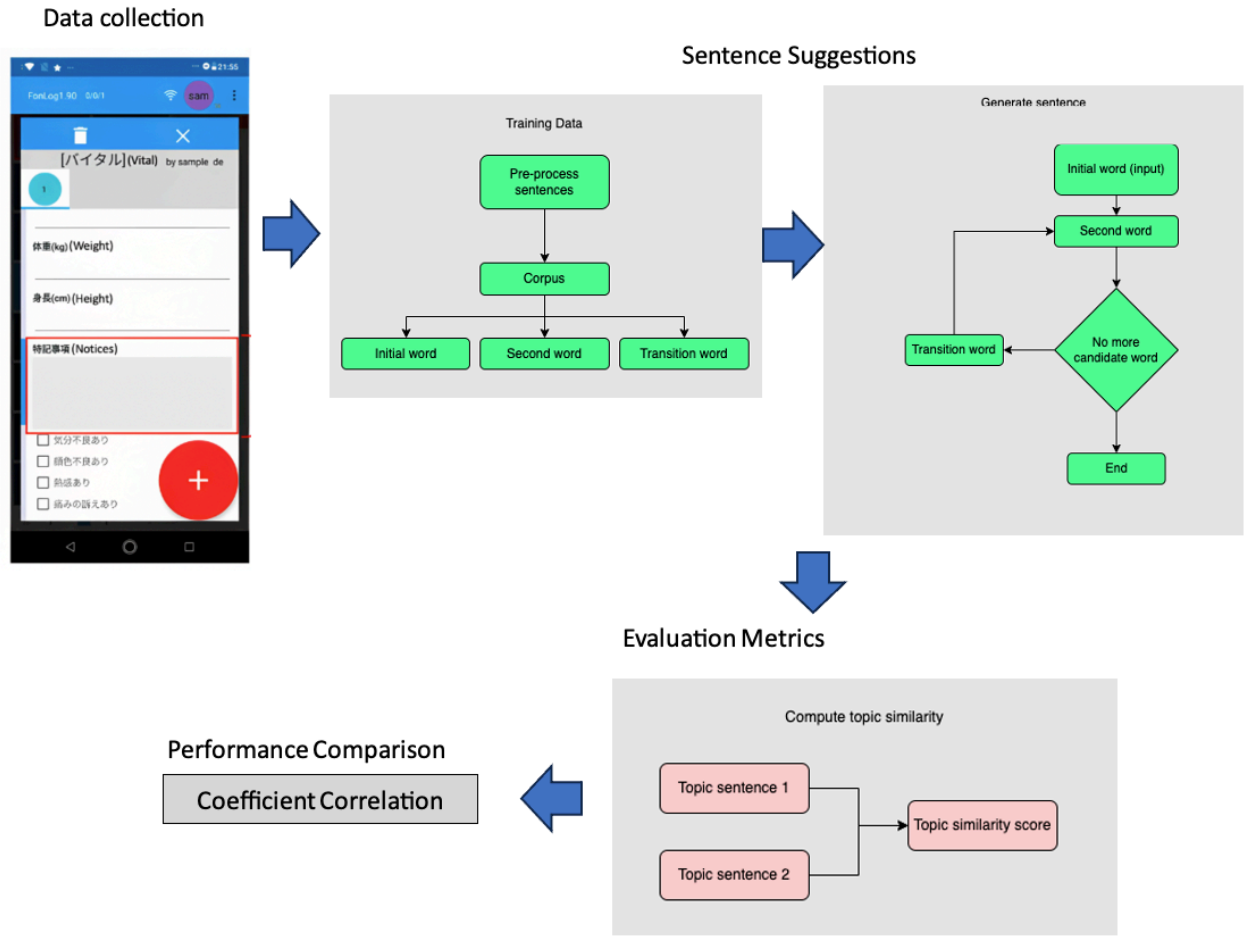

Fig. 4.1: Overview of Proposed Sentence Suggestion Evaluation in Nursing Care Record Application

In the realm of caregiving, the data collected by caregivers through the use of FonLog in the 特記事項 column of vital activity types serves as the ground truth. This data serves as the foundational information processed by the Markov model to create a robust model for sentence suggestion. The sentences generated by the sentence suggestion undergo a rigorous comparison with the ground truth, with evaluations carried out by caregivers serving as the benchmark. Subsequently, the sentences from both the sentence suggestion and ground truth are subjected to various model evaluation metrics. The scores assigned by caregivers and the evaluation metrics are then juxtaposed using correlation coefficients. This comparative analysis provides insights into which evaluation metrics align more closely with the assessments made by caregivers, shedding light on the efficacy of the model in replicating human judgment. Through this process, a more nuanced understanding emerges regarding the reliability and accuracy of the Markov model in sentence suggestions for caregiving contexts.

4.2 Data Collection Tools

In the general overview of the patient information recording process by caregivers concerning vital activity, caregivers access an integrated mobile care record application within the FonLog system. Subsequently, the caregiver initiates the patient identification process by selecting unique identification data, such as the patient's identification number. This ensures that the recorded information is associated with the correct patient record. Following this, the caregiver proceeds to choose the vital activity type, within which various columns are available for input. These columns typically encompass parameters such as blood pressure, heart rate, body temperature, and other vital metrics. The caregiver may also utilize an additional notes section (特記事項) to include specific remarks or crucial information about the patient. This feature provides supplementary space for details that may not be captured within standard fields. Once the necessary data is input, the caregiver saves the recorded information within the application, completing the documentation process. This systematic approach ensures accuracy, thoroughness, and secure storage of patient-related data, contributing to the overall efficiency of healthcare record management.

Fig. 4.2: Vital Activity Type in FonLog Application

FonLog's special notes (特記事項) input is intended to capture more elderly information to allow caregivers to report specific patient conditions during activities. Caregivers can

provide information in their language through notices, which provide a free-form input field. By providing caregivers with sentence suggestions for filling in the notice input, the caregiver's task will undoubtedly be more efficient and effective in terms of time and quality of records. Figure 4.3 shows the notices input for the vital activity type, which records extra information about the patient's vital activity.

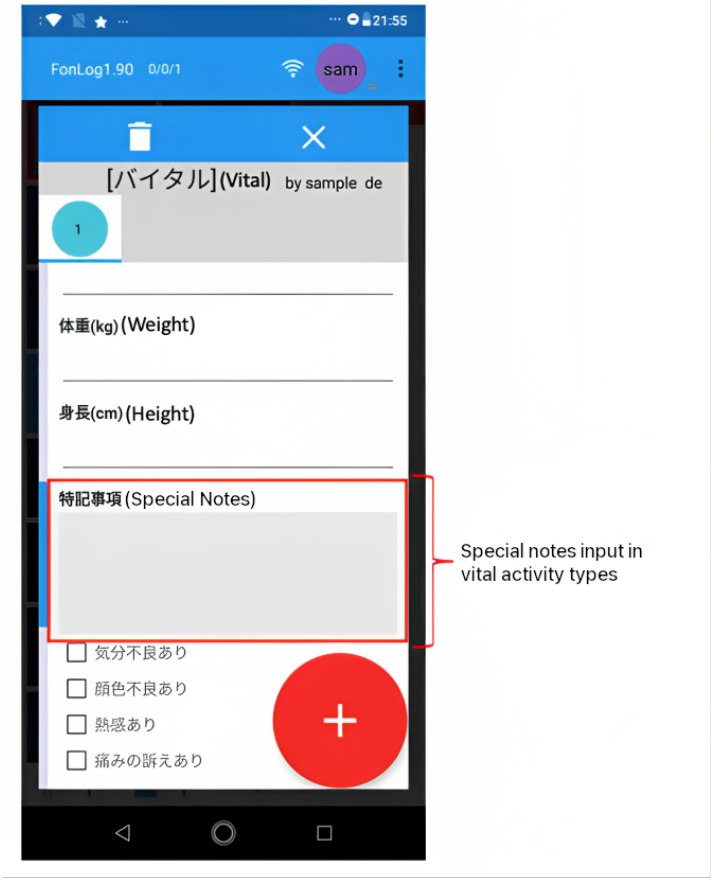

Fig. 4.3: Special notes (特記事項) in Vital Activity Type

In the context of nursing care record application, the column 特記事項 (special notes) within the vital activity type serves as a crucial component utilized by caregivers to provide detailed reports concerning the essential aspects of a patient's well-being. Caregivers leverage this column to document and communicate noteworthy information, enabling them to convey nuanced details about a patient's vital signs, medical condition, and any pertinent observations[23]. The inclusion of special notes in the vital activity type column enhances the comprehensiveness of patient reporting, facilitating a more thorough understanding of the individual's health status[52]. This structured approach not only aids in real-time patient monitoring but also contributes to a comprehensive and cohesive healthcare record. The emphasis on the 特記事項 column underscores its significance as

a dedicated space for caregivers to articulate specific details that may require attention, facilitating effective communication and coordination within the healthcare ecosystem.

4.3 Expert-Generated Human Evaluation Metrics

In this study, human evaluation was conducted to compare machine performance results with the human perspective assessment. To develop the model's performance evaluation, we generated a total of 390 sentence suggestions using both the Markov model and ChatGPT. These suggestions were then paired with 390 sentences from the care record dataset, serving as the ground truth. To obtain assessments based on meaningfulness and utility, we engaged the participation of three experienced caregivers from an elderly care facility in Japan. These caregivers provided their evaluations through the distribution of questionnaires, covering inquiries designed to assess the relevance, clarity, and accuracy of the sentence suggestions generated by the model. In detail, the annotators gave a similarity assessment between the sentence suggestion and the ground truth using the Absolute Category Rating [21] [11] approach as follows:

Table 4.1: Assessment of similarity between sentence suggestions and ground truth

Indicators of assessment	**Description**
Excellent	perfectly similar with a score of 100%
Good	sufficiently similar with a score of 75%
Fair	less similar with a score of 50%
Poor	not similar with a score of 25%
Bad	totally not similar with score 0%

Chapter 5

Evaluation

In our research efforts, the evaluation section is a critical point where the efficacy and relevance of our proposed methodologies are carefully analyzed. This section serves as a crucible in which we rigorously evaluate the generated sentence suggestions for nursing care records, recognizing the complexity of healthcare documentation. As we navigate through this evaluation phase, we face the dual challenge of addressing the lack of dedicated metrics for this specific domain and ensuring that our approach meets the imperative benchmarks of accuracy, coherence, and relevance within the nuanced context of nursing care records. The subsequent analysis and insights derived from our evaluation process not only contribute to the refinement of our proposed model, but also illuminate the broader landscape of natural language generation in healthcare. In this section, we delve into the intricacies of our evaluation framework, acknowledging the existing limitations of metrics and delineating the nuanced considerations required for effective and domain-specific evaluation.

5.1 Goal

The overall goal of the EmbedHDP proposal is to improve the metrics for evaluating the quality of sentence suggestions in nursing records. This initiative is driven by the recognition of the unique characteristics of nursing records, which vary in sentence length and contain medical terminology. The primary objective is to provide a systematic means of evaluating the sentence suggestions generated, particularly in the context of reporting the health status of elderly people. Given the critical nature of this information, it is imperative to ensure the accuracy and precision of the generated content in order to mitigate the potential risks associated with erroneous data.

EmbedHDP is considered successful when the evaluation results closely match or reflect those of the human evaluation by caregivers. Evaluation of its effectiveness depends on achieving a high degree of correlation between the model's scores and those of human scores. The use of coefficient correlation serves as a quantitative measure of the closeness and agreement of these scores. This agreement is critical because sentence suggestions play a critical role in conveying accurate health-related information, especially for the elderly. Any inaccuracies or discrepancies in the generated content can pose significant risks.

The realization of the intended goal of EmbedHDP has been achieved through a careful integration of various elements and considerations. The proposed framework shows a remarkable ability to evaluate the quality of sentence suggestions in care records, which is challenging due to varying sentence lengths and the inclusion of medical terminology. The effectiveness of EmbedHDP is exemplified by its ability to approximate or closely match expert judgments, establishing a commendable degree of correlation. The adoption of coefficient correlation serves as a quantitative metric that validates the model's ability to align its ratings with those of domain experts.

The success of EmbedHDP in this context is promising because of the critical role of sentence suggestions in conveying accurate and reliable information about the health status of the elderly. Achieving the goal of the model is paramount, as inaccuracies in health-related information can have serious consequences. EmbedHDP has proven to be a robust and valuable tool in the field of healthcare documentation, helping to generate trustworthy sentence suggestions within care records. This achievement is a testament to the model's adaptability and effectiveness in dealing with the nuances of healthcare language, ultimately increasing the reliability and utility of the generated content in the reporting of elderly health conditions.

5.2 The Design of Evaluation

The design of our evaluation methodology is meticulously structured to ensure a comprehensive and reliable assessment of sentence suggestions within nursing care records, utilizing both expert judgment and quantitative metrics. To establish a benchmark, the caregivers actively engaged in elderly care facilities were selected as evaluators. Three caregivers participated in the evaluation process, each contributing valuable insights through a carefully crafted questionnaire comprising 390 pairs of words, comparing sentence suggestions against corresponding ground truth sentences.

The recorded data within the nursing care records application provides ground truth sentences that serve as reference points to determine the accuracy and fidelity of the generated sentence suggestions. Caregivers diligently compared and assessed each pair of sentence suggestions and ground truth sentences according to the criteria outlined in subsection 3.3.

The caregivers' evaluations serve as a benchmark for the quality of sentence suggestions, offering a human-centered perspective on the generated sentences. These assessments are then compared with alternative evaluation metrics, including EmbedHDP, which we propose as a novel metric, along with established metrics such as BERTScore, cosine similarity, ROUGE, and BLEU.

The coefficient correlation is used to measure the agreement between caregivers' evaluations and those provided by each evaluation metric. A higher coefficient correlation score indicates a closer alignment between the metric's assessments and the judgments made by caregivers. This alignment is important because it shows how well the evaluation metrics, including EmbedHDP, match the subjective assessments of caregivers.

The correlation coefficient is used to measure the agreement between caregivers' evaluations and those provided by each evaluation metric. The correlation coefficient is a measure of the degree of linear relationship between two variables x and y [2]. It ranges from -1 to $+1$, where -1 indicates a perfect negative linear relationship, $+1$ indicates a perfect positive linear relationship, and 0 indicates no linear relationship. A higher correlation coefficient score indicates a closer alignment between the metric's assessments and the caregivers' judgments. This alignment is important because it shows how well the evaluation metrics, including EmbedHDP, match the caregivers' judgments.

Essentially, a higher correlation coefficient score indicates that the evaluation metrics are approaching the judgments made by caregivers. The purpose of using these metrics is to establish an objective and standardized way of evaluating sentence suggestions in nursing care record applications. This methodology improves the reliability and objectivity of the evaluation process by reducing dependence on caregiver assessments. The assessment ensures that the model's ability to generate accurate and contextually relevant sentences is rigorously and quantifiably evaluated.

5.3 Filtering Data Sample

In Section 3, we discussed the limitations of EmbedHDP when handling relatively long sentences. Due to the potential challenge posed by longer sentences, which may contain more intricate information, it can be challenging for a model to capture the comprehensive semantic meaning of the entire sentence effectively. We have implemented a data filtering criterion for testing our proposed evaluation model in response to this consideration. Specifically, we stipulate that sentences comprised of 13 words or fewer will be used in the testing phase. This limitation is imposed to ensure the evaluation model is assessed under conditions where sentences are relatively concise. Focusing on shorter sentences facilitates a more targeted evaluation of the model's ability to understand and generate content with optimal relevance and precision within a constrained linguistic scope. The filtration process can be automated with the following pseudo-code:

Algorithm 4: Filtering Data Sample

Input: sentence1 ← sentence similarity, sentence2 ← ground truth

Output: 13-word or fewer sentences

if *len(sentence1) > 13 or len(sentence2) > 13:* **then**

_

return Both sentences eliminated

Based on the aforementioned conditions, the initial dataset, which originally comprised 390 data, has been reduced to 320 data due to the imposed criteria. Additionally, 70 data points have been identified as outliers and subsequently excluded from the dataset. In statistical analysis, identifying and handling outliers is a common practice to ensure the robustness and reliability of the data. Outliers, which are data points significantly different from most of the dataset, can substantially impact statistical measures. By excluding these outliers based on the specified criteria, the dataset has been refined to a more representative and manageable size, consisting of 320 data entries. This process contributes to more accurate analysis and interpretation of the dataset, aligning with best practices in data preprocessing.

5.4 Results

Based on the comprehensive exposition above, we have conveyed that EmbedHDP is a potential solution for evaluating models applied to care record sentences, addressing two primary challenges. Furthermore, we have substantiated that EmbedHDP has successfully yielded assessments that align more closely with expert opinions than other evaluation metrics.

The challenges in evaluating care record sentences, such as diverse sentence structures and specialized medical terminology, necessitate a model that can discern nuances effectively. EmbedHDP, through its incorporation of hierarchical Dirichlet processes and domain-specific dictionaries, demonstrates a capacity to navigate these intricacies.

The comparison with current evaluation metrics underscores the superiority of EmbedHDP in capturing the nuanced nature of care record sentences. Its success in producing evaluations that closely approximate expert opinions reflects its potential to contribute to more accurate and meaningful assessments for sentence suggestion in care record application. Table 5.1 below shows that EmbedHDP outperforms current evaluation metrics in coefficient score on 320 test data.

Table 5.1: EmbedHDP outperforms other evaluation metrics

Evaluation Metrics	**Correlation Coefficient**
EmbedHDP	**0.61**
BERTScore	0.58
ROUGE	0.57
Cosine Similarity	0.59
BLEU	0.53

EmbedHDP has surpassed current evaluation metrics when assessing the quality of sentence suggestion generation against the corresponding ground truth. Employing coefficient correlation parameters in human evaluation, EmbedHDP outperforms other evaluation methods with a score of 61%, followed by cosine similarity at 59%, and BERTScore at 58%. This substantiates the effectiveness of EmbedHDP as the proposed primary evaluation metric for assessing sentence suggestion generation in care records. Notably, the observed higher linear relationship between EmbedHDP and human scores compared to

other evaluation models underscores its robust performance in capturing the nuances of human expert opinions.

With 70 identified outliers, we classify them as one of the limitations in both EmbedHDP and other evaluation metrics. Outliers can pose a significant challenge in data evaluation and analysis, including utilizing EmbedHDP and current evaluation metrics. Outliers can impact evaluation results significantly, particularly if the model or metric is not designed to handle extreme variability. In the context of EmbedHDP, identifying and addressing outliers may become a focus of future development to enhance the model's robustness against unusual data variations. The following Table 5.2 shows the correlation coefficient for 70 data as outliers.

Table 5.2: Limitation of EmbedHDP to sentences of 14 or more words

Evaluation Metrics	**Correlation Coefficient**
EmbedHDP	0.25
BERTScore	**0.35**
ROUGE	0.34
Cosine Similarity	**0.35**
BLEU	0.34

5.5 Benchmarking Method

In the preceding section, we delved into the mechanisms and limitations inherent in current evaluation metrics. Additionally, we explored the challenges posed by care record sentences and elucidated how our proposed evaluation model, EmbedHDP, is poised to address these challenges. Examining current evaluation metrics provided insights into their operational mechanisms and constraints. This understanding sets the stage for introducing and justification our proposed model, EmbedHDP, which offers a novel approach to evaluating sentence suggestions within the context of care record sentences. By acknowledging and addressing the specific challenges posed by the diverse sentence structures and specialized medical terminology in care records, EmbedHDP aims to provide a more nuanced and contextually relevant evaluation. The following example will illustrate how EmbedHDP can address some of the limitations inherent in current evaluation metrics when facing the challenges posed by nursing care records:

1. The example of BERTScore limitations when evaluating short sentences (diverse sentence structures) and those containing medical information (specialized medical terminology).

Table 5.3: An example of BERTScore limitation.

Sentence Suggestion	**Ground Truth**
頻繁な少量の排尿。 (Frequent small amount of urination)	排便中量あり。 (There was a large amount during defecation.)

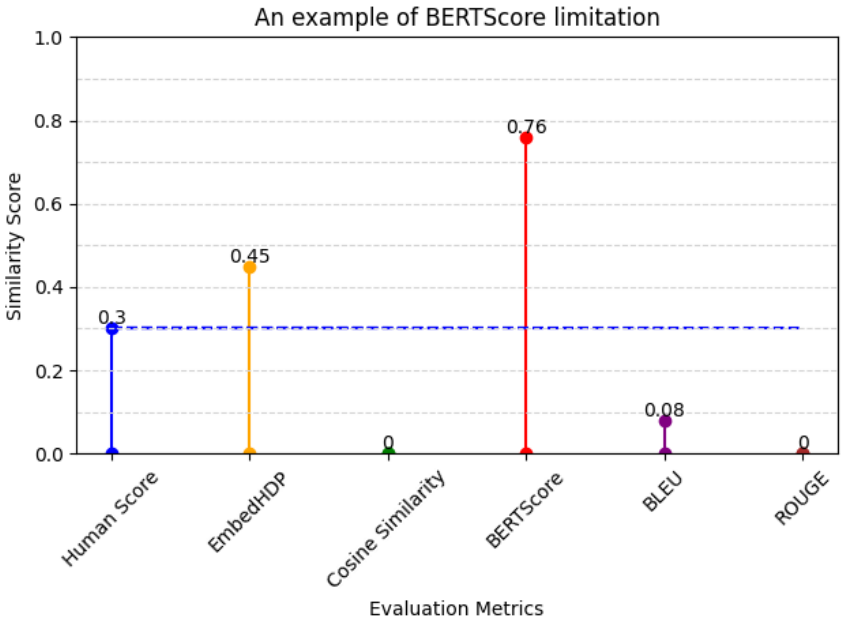

Fig. 5.1: Metrics and human evaluation assessment of BERTScore limitation in sentences

2. The example of Cosine Similarity limitations when evaluating short sentences (diverse sentence structures) and those containing medical information (specialized medical terminology).

Table 5.4: An example of Cosine Similarity limitation.

Sentence Suggestion	**Ground Truth**
吐き気あり報告入れる。 (report nurse)	吐き気訴えあり。 (complaints of nurse)

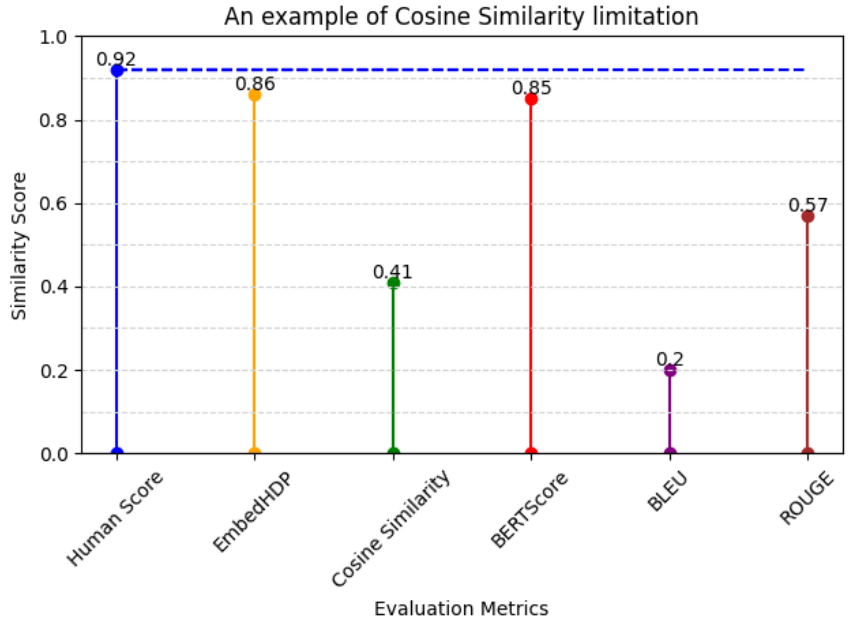

Fig. 5.2: Metrics and human evaluation assessment of cosine similarity limitation in sentences

3. The example of ROUGE limitations when evaluating different structures of sentences (diverse sentence structures).

Table 5.5: An example of ROUGE limitation.

Sentence Suggestion	**Ground Truth**
気分訴えなし (no mood complaints)	気分不良はないと本人言われる (he says he doesn't feel unwell)

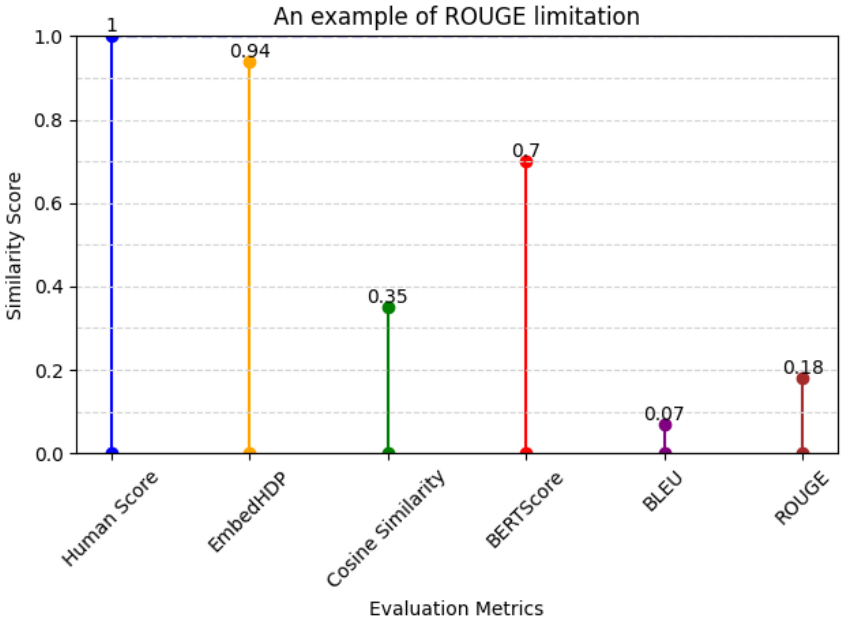

Fig. 5.3: Metrics and human evaluation assessment of ROUGE limitation in sentences

4. The example of BLEU limitations when evaluating different structures of sentences (diverse sentence structures) and those containing medical information (specialized medical terminology).

Table 5.6: An example of BLEU limitation.

Sentence Suggestion	**Ground Truth**
熱があったので、看護師に報告して中止しました。 (I had a fever, so I informed the nurse and cancelled the session)	熱発の為、ナースに報告し中止。 (Due to fever, the nurse was informed and the procedure was discontinued)

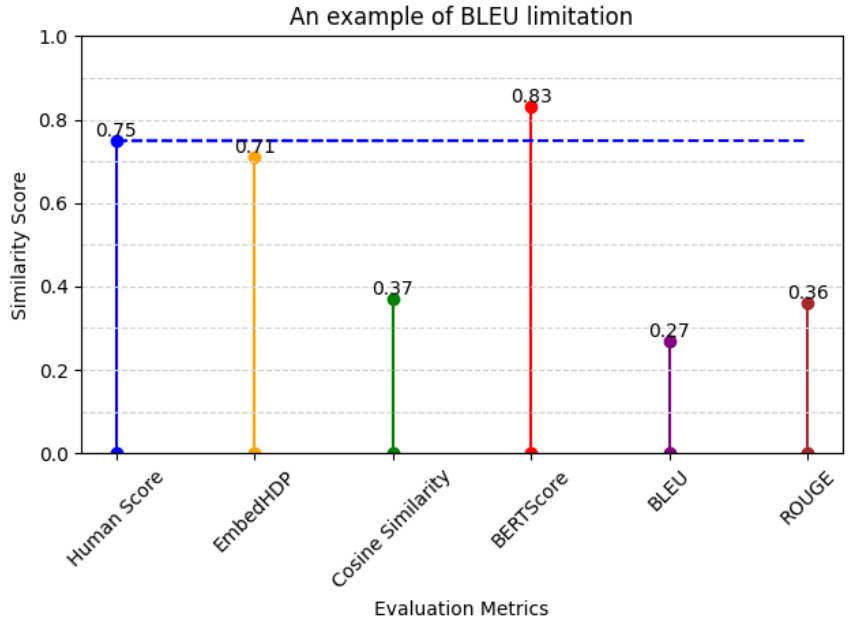

Fig. 5.4: Metrics and human evaluation assessment of BLEU limitation in sentences

5.6 Discussion

In this paper, our goal is to provide evaluation metrics designed to assess the quality of sentence suggestions in nursing care record applications, specifically tailored to record information related to the elderly. Nursing care records present specific challenges, including diverse sentence structures and specialized medical terminology. We present evaluation metrics that address the two main challenges in nursing care records by employing a methodology that computes topic similarity using word embedding vectors. This innovative approach aims to overcome the challenges of diverse sentence structures and specialized medical terminology in nursing notes. By harnessing the power of word embedding vectors, our proposed evaluation metrics strive to capture the semantic nuances and context-specific information inherent in healthcare-related texts, allowing for a more accurate and contextually relevant evaluation of sentence suggestions in the domain of care records.

It is important to recognize that while EmbedHDP demonstrates remarkable capabilities, it is subject to certain limitations that require careful analysis for future refinement. A significant challenge lies in EmbedHDP's limitations in analyzing the similarity between two sentences within health records, particularly when the sentences extend to 14 words

or more. This decision is based on the observation that longer sentences tend to encapsulate more complex and overlapping information. As a result, EmbedHDP needs to be extended and enhanced to address the challenges posed by longer sentences effectively. This strategic evolution aims to strengthen EmbedHDP's ability to handle more complex linguistic structures within care records, ensuring its continued effectiveness in generating accurate and contextually relevant sentence suggestions.

As shown in Table 2.1 and Table 2.2, there are 31 activity types within the FonLog nursing record application for which sentence suggestions must be generated. However, for our current study, we used only the dataset associated with the vital (バイタル) activity type. While this subset is considered representative of the overall structure of care record sentences, it is recognized that additional conditions or contexts within other activity types could further enrich EmbedHDP's training.

The applicability of EmbedHDP to domains other than nursing records requires careful consideration, particularly in evaluating its adaptability to different data characteristics.EmbedHDP has effectively evaluated sentences with short and seemingly incomplete structures, where sentences may lack subjects or objects. This means that domains with similar characteristics can be adapted. The model's ability to handle incomplete sentences is consistent with scenarios where linguistic structures may vary.

In addition, the relevance of the task should be reviewed, as EmbedHDP was initially designed to evaluate sentence suggestions within nursing records. Specific domain-related words in care records are acceptable for the model's performance, especially when evaluating the similarity between sentence suggestions and ground truth within the context of nursing care records. The incorporation of word embeddings and dictionaries proves to be instrumental in capturing the nuances of domain-specific language. A key consideration is the involvement of domain experts who can provide valuable insights to improve the relevance and effectiveness of the model. Therefore, before extending the application of EmbedHDP to different domains, a thorough understanding of data characteristics, task relevance, and expert involvement is essential for successful adaptation.

5.7 Conclusion

In this study, the researcher proposes an evaluation metric that is close to human judgment and can achieve a high degree of correlation by using correlation coefficients. In this way, we aim to eliminate human involvement in the evaluation of sentence suggestions in

nursing care record applications. However, we would like to emphasize that the role of human judgment remains very important in the context of this research. Nursing care records have special characteristics such that most of the information produced can only be understood in depth by nurses, caregivers, or parties directly involved in the provision of health care services. This becomes even more important when considering that the information produced is directly related to the health of the elderly, where misinformation can have serious and even life-threatening consequences.Therefore, although the proposed evaluation metrics can reduce the burden of human involvement, evaluation by health professionals remains a key point to ensure the accuracy and safety of the resulting information.

Three important variables serve as the basis of EmbedHDP, contributing to its robust functionality in evaluating sentence suggestions in nursing care records. Firstly, the hierarchical approach establishes the framework for handling data clustering within a hierarchical structure, offering adaptability to the inherent complexity and uncertainty present in the data. This hierarchical model facilitates the representation of nuanced relationships between topics, thereby enhancing the model's capacity to capture intricate patterns within the data.

Secondly, word embeddings are key to generating vectors with expansive semantic values. The embedding matrix systematically maps words or tokens to continuous vectors, fostering a distributed representation that encapsulates subtle semantic relationships between words. This approach enhances the model's ability to grasp contextual nuances, contributing to evaluating coherent and contextually relevant sentence suggestions.

Lastly, combining a dictionary with the Dirichlet distribution plays a pivotal role in determining the existence and prevalence of topics within each document. The dictionary provides a repository of words systematically paired with the Dirichlet distribution, guiding the allocation of topics and influencing the overall thematic composition of the generated sentences. This synergistic interplay of hierarchical structures, word embeddings, and dictionary-based topic determination collectively forms the foundation of EmbedHDP, ensuring its adaptability, semantic richness, and effectiveness in generating contextually relevant sentence suggestions.

EmbedHDP has demonstrated a commendable leap forward in outperforming current evaluation metrics, as evidenced by its correlation coefficient of 61%. This achievement positions EmbedHDP as an effective evaluation metric for assessing the quality of sentence

suggestions within nursing care records. By outperforming established benchmarks, EmbedHDP represents a significant advancement in natural language generation, particularly in the complex landscape of healthcare documentation. In direct comparison, the cosine similarity metric yielded a score of 59%, while BERTScore achieved 58%, ROUGE at 57%, and BLEU at 53%. These results highlight EmbedHDP's ability to align with evaluations provided by caregivers, showcasing its enhanced capacity to capture the nuances of nursing care records and generate contextually relevant sentence suggestions.

This success demonstrates EmbedHDP's adaptability and effectiveness in handling healthcare language, which is characterized by diverse sentence structures and specialized medical terminology. The robust quantitative measure correlation coefficient indicates that EmbedHDP is skilled at approximating expert evaluations and outperforms other widely used metrics in this context.

EmbedHDP has successfully surpassed established metrics, addressing the critical need for specialized evaluation tools in the healthcare domain. Conventional metrics, such as cosine similarity, ROUGE, and BLEU, have limitations in capturing the domain-specific nuances of nursing care records, which emphasizes the necessity for tailored approaches. EmbedHDP's achievement, demonstrated by an enhanced correlation coefficient score, confirms its reliability and usefulness as a metric for evaluating sentence suggestions. This reduces reliance on assessments and enhances the objectivity and standardization of the evaluation process in nursing care record applications.

Chapter 6

Discusscion and Future Work

6.1 Discussion

In an effort to improve the efficiency and accuracy of sentence suggestion generation in nursing care records, this study aims to address several critical issues to improve the overall quality of automated language models. The unique challenges posed by the intricate and non-standard sentence structures, coupled with the specialized medical terminology inherent in healthcare documentation, necessitate a comprehensive investigation of the capabilities of existing models and the development of tailored evaluation metrics. Grounded in the context of elderly care, where accurate and contextually relevant reporting is paramount, our research aims to bridge the gap between automated sentence suggestions and the nuanced language of original care records.

The primary goal is to evaluate the effectiveness of current sentence suggestion systems and propose improvements that are more closely aligned with expert opinion and domain-specific requirements. By examining the limitations of existing models, understanding the complexities introduced by varying sentence structures and medical terminologies, and leveraging the insights of caregivers and healthcare professionals, this research seeks to contribute to the refinement and sophistication of language models tailored for nursing care records. The ensuing discussion explores the methods employed, the insights gained, and the advancements proposed, laying the groundwork for a nuanced understanding of the intricate language nuances within the realm of elder care documentation.

How effectively does the current sentence suggestion generation system capture the nuances of original sentences in care records?

The effectiveness of current sentence suggestion systems in capturing the nuances of

original sentences in nursing records is currently limited. Existing systems face challenges in dealing with the complex and non-standardized sentence structures prevalent in nursing care records. The variety of sentence lengths, coupled with the inclusion of specialized medical terminology, presents obstacles to traditional language models. These models, designed for conventional grammatical structures, may have difficulty accurately interpreting and preserving the subtle nuances embedded in the original sentences.

In addition, the specialized vocabulary used in healthcare documentation, ranging from specific drug names to detailed descriptions of medical conditions and treatment procedures, adds another layer of complexity. The nuances of healthcare language, specific to each patient's medical history and treatment plan, require a level of sophistication that current phrase suggestion systems may not fully achieve.

Comparative analysis with conventional evaluation metrics such as cosine similarity, BERTScore, ROUGE, and BLEU highlights the limitations of existing approaches in capturing the nuanced nature of nursing care records. The proposed EmbedHDP, with its commendable coefficient correlation score of 61%, indicates a notable advance in aligning with the subtleties embedded in the original sentences, suggesting a more effective approach.

In conclusion, while the current sentence suggestion generation systems demonstrate some capability, there is room for improvement, particularly in addressing the complex linguistic structures and specialized terminology inherent in nursing care records. The insights from the research suggest that more context-aware and comprehensive models, such as EmbedHDP, hold promise in enhancing the effectiveness of sentence suggestion systems in capturing the rich nuances of original sentences in care records.

How does evaluating the sentence suggestion generation model differ from expert opinions in the context of care records?

Sentence suggestion model evaluation differs from expert opinion in the context of care records in several key aspects. Expert opinion, often provided by caregivers or healthcare professionals, brings a nuanced and contextual understanding of the nuances within care records that automated models may not fully capture. Here are a few points of differentiation:

1. Subjectivity and contextual understanding: Expert opinions are subjective and based on contextual understanding of individual patient histories, treatment plans, and specific healthcare scenarios. Caregivers use their expertise to assess gener-

ated sentences' appropriateness, relevance, and accuracy based on their contextual knowledge. In contrast, automated models rely on pre-defined algorithms and may struggle to capture the nuanced context inherent in care records.

2. Handling Non-Standard Sentence Structures: Care records often exhibit non-standard sentence structures due to the varied nature of elderly histories, medical observations, and treatment plans. The caregivers, accustomed to these variations, can effectively interpret and assess the meaning of such sentences effectively. Automated models, especially traditional language models, may face challenges in accurately handling non-standard structures.
3. Specialized medical terminology: Nursing care records contain specialized medical terminology, from drug names to detailed descriptions of conditions and procedures. Caregivers deeply understand this terminology and can evaluate the accuracy and appropriateness of the language used. Automated models can struggle with the domain-specific nuances of medical discourse.
4. Holistic, elderly-centred evaluation: Caregivers' opinion often involves a holistic and elderly-centred evaluation that considers not only grammatical correctness but also the overall coherence and relevance of the sentences generated within the broader healthcare context. Automated models may prioritize grammatical rules and pre-defined metrics, potentially missing the broader elderly-centered perspective.
5. Flexibility in interpretation: Caregivers can adapt their interpretation based on the unique characteristics of each elder's care records. They can recognize the subtle differences in tone, intent, and urgency, which can be challenging for automated models to replicate without extensive training on diverse datasets.
6. Include for caregiver insights: Involving caregivers in the evaluation process allows for incorporating their insights, preferences, and professional judgment into the evaluation. This collaborative approach ensures that the evaluation aligns with the practical needs and expectations of those directly involved in elder care.

While automated models, such as EmbedHDP, provide a quantitative and systematic approach to evaluation, the nuanced and subjective nature of caregiver opinion remains essential to ensuring the appropriateness and effectiveness of sentence suggestions within nursing care records. Integrating both perspectives can lead to a more comprehensive and reliable evaluation process.

What methodologies can be employed to assess and enhance the overall quality of sen-

tence suggestions generated in care record applications?

In order to evaluate and improve the overall quality of sentence suggestions generated in nursing applications, several methods can be used to gain insight from the information provided:

1. Expert evaluation: Involving healthcare professionals and caregivers with expertise in elder care and healthcare documentation is critical. Their subjective evaluations can provide valuable insight into the appropriateness, accuracy, and contextual relevance of generated sentence suggestions.
2. Human-centric metrics: Incorporating human-centric metrics in the evaluation process, such as coherence, relevance, and overall comprehension, ensures that the generated sentences align with the broader healthcare context and elderly-specific nuances.
3. Specialized evaluation metrics: Developing and utilizing evaluation metrics tailored to the unique characteristics of nursing care records, including non-standard sentence structures and specialized medical terminology. EmbedHDP, with its emphasis on coefficient correlation, represents a step in this direction.
4. Context-aware language models: Use advanced natural language processing models, specifically designed to handle the nuances of care records. EmbedHDP, as proposed in the research, is a context-aware language model capable of capturing the nuances of healthcare language.
5. Iterative model refinement: Implement an iterative refinement process for language models, that incorporates feedback from expert evaluation and real-world use. This continuous improvement loop ensures that the model adapts to evolving healthcare documentation needs.
6. Benchmarking against ground truth: Use of benchmark datasets containing ground truth sentences derived from actual care records. Comparing the generated suggestions to these ground truth sentences provides a tangible measure of accuracy and relevance.
7. Incorporating domain-specific knowledge: Integrating domain-specific knowledge, such as medical ontologies and dictionaries, into language models. This ensures that the models have a robust understanding of the specialized vocabulary and terminology used in healthcare documentation.
8. Collaborative model development: Involves interdisciplinary collaboration between

natural language processing experts, healthcare professionals, and caregivers. This collaborative approach ensures that the language models are developed with a deep understanding of both linguistic nuances and practical healthcare requirements.

9. Utilizing Time Series Approaches: Exploring time series approaches, as demonstrated by Caballero and Akella, to predict health states from nursing care record applications. This approach allows the models to consider temporal aspects and changing health states in their predictions.

By combining these methods, a comprehensive evaluation framework can be established to continuously assess and enhance the quality of sentence suggestions in nursing care record applications. This holistic approach ensures that the language models not only adhere to linguistic standards but also meet the specific and evolving needs of healthcare professionals and caregivers in their daily practice.

Why did I choose to design EmbedHDP as the evaluation metric for sentence suggestion in care record applications?

The design process of EmbedHDP followed a comprehensive methodology that included several key steps aimed at creating an improved formula for evaluation metrics. The first steps included expert evaluation, exploration of specialized evaluation metrics, integration of context-aware language models, and benchmarking against ground truth sentences. The journey began with an interest in a topic recognition approach that focused on computing the similarity between the topics present in two sentences. Initially using Latent Dirichlet Allocation (LDA), the methodology had a limitation in that the number of topics had to be predetermined before training the sentences. To overcome this limitation, the Hierarchical Dirichlet Process (HDP) was introduced, exploiting its hierarchical advantages. Although EmbedHDP showed better evaluation results than LDA, it was still unable to surpass the performance of current evaluation metrics.

As we delved deeper into the workings of the existing evaluation metrics, we realized the critical role of word embeddings in understanding the context of individual words or sentences. This revelation became particularly important in the light of the prevalent problem of medical terminology, which is often encountered in care records. As a result, the methodology evolved, moving away from HDP and replacing the Bag-of-Words (BoW) process during corpus generation with word embeddings. The aim of this adaptation was to improve the understanding of context, especially in the complex domain of healthcare, and to address the challenges posed by specialized medical terminology. The development

of EmbedHDP reflects a strategic evolution that incorporates advances in natural language processing and harnesses the power of word embeddings to achieve a more nuanced and effective evaluation metric for sentence suggestion in nursing care record applications.

How did you find limitations in your research?

Several limitations were identified in this research, and one of the prominent challenges was the constraint on the number of caregivers involved, which was limited to three individuals. This limitation introduced complexities in establishing a robust benchmark for evaluation. For instance, when the three caregivers provided divergent assessments for a sample sentence, such as Caregiver A giving a rating of 0.75, Caregiver B assigning a score of 0.5, and Caregiver C providing a rating of 0.25, the challenge arose in reconciling these varied evaluations.

In scenarios where caregivers present contrasting opinions on a given sample sentence, the resulting discrepancies in the assigned scores highlight the subjectivity inherent in human evaluations. In the mentioned example, the discrepancies could stem from differences in individual interpretations, personal preferences, or varying levels of familiarity with the context. This subjectivity poses a challenge in establishing a definitive benchmark, as it becomes challenging to ascertain the "correct" evaluation for a given sentence.

Furthermore, another potential limitation arises when two caregivers reach a consensus on assigning high scores to a sample sentence, while a third caregiver assigns a lower score. In-depth analysis is necessary to understand the reasons behind such disparities. Possible explanations include variations in individual perceptions of clarity, relevance, or accuracy. Additionally, differences in expertise, experience, or contextual understanding among caregivers may contribute to divergent evaluations.

Addressing these limitations requires careful consideration of the inherent subjectivity in human evaluations. While human assessments provide valuable qualitative insights, their variability underscores the need for complementary quantitative evaluation metrics, such as EmbedHDP, to offer a more objective and standardized measure of the quality of sentence suggestions in care record applications. The combination of both subjective and objective evaluation approaches contributes to a comprehensive understanding of the model's performance, acknowledging the challenges associated with the inherently subjective nature of language evaluation.

What are the notable strengths and positive aspects of this research?

The research undertaken encompasses several commendable aspects that contribute to

its significance and potential impact. Firstly, the introduction of EmbedHDP as an evaluation metric for sentence suggestion in nursing care record applications represents a noteworthy innovation. EmbedHDP evolved from a meticulous methodology that integrates expert evaluations, specialized metrics, context-aware language models, and benchmarking against ground truth. This multi-faceted approach reflects a comprehensive effort to address the nuances and complexities inherent in nursing care records.

Furthermore, the research demonstrates a keen awareness of the limitations associated with human evaluations, particularly the challenges posed by the subjectivity of caregiver assessments. By acknowledging the potential discrepancies among caregiver ratings and the inherent difficulty in establishing a definitive benchmark, the research adds a layer of transparency to the evaluation process. This recognition of limitations contributes to a nuanced understanding of the complexities involved in evaluating sentence suggestions in care records.

Moreover, the evolution of EmbedHDP, transitioning from traditional topic modeling with HDP to incorporating word embeddings, showcases adaptability and responsiveness to the unique characteristics of healthcare language. This shift is particularly pertinent in addressing the intricacies of medical terminology present in nursing care records, demonstrating a commitment to enhancing the model's relevance and effectiveness within the healthcare domain.

Additionally, the research places emphasis on the significance of evaluation metrics in bridging the gap between subjective human assessments and objective quantitative measures. The proposal of EmbedHDP as an evaluation metric seeks to mitigate the challenges posed by the limited number of caregivers and the inherent subjectivity in their evaluations. This approach offers a standardized and quantitative measure, reducing dependency on individual interpretations and providing a more reliable means of assessing the quality of sentence suggestions.

How durable is EmbedHDP in terms of its long-term effectiveness and stability?

The long-term durability and stability of EmbedHDP is an essential consideration in assessing its effectiveness as a metric for evaluating sentence suggestion in nursing documentation. While EmbedHDP introduces innovative approaches, such as incorporating word embeddings to address the intricacies of medical terminology, its robustness in dealing with sentence nuances, especially those beyond 14 words, presents a significant challenge. The inherent limitation of EmbedHDP to longer sentences warrants careful consideration,

as sentences in nursing records can often extend beyond this threshold.

The research acknowledges the challenge posed by longer sentences, emphasizing that they tend to contain more complex information and overlap, making their analysis more complicated. This limitation points to the need for future analysis and enhancements to ensure EmbedHDP's adaptability and effectiveness in handling extended sentences commonly found in healthcare documentation. The long-term sustainability of EmbedHDP depends on its ability to evolve and address such challenges to ensure its continued relevance and stability in the dynamic landscape of nursing care record applications.

As EmbedHDP undergoes further refinement and adaptation to accommodate longer records, the research community's commitment to ongoing development and improvement will be critical to its continued effectiveness. The durability of EmbedHDP, in the context of its long-term application, will depend on its ability to overcome limitations and evolve with the evolving needs and complexities of nursing care records. Therefore, a continued commitment to research and development, addressing identified limitations, and adapting to emerging challenges will be critical to ensuring the continued effectiveness and stability of EmbedHDP in the nursing care record application domain.

What aspects are currently under evaluation in the ongoing assessment?

It is important to note that certain limitations have been identified during this evaluation process. One significant limitation revolves around the handling of longer sentences, specifically those longer than 14 words. The analysis shows that EmbedHDP faces challenges in effectively capturing the intricacies of longer sentences, which may impact its ability to provide accurate and nuanced evaluations for such cases. This limitation prompts further investigation and refinement to improve EmbedHDP's adaptability to extended sentences commonly found in healthcare documentation.

In addition, another critical condition to evaluate concerns situations where one sentence is a subset of another sentence. This scenario introduces additional complexity, as the model must navigate the hierarchical relationships between sentences and accurately evaluate their similarities or differences. Evaluating the model's performance in handling subsets and ensuring that it can distinguish nuanced differences in such cases becomes a critical aspect of ongoing evaluation and refinement.

In essence, the evaluation process has examined EmbedHDP across multiple dimensions, highlighting its strengths and identifying areas for further attention and development. As the research progresses, addressing the identified limitations and refining the model's

capabilities will be critical to achieving a more comprehensive and effective metric for evaluating sentence suggestions in nursing care records.

What are the findings derived from the proposed design?

The findings derived from the proposed design reveal a series of strategic modifications aimed at enhancing the overall performance and effectiveness of the evaluation metric, EmbedHDP. Notably, a pivotal adjustment involved substituting the ineffective Bag-of-Words (BoW) approach with word embeddings. This transition proved essential in improving the model's capacity to comprehend the semantic information embedded within sentences, marking a significant advancement in the understanding of the intricacies of nursing care records.

In addition, careful consideration was given to the optimization of hyperparameters tuned to the corpus generated by word embeddings. The need for compatibility between the corpus and the hyperparameterized Hierarchical Dirichlet Process (HDP) training process required careful adaptation to ensure seamless integration. Despite the advantages introduced by the word embeddings, a transformation back to the BoW format was necessary after the corpus acquisition, highlighting the complex interplay between these components in the design.

Furthermore, the refinement extended to the composition of the dictionary, which initially included a compilation of corpora of sentence suggestions and their corresponding ground truth sentences. This dictionary underwent a transformative shift to encapsulate a collection of words specifically representing medical terminology prevalent in nursing records. This customized dictionary improved the model's ability to recognize the nuances of healthcare language and addressed the challenge posed by medical terminology, contributing to more nuanced and accurate scoring.

Essentially, these results underscore the careful adjustments made to the design, emphasizing a paradigm shift from conventional approaches to a more sophisticated and domain-specific methodology. The amalgamation of word embeddings, hyper-parameter optimization, and a specialized dictionary has culminated in a more robust and context-aware EmbedHDP poised to improve sentence suggestion evaluation in nursing care records.

Is the dataset designed to be universal, or does it possess limitations in its applicability?

The dataset used in this study comes from a collection of data collected through the use of the care record application known as FonLog. It's imperative to note that this dataset is inherently closed, as it is associated with sensitive information about elderly individuals

within healthcare facilities. Unlike universal datasets, the characteristics of this dataset are specifically tailored to include information relevant to the elderly population.

In addition, the nature of the dataset is unique and not generally applicable to general contexts. It delves into the intricacies of data related to the elderly, providing a nuanced perspective that includes elements such as medical terminology commonly found in care records. To understand the nuances of the sentences in this dataset, a prior understanding of care records is essential, emphasizing the specialized and domain-specific nature of the information contained in the dataset. As a result, the design of the dataset is not intended for universal applicability but rather for the unique and specialized domain of elderly care records within healthcare facilities.

What specific aspects require further attention or improvement based on the findings obtained?

Based on the results obtained, certain aspects that require further attention and potential improvement emerge. One critical area that requires careful consideration is the limited number of nurses involved in the assessment process. Relying on only three caregivers may present challenges in establishing a robust benchmark due to potential variations in individual ratings. The diversity of opinion among caregivers, as evidenced by the varying scores for a given example sentence, underscores the need for a larger and more diverse pool of caregivers to increase the reliability and representativeness of the benchmark.

In addition, current evaluation metrics, including EmbedHDP, have limitations when dealing with longer sentences, particularly those containing 14 words or more. The inherent complexity and overlap of information in longer sentences present a challenge to effective analysis. Addressing this limitation will be critical to ensuring the long-term effectiveness and stability of EmbedHDP, especially in scenarios where large and detailed sentences are prevalent in care records.

Furthermore, the transition from Bag-of-Words (BoW) to word embeddings in the corpus generation process is a notable improvement. However, there is room for refinement in tuning the hyperparameters to match better the corpus characteristics resulting from the word embeddings. This adaptation is crucial for optimizing the training of the Hierarchical Dirichlet Process (HDP) on the generated corpus. The inclusion of medical terminology in the dictionary is commendable but could benefit from ongoing curation to improve its representation and coverage of relevant terms.

In summary, future efforts should focus on expanding the pool of caregivers for evalu-

ation, addressing limitations related to sentence length, fine-tuning hyperparameters for optimal compatibility with word embeddings, and refining the medical terminology dictionary. These targeted improvements will help increase the robustness and effectiveness of the proposed EmbedHDP evaluation metrics in the context of care record applications.

What insights or conclusions can be drawn from the initial exploration of these aspects?

The initial exploration of various aspects in the study yields valuable insights and conclusions. Firstly, the limited number of caregivers participating in the evaluation process highlights the need for a more extensive and diverse pool to establish a comprehensive benchmark. The observed variations in evaluation scores among three caregivers emphasize the subjectivity inherent in human assessments, necessitating a broader range of perspectives for a more robust benchmark.

Regarding the limitations associated with sentence length, particularly those exceeding 14 words, the findings underscore the challenges in effectively capturing the nuances of longer sentences. This insight suggests that a tailored approach may be necessary for the sentence suggestion generation system to effectively handle more extensive and complex information in nursing care records.

The transition from Bag-of-Words (BoW) to word embeddings in the corpus generation process reveals promising improvements in understanding the semantic information within sentences. However, the initial exploration suggests a need for further fine-tuning of hyperparameters to optimize compatibility with the characteristics of the generated corpus. This refinement aims to enhance the training effectiveness of the Hierarchical Dirichlet Process (HDP) on the word embeddings corpus, contributing to better overall performance.

The incorporation of medical terminology into the dictionary signifies a positive step toward contextualizing the language models for nursing care records. The initial exploration suggests that ongoing curation and expansion of the medical terminology dictionary may be beneficial for a more comprehensive representation of relevant terms.

In conclusion, the initial exploration highlights the importance of diversifying caregiver evaluations, addressing challenges associated with sentence length, fine-tuning hyperparameters for optimal training, and continually refining the medical terminology dictionary. These insights pave the way for further research and refinement to enhance the proposed sentence suggestion generation model's effectiveness and applicability in nursing care record applications.

6.2 Future Works

Future works will leverage established models such as BERTScore and fine-tune their contextual embeddings to align with the specific characteristics of care record sentences. The comparison with EmbedHDP will provide insights into the strengths and weaknesses of each model. Additionally, efforts will be directed towards refining both evaluation models by gaining a deeper understanding of their mechanisms. This iterative process of refinement and comparison contributes to the continuous improvement and adaptation of evaluation techniques for sentence suggestions in the care records domain.

Future research can also explore optimising the use of Large Language Models (LLMs) as evaluation metrics, especially in the context of sentence suggestions in nursing care record applications. The main question that needs to be answered is how we can more effectively utilize LLM capabilities to assess and improve the quality of sentence suggestions in nursing care note applications. Previously, research has shown that LLM has been used to provide helpful feedback on research papers and compared it to human feedback from peer reviewers[29]. The results of this study show that the overlap between input from GPT-4 and human input is comparable to the overlap between two human reviewers and that GPT-4 tends to identify key or common issues raised by multiple reviewers. Additionally, the study notes that GPT-4 emphasizes some aspects of feedback more than humans, such as suggesting more experiments or data sets and providing more positive feedback. These findings may provide a foundation for future research to identify ways LLM can be specifically optimized to evaluate better sentence suggestions in implementing nursing care notes.

Chapter 7

Conclusion

This study aims to provide evaluation metrics to better capture semantic information in sentence suggestions in care notes and produce ratings that reflect human evaluations. Because in providing a sentence suggestion system for elderly care records, the accuracy of evaluation metrics is critical. As the resulting sentence suggestions are closely related to important information about the elderly, and any inaccuracies may pose potential risks.

Current evaluation metrics fail to fully reflect their ability to analyze the quality of suggested sentences, which is critical for their implementation in nursing care recording systems. For instance, BERTScore has difficulty evaluating the domain of nursing care notes effectively and consistently assessing the quality of the resulting sentence suggestions above 60%. Additionally, cosine similarity, which is widely used, represents limitations in word order, leading to potential misjudgments of semantic differences in sets of similar words. Then, researchers also found that ROUGE relied on lexical overlap but tended to ignore semantic accuracy. Another point is BLEU ignores semantic coherence in its evaluation.

We proposed a new evaluation metric, EmbedHDP, with a word embedding approach combined with HDP. After conducting experiments on 320 original data from elderly facilities, EmbedHDP outperformed other evaluation metrics with a coefficient score of 61%. The cosine similarity has a coefficient score of 59% and a BERTScore of 58%. However, it must be acknowledged that EmbedHDP still has limitations and requires careful analysis for future development. One significant challenge lies in EmbedHDP's vulnerability when analyzing similarities between two care record sentences, especially when the sentences are relatively long, containing 14 words or more. EmbedHDP has difficulty capturing sentence relationships effectively in such scenarios.

Another potential approach for enhancing the evaluation of sentence suggestions in care record applications involves fine-tuning existing metrics, particularly focusing on well-established evaluation models like BERTScore. This strategic approach acknowledges the necessity for domain-specific evaluations that can accurately capture the intricacies of language within care records. While fine-tuning process requires a significant time investment, especially in tasks such as generating contextual embeddings specific to the care record domain, the potential benefits are substantial. The resulting fine-tuned metrics have the potential to provide a more nuanced and precise assessment of the generated content, contributing to the continual refinement of natural language processing techniques tailored for the intricacies of healthcare-related texts. In care records, where language is highly specialized and context-dependent, adapting evaluation metrics like BERTScore through fine-tuning is a strategic move.

Acknowledgement

First of all, praise Allah to ease, protect, and help me in all circumstances.

In this section, I would like to thank my supervisor, Prof. Sozo Inoue Sensei, for his constant support, kindness, and guidance. I always enjoyed the time spent having discussions and brainstorming sessions with him. He always engaged and motivated me to think more and go outside the box while also making me excited about the research. His knowledge and understanding of the topic area aided me throughout this research. I would also like to thank my co-supervisors, Prof. Keichii Horio Sensei, Prof. Kaori Yoshida Sensei and Prof. Kazukata Shimada Sensei, for their guidance and support. I also say Thank you to Prof. Muhammad Atiqur Rahman Ahad Sensei, his support and insight also helps me able to go through this hard time.

Furthermore, I would also like to thank our lab secretary, Miki Uchida San, for her kindness in helping me handle organizational work for all things related to my research. As Miki Uchida San family always greatly helps my small family, I also owe her so much. I am also thankful to my lab members for their cheer and support during my time as a lab member. Here, I would like to convey special thanks to Haru Kaneko, John Noel Victorino, and Christina Garcia for actively helping me with my research work.

Moreover, I would like to express gratefulness and gratitude to my beloved wife for always supporting me with her delicious food that prevented me from starving during my working time. Also, for my son, thanks a lot for being a nice and kind baby who always cheers me up when I come back from campus.

Lastly, I thank my family, especially my parents. I know I can handle all the struggle because of their du'a as a support from afar. Without their constant support, I could not complete this tough journey.

Publications

Journal

- **Hamdhana, D.**, Kaneko, H., Victorino, JN., and Inoue, S. "Improved Evaluation Metrics for Sentence Suggestion in Nursing and Elderly Care Record Applications". In: *Healthcare*(2024). https://doi.org/10.3390/healthcare12030367

International Proceedings

- **Defry Hamdhana**. 2020. Mobile application for caregiver in collecting statistical data of BPSD attack focused on macro activities: PhD forum abstract. In Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys '20). Association for Computing Machinery, New York, NY, USA, 807–808. DOI: https://doi.org/10.1145/3384419.3430577
- Fikry M., **Hamdhana D.**, Lago P., Inoue S. (2021) Activity Recognition for Assisting People with Dementia. In: Ahad M.A.R., Mahbub U., Rahman T. (eds) Contactless Human Activity Analysis. Intelligent Systems Reference Library, vol 200. Springer, Cham. https://doi.org/10.1007/978-3-030-68590-4_10
- **Hamdhana, D.**, Garcia, C., Nahid, N., Kaneko, H., Alia, S. S., Hossain, T. and Inoue, S.: Summary of the Fourth Nurse Care Activity Recognition Challenge - Predicting Future Activities, Human Activity and Behaviour Analysis: Advances in Computer Vision and Sensors, CRC press (2022)
- Kaneko, H., Victorino, JN., Garcia, C., **Hamdhana, D.**, Fikry, M, Nahid, N., Hossain, T. Shibata, T. and Inoue, S. Summary of the Fifth Activity and Behavior Computing Challenge – Forecasting Wearing-Off Phenomenon. 5th International Conference on Activity and Behavior Computing (September 7th-9th, 2023)

Presentations and Local Publications

- **Hamdhana, D.**, Kaneko, H., Victorino, JN., and Inoue, S.: Toward Automatic Sentence Suggestion in Care Record Mobile Applications, SAES (2022) Abstract submitted, paper submission Dec. 2022
- **Hamdhana, D.**, Garcia, C., Nahid, N., Kaneko, H., Alia, S. S., Hossain, T. and Inoue, S.: Brief Summary of the 4th Nurse Care Activity Recognition Challenge, IPSJ-SIG ASD (2022). Paper submitted, Presentation December 2022

Appendices

This appendix presents a compilation of 390 pairs of sentence suggestions that have been meticulously paired with their respective ground truths. To measure the quality of these sentence suggestions, an expert score was derived as the average of the ratings provided by three caregivers. These caregivers rated the sentences using a questionnaire that gave them the option of rating the suggestions as "excellent," "good," "fair," "poor," or "bad. The expert score serves as a benchmark. The evaluation metrics are calculated for their proximity to the expert score using the correlation coefficient score.

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
1	頭頭痛の訴えある	頭痛の訴えあり	0.92	0.67	0.96	0.41	0.40	1.00
2	頭痛頭痛の訴えあり。	頭痛の訴えあり。	0.92	0.89	0.94	0.95	0.82	0.91
3	頭痛訴えありああナース報告入れる	頭痛訴えありああナース報告入れる	1.00	1.00	1.00	1.00	1.00	1.00
4	一服用して頂くください	一包服用して頂く。	0.50	0.80	0.80	0.67	0.70	0.52
5	起きないそう。	起き上がれないそう。	0.25	0.50	0.81	0.71	0.68	1.00
6	ああ往診変わりなし	ああ医院往診変わりなし	0.92	0.89	0.90	0.87	0.76	0.93
7	泡沫の唾液あり。	泡沫状の唾液あり。	1.00	0.89	0.94	0.82	0.78	1.00
8	悪寒消失まだ解熱はないポカリスエット 100cc 召し上がって頂くください	悪寒が治まった 熱がまだ下がらない ポカリスエットを 100ml 飲む。	0.50	0.37	0.75	0.48	0.28	0.67
9	ああ電話とられ状況を報告を行う誤嚥があったようでゴロ音があり顔色不良のため再建する	ああナース電話とられ状況を報告する。	0.33	0.48	0.83	0.54	0.34	0.56
10	上腕再検。	上腕で再検。	1.00	0.80	0.91	1.00	0.65	1.00
11	上腕再検。	上腕で再検。	1.00	0.80	0.91	1.00	0.65	1.00
12	吐き気あり報告入れる	吐き気訴えあり	0.92	0.57	0.85	0.41	0.20	0.86
13	クーリング中	クーリング継続	1.00	0.67	0.86	0.50	0.40	1.00

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
14	エリア全体が 5 〜 10 秒ごとにパルスを発します。	患者の脈拍は 5 〜 10 秒ごとです。	0.42	0.24	0.77	0.58	0.38	0.53
15	昼食ホールに誘導するが腰痛と食欲不振を訴え臥床を希望される頭部挙上にて様子観察する	患者は昼食のためにホールに案内されたが、腰痛と食欲不振を訴えた。	0.75	0.52	0.79	0.59	0.45	0.80
16	入居 88%spo2 測定不可。	入居時 88%spo2 のため再建する	0.17	0.62	0.80	0.67	0.41	0.23
17	気分訴えなし	気分不良訴えなし	1.00	0.86	0.91	0.82	0.70	1.00
18	気分訴えなし	気分不良はないと本人言われる	1.00	0.18	0.70	0.35	0.07	0.94
19	風邪た熱がある起き上がるとめまいがあり顔色不良のため再建する	風邪引いた熱があると訴えあり。	0.42	0.58	0.77	0.35	0.34	0.56
20	昼食て来られずきついと訴えられる。	昼食起きて来られずきついと訴えある	0.75	0.84	0.85	0.89	0.65	0.87
21	昼食汁物飲まれ嘔吐される頭部挙上にて様子観察する	昼食時汁物飲まれ嘔吐される	0.40	0.70	0.90	0.63	0.53	0.88
22	ああ医院往診変わりなし	ああ医院往診される。	0.92	0.60	0.79	0.87	0.47	1.00
23	お腹痛い訴え有り。	お腹が痛いの訴えあり。	1.00	0.60	0.87	0.87	0.55	0.82
24	胃痛い訴え有り。	胃が痛いか分からないと訴えられる。	0.33	0.46	0.75	0.29	0.09	0.34

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
25	ああ来てくださり対応して頂くください	ああナース来てくださり対応してください。	0.92	0.90	0.86	0.80	0.66	0.99
26	体測定行う。	体バイタル測定行う。	0.50	0.80	0.88	0.87	0.72	0.69
27	また o2 開始する。頭部ギャッジアップ行う	また１ o2 開始する。30 分後にて 93 から 95 %に上昇する	0.17	0.48	0.70	0.43	0.32	0.34
28	また口周りに乾燥して頂くください	また、口周りに乾燥した食渣張り付きあり。	0.25	0.56	0.77	0.51	0.43	0.88
29	さむけすると訴えられる。	さむけがすると言われる。	0.83	0.62	0.89	0.50	0.52	0.82
30	介助ことに対して不穏みられる。	介助することに対して不穏みられる。	0.92	0.91	0.89	1.00	0.84	0.90
31	居室過ごされる	居室にて良眠される	0.75	0.44	0.81	0.41	0.23	1.00
32	居室過ごされる	居室にて休まれる	1.00	0.33	0.84	0.50	0.51	1.00
33	居室過ごされる	居室にて過ごされる	0.92	0.86	0.90	1.00	0.73	1.00
34	フロア過ごされる	フロアにて傾眠される	0.75	0.50	0.82	0.50	0.32	0.68
35	朝右口角下がり、言葉が少し聞きづらい等から水分補給で様子を見ていない	朝から右口角下がり、言葉が少し聞きづらい等から水分補給で様子を見ていたが午後より呂律難あり。	0.33	0.79	0.89	0.90	0.66	0.83
36	経管	経管栄養	0.50	0.80	0.85	0.71	0.39	1.00

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
37	ああ来館。可動域訓練実施	ああ PT 来館。可動域訓練実施	0.75	0.92	0.93	0.85	0.75	0.73
38	排尿居室にて過ごされる	排尿あり、パッド交換居室にて良眠される	0.25	0.53	0.77	0.47	0.23	0.81
39	四肢する	四肢施行する	0.75	0.80	0.88	0.71	0.59	1.00
40	水分使用あり	水分トロミ使用あり	0.25	0.75	0.83	0.82	0.66	0.87
41	ズボンクッション交換	ズボンとクッション交換	1.00	0.86	0.89	1.00	0.70	1.00
42	排尿居室にて過ごされる	排尿あり、パッド交換居室にて良眠される	0.42	0.53	0.77	0.47	0.23	0.81
43	受診ため取り置き対応中	受診のため取り置き対応中	0.83	0.86	0.92	1.00	0.81	0.82
44	フルーツレクトロミ使用あり	フルーツレク水分トロミ使用あり	0.50	0.44	0.87	0.35	0.30	0.52
45	中止　大建中湯 1 包	中止薬剤　大建中湯 1 包	0.67	0.92	0.87	0.93	0.80	0.55
46	左足腱部表皮剥離大きい	左足アキレス腱部表皮剥離大きい	0.75	0.80	0.92	0.93	0.82	0.76
47	ああ来館。可動域訓練実施	ああ PT 来館。可動域訓練実施	0.75	0.92	0.93	0.85	0.75	0.73
48	部屋男の人が数人来るんですとやや不穏みられるフロアにて過ごされる	部屋に男の人が数人来るんですとやや不穏みられるフロアにて過ごされる	0.83	0.97	0.97	1.00	0.95	0.79
49	ベッドセンサーコールにて訪室。トイレに座っている水疱は破れておらず。その後ワセリンを塗るられる。	ベッド下センサーコールにて訪室。トイレに座っている。終わったあと三階にお連れする。	0.75	0.40	0.83	0.52	0.47	0.75

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
50	ガスする	ガス抜き	0.25	0.50	0.86	0.71	0.34	0.84
51	左一趾ゲーベンクリーム足しメロリンガーゼあてておく。	左第一趾ゲーベンクリーム塗布	0.50	0.40	0.88	0.58	0.40	0.58
52	今チクチクもしません、と。	今はチクチクもしません、と。	0.83	0.91	0.92	1.00	0.88	0.84
53	他臀部のデュオアクティブドレッシングは汚れ等なく正常に貼られている水疱は破れておらず。その後ワセリンを塗るられる。	他､臀部のデュオアクティブドレッシングは汚れ等なく正常に貼られている。様子見。	0.25	0.69	0.89	0.73	0.56	0.93
54	排尿あり、パッド交換居室にて休まれる	排尿あり、パッド交換居室にて休まれる	1.00	1.00	1.00	1.00	1.00	1.00
55	おむつ介助行う居室にて過ごされる	おむつ更衣介助行う。	0.50	0.55	0.83	0.67	0.34	0.59
56	左部内側、熱傷あり。	左大腿部内側、熱傷あり。	0.67	0.86	0.93	0.26	0.84	0.89
57	左一趾ゲーベンクリーム足しメロリンガーゼあてておく。	左第一趾ゲーベンクリーム塗布	0.58	0.40	0.88	0.58	0.40	0.58
58	ポステリザン注入する本日脱肛増大	ポステリザン軟膏注入する本日脱肛増大	0.50	0.92	0.96	0.91	0.85	0.51
59	眠前内服時　居室にて過ごされる	眠前薬内服時　居室にてスタッフとお話される	0.75	0.63	0.77	0.72	0.53	1.00
60	熱発為中止	熱発の為、ナースに報告し中止。	0.50	0.55	0.67	0.77	0.11	0.77

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
61	居室は、よく眠るための居心地の良い場所です。	居室にて良眠される	0.50	0.13	0.75	0.52	0.10	0.68
62	居室で良い休息を。	居室にて休まれる	0.50	0.25	0.86	0.41	0.14	1.00
63	居室で過ごす時間は、リラックスして過ごすのに最適です。	居室にて過ごされる	0.33	0.13	0.85	0.75	0.09	0.48
64	フロアにて傾眠されると	フロアにて傾眠される	0.92	0.91	0.97	1.00	0.84	1.00
65	朝から口の端が下がり、言葉が少し聞きづらかったため、水分補給で様子を見ていましたが、午後からは呂律が回りにくくなってしまいました。	朝から右口角下がり、言葉が少し聞きづらい等から水分補給で様子を見ていたが午後より呂律難あり。	0.67	0.63	0.89	0.65	0.49	0.77
66	あああ PT さんが来館され、可動域訓練を実施してくださいました。	あああ PT 来館。可動域訓練実施	1.00	0.67	0.85	0.88	0.31	0.81
67	排尿があり、パッド交換を行った後、患者様は居室で心地よくお眠りになっています。	排尿あり、パッド交換居室にて良眠される	0.58	0.33	0.85	0.52	0.22	0.71
68	水分摂取に際して、水分トロミを使用するようにしてください。	水分トロミ使用あり	0.50	0.44	0.82	0.77	0.16	0.67
69	患者様のズボンとクッションを取り替えます。	ズボンとクッション交換	0.50	0.50	0.80	0.77	0.29	0.28

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
70	患者受診のため、取り置き中。	受診のため取り置き対応中	0.58	0.75	0.86	0.80	0.62	0.61
71	フルーツレク飲用中、水分トロミ使用。	フルーツレク水分トロミ使用あり	0.75	0.83	0.89	0.82	0.58	0.80
72	薬剤中止、大建中湯 1 包で対応。	中止薬剤　大建中湯 1 包	0.75	0.88	0.85	0.86	0.62	0.73
73	左足アキレス腱部に大きな表皮剥離があります。	左足アキレス腱部表皮剥離大きい	0.75	0.57	0.89	0.86	0.48	0.86
74	PT 来館、可動域訓練実施。	ああ PT 来館。可動域訓練実施	0.75	0.92	0.83	0.93	0.80	0.58
75	数人の男性が部屋に来たため、やや不安に感じながらフロアで過ごしています。	部屋に男の人が数人来るんです…とやや不穏みられるフロアにて過ごされる	0.75	0.36	0.78	0.73	0.32	0.74
76	ベッド下のセンサーアラートで訪問しました。トイレで用を足した後、3 階に案内しました。	ベッド下センサーコールにて訪室。トイレに座っている。終わったあと三階にお連れする。	0.67	0.29	0.77	0.36	0.32	0.64
77	左足の第一趾にゲーベンクリームを塗布しました。	左第一趾ゲーベンクリーム塗布	0.75	0.50	0.89	0.77	0.39	0.77
78	今はチクチク感がないようです。	今はチクチクもしません、と。	0.58	0.36	0.81	0.67	0.34	0.41
79	他にも臀部のデュオアクティブドレッシングも正常に貼られており、汚れもないことから、今後の経過観察となります。	他､臀部のデュオアクティブドレッシングは汚れ等なく正常に貼られている。様子見。	0.75	0.51	0.89	0.64	0.41	0.78

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
80	排尿があり、休息のため居室でパッド交換が行われました。	排尿あり、パッド交換居室にて休まれる	0.58	0.50	0.86	0.68	0.33	0.58
81	おむつを交換するため、介助を行いました。	おむつ更衣介助行う。	0.58	0.33	0.77	0.67	0.20	0.65
82	左太ももの内側にやけどの跡があります。	左大腿部内側、熱傷あり。	0.75	0.29	0.76	0.20	0.12	0.85
83	左足第一趾にゲーベンクリームを塗りました。	左第一趾ゲーベンクリーム塗布	0.67	0.40	0.89	0.62	0.40	0.67
84	ポステリザン軟膏を注入した結果、脱肛が改善されました。	ポステリザン軟膏注入する本日脱肛増大	0.67	0.40	0.86	0.62	0.24	0.69
85	患者様は、眠前薬を内服された後、居室でスタッフとお話されました。	眠前薬内服時　居室にてスタッフとお話される	0.75	0.59	0.87	0.67	0.35	0.83
86	色塗りに取り組んで過ごされています。	色塗りに取り組まれるフロアにて過ごされる	0.67	0.42	0.80	0.89	0.35	0.84
87	100ml の尿量が測定されました。	尿量測定 100ml	0.67	0.67	0.84	0.89	0.30	0.57
88	熱があったので、看護師に報告して中止しました。	熱発の為、ナースに報告し中止。	0.75	0.36	0.83	0.37	0.27	0.71
89	リスペリドンの服薬介助を行いました。	リスペリドン服薬介助す	1.00	0.55	0.86	0.77	0.29	0.98

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
90	食事中に口周りに食べカスがつき、乾燥しています。	また、口周りに乾燥した食渣張り付きあり。	0.58	0.35	0.77	0.43	0.36	0.73
91	眠前薬の代わりに、お湯 50ml を飲むことをおすすめします。	眠前薬 + 白湯 50ml フラッシュする。	0.42	0.38	0.80	0.53	0.24	0.48
92	居室にて適切な環境を整え、質の高い睡眠を実現します。	居室にて良眠される夜間 2 時間ごとの巡視行う。	0.00	0.23	0.73	0.12	0.18	0.68
93	ラコールを投与した後、食事の後に薬を注入し、最後に温かいお湯で洗い流しました。	ラコール 1P 注入 + 食後薬注入 + 白湯フラッシュ施行し終了。	0.75	0.26	0.72	0.25	0.14	0.74
94	肺音聴取すると左肺野で荒々しい呼吸音が聞こえました。	肺音聴取すると右肺野よりフツフツと水泡音あり。	0.50	0.52	0.82	0.64	0.44	0.89
95	夕食終了後、サプリメントを 1 錠追加で摂取しています。	夕食後薬センノシド 1T 追加内服される	1.00	0.27	0.72	0.50	0.23	0.90
96	居室にて趣味に没頭する	居室にて良眠される	0.25	0.33	0.85	0.33	0.26	0.89
97	特殊な方法で入眠するために、時間ごとにマットレスの交換を行う	特変なく入眠される時間ごとのパット交換行う	0.25	0.42	0.81	0.57	0.33	0.90
98	特殊な方法で入眠される	特変なく入眠される	0.67	0.57	0.89	0.33	0.47	0.80
99	水分を抜いたほうが良いです	水分トロミ使用あり	0.25	0.17	0.68	0.29	0.10	0.57
100	一人で泳ぐのは避けてください。	一人で歩かないお願いす。	0.30	0.17	0.81	0.29	0.25	0.80

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
101	職員二人が協力してベッドの交換と体位変更を行う	職員二人介助にてパット交換体位交換する	0.80	0.29	0.82	0.59	0.29	0.86
102	ゲンタシンを塗ると、傷が速く治ります。	ゲンタシン擦り込み、カットバン保護する。	0.42	0.13	0.77	0.20	0.14	0.42
103	右第 3 趾爪、炎症が起こっている。	右第 3 趾爪、内出血している。	0.50	0.62	0.88	0.71	0.69	0.83
104	入浴中に驚きました。	入浴に行かれてる。	0.25	0.44	0.78	0.29	0.29	0.72
105	ああ歯科の治療を避けられる。	ああ歯科の治療を受けられる	0.25	0.86	0.94	0.75	0.79	0.86
106	排便量が少ないです。	排便中量あり。	0.50	0.50	0.83	0.41	0.21	0.51
107	端座位になろうとしたが、時間を伝えて立ち上がるようにお願いします。	端座位になろうとされている為時間お伝えし臥床して頂く。	0.50	0.53	0.79	0.62	0.48	0.82
108	廊下でランニングする。	廊下にてリハビリされる。	0.50	0.22	0.81	0.50	0.27	0.81
109	右前腕にはバンドが巻かれているので、状況を確認する必要があります。	右前腕剥離部テープついているので、様子見る。	0.67	0.30	0.79	0.29	0.32	0.91
110	バルーンに液体 150ml 注入済み	バルーンに排尿 150ml あり	0.67	0.62	0.89	0.61	0.55	0.64
111	独語あり静かに話される	独語あり大声を出される	0.50	0.50	0.85	0.33	0.35	0.86
112	車椅子にてリハビリテーションを行われる	車椅子にて居室にて良眠される	0.25	0.43	0.81	0.45	0.39	0.67

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
113	車椅子か使わないで歩こうと頑張っている。	車椅子から立ち上がろうと何度もされる。	0.42	0.33	0.76	0.37	0.24	0.49
114	居室内にて靴が散乱している	居室内にてお茶っ葉を床にこぼされる	0.25	0.22	0.81	0.45	0.28	0.88
115	排尿あり、便秘の兆候見受けられ、適切な下剤の処方を検討する必要があります。	排尿あり、泥状便少量あり、清拭する夜間 2 時間ごとの巡視行う。	0.42	0.29	0.72	0.10	0.31	0.69
116	利尿剤を処方することを検討する	利尿剤中止となる	0.00	0.31	0.84	0.58	0.20	0.77
117	おやつ後、他の入居者様と楽しく会話をされるために、共有スペースに移動されます	おやつ後他入居者様に、ご立腹され居室にて過ごされる	0.50	0.43	0.80	0.60	0.40	1.00
118	ポータブルトイレで快便体験	ポータブルトイレで排尿あり	0.67	0.55	0.86	0.58	0.58	0.86
119	眠前薬を服用すると、心地よい眠りに誘われます。	眠前薬センノシド 1 錠内服拒否される	0.50	0.19	0.70	0.31	0.19	0.67
120	立位が難しいので、トイレでの排尿はお手伝いいたします。	立位できない、パッド排尿お願いする	0.50	0.29	0.77	0.41	0.16	0.81
121	居室にて趣味に没頭される	居室にて良眠される	0.50	0.62	0.87	0.33	0.39	0.68
122	おやつの時間にお茶を飲むと、全部吐き出してしまうそうです。	おやつの時のお茶を飲むときむせてお茶のみ嘔吐される	0.83	0.50	0.79	0.50	0.36	0.92
123	バスルームでおしっこする	浴室にて浣腸行う	0.83	0.00	0.75	0.00	0.00	0.80

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
124	スタッフと共にケアプランを確認しながら時間を過ごされる	スタッフとお話されるフロアにて過ごされる	0.75	0.30	0.82	0.41	0.31	0.77
125	入浴後、全身にパウダーをふりかけます。	入浴後体全体に保湿剤塗布する。	0.67	0.27	0.78	0.34	0.32	0.65
126	左下腿の皮下出血部から腫れが生じています。	左下腿の皮下出血部より出血あり。	0.67	0.47	0.92	0.80	0.44	0.73
127	私は 2 時間後に夕食を食べましたが、その後は眠れなくなりました。	夕食 2 時間以上かけ召し上がり、その後臥床介助。	0.50	0.11	0.70	0.47	0.20	0.63
128	頻繁な少量の排尿。	排便中量あり。	0.33	0.00	0.76	0.00	0.08	0.45
129	居室内にて、椅子がひっくり返る	居室内にてお茶っ葉を床にこぼされる	0.25	0.27	0.82	0.45	0.25	0.68
130	部屋の照明を調整してリラックスしましょう	居室にて良眠される時間ごとのパット交換行う	0.00	0.10	0.69	0.00	0.06	0.84
131	リハビリ帰館される	リハビリ帰館される	1.00	1.00	1.00	1.00	1.00	1.00
132	ラコール注入 + 食後薬＋白湯 50ml 注入開始する	ラコール 1P 注入＋食後薬注入＋白湯フラッシュ施行し終了	0.50	0.55	0.85	0.75	0.50	0.45
133	白湯 +nacl4g 注入し終了する	白湯 500ml 滴下後、昼食後薬注入し、50ml フラッシュし終了する	0.42	0.43	0.75	0.29	0.25	0.42
134	ミックス塗布防水テープ保護する	ミックス軟膏、両ソケイ部に塗布	0.25	0.33	0.74	0.37	0.22	0.69

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
135	発赤改善して終了する	発赤は変わらずあり	0.00	0.18	0.83	0.41	0.13	0.56
136	敬老日レク、メッセージカード	敬老の日、ケーキを提供する	0.50	0.20	0.76	0.45	0.14	0.51
137	発赤にベタメタゾン塗布	発赤部分にベタメタゾン塗布	0.67	0.89	0.93	0.87	0.72	0.51
138	ハロウィンお菓子を召し上がる	ハロウィンでお菓子を召し上がる	0.92	0.89	0.97	1.00	0.81	1.00
139	今朝時ひだり腕痛いと訴える左肩甲骨から腕にかけ痛みは腿の筋肉を鍛える	今朝離床時ひだり腕痛いと訴える左肩甲骨から腕にかけ痛みと運動制限がある	0.83	0.65	0.89	0.82	0.65	0.88
140	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後体全体に保湿剤塗布してもらっている	0.58	0.29	0.78	0.25	0.21	0.51
141	発赤や掻き傷などは出来ておらず	発赤疹や掻き傷などは出来ておらず	0.83	0.95	0.96	0.87	0.89	0.81
142	人工知能は未来の医療に革命をもたらす可能性がある。	人工透析の為、アバ病院へ送迎を行う	0.50	0.26	0.72	0.13	0.16	0.68
143	レントゲン行う	レントゲン撮影あり	0.67	0.40	0.85	0.50	0.27	0.77
144	白湯 +nacl4g 注入し終了する	白湯 400ml+NaCl4 ｇ注入しフラッシュして終了する	0.25	0.67	0.83	0.84	0.42	0.26
145	訪室と転倒した方が良いとの連絡あり	訪室すると　転倒したと言われる	0.75	0.60	0.78	0.45	0.41	0.78
146	左手甲青い小指の方が良いとの連絡あり	左手の甲青い小指の方	0.42	0.50	0.86	0.85	0.53	0.72

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
147	眠前飲むかどうかわからないので飲んで心臓の状態が良いとの連絡あり	眠前薬内服	0.25	0.09	0.74	0.18	0.04	0.34
148	体の発疹あるところに保湿剤塗布する経過は良好との連絡あり	体全体の発疹あるところにウレパールクリーム塗布	0.42	0.43	0.82	0.43	0.36	0.39
149	指示軟膏塗布防水テープ保護する	指示の軟膏塗布する	0.42	0.67	0.86	0.71	0.54	0.53
150	足趾乾燥して終了する	足趾間乾燥している	0.50	0.62	0.90	0.67	0.61	0.54
151	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後両下肢に保湿剤塗布し着圧ソックス装着する	0.50	0.35	0.79	0.22	0.27	0.62
152	腹部部に軟膏塗布防水テープ保護する	腹部の発赤あるリンデロン軟膏塗布	0.67	0.40	0.80	0.51	0.31	0.81
153	右部小さな表皮剥離する	右第３趾処置する	0.50	0.25	0.78	0.20	0.34	0.66
154	排尿中に突然電話が鳴り、驚いて洋服に水をこぼしてしまいました。	排尿あり、洋服に更衣介助	0.33	0.29	0.75	0.30	0.14	0.71
155	隔離の為中止	隔離中の為中止	0.83	1.00	0.93	0.87	0.75	0.68
156	全介助を実施する	全更衣介助、リネン交換する	0.75	0.55	0.72	0.52	0.42	0.73
157	トイレ洗面所のところで転倒されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	トイレにて排尿と軟便少量あり	0.25	0.08	0.75	0.16	0.04	0.86

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
158	ソケイに軟膏塗布防水テープ保護する	ソケイ部にミックス軟膏塗布 右内踝部にゲーベン塗布 右肘部にゲーベン塗布	0.42	0.38	0.72	0.39	0.15	1.00
159	右横に軟膏塗布防水テープ保護する	右大転子部、臀部褥瘡跡変化なし	0.50	0.00	0.67	0.00	0.00	0.68
160	右入浴後熱傷部に軟膏塗布防水テープ保護する	右足首～足背にかけて浮腫あり	0.58	0.00	0.72	0.16	0.13	0.79
161	昨日お風呂上がりにシーネ固定をする前にも散歩, 熱心に取り組まれる	昨日、お風呂上がりにシーネ固定をする前には腫れてはいなかった	0.25	0.61	0.87	0.68	0.54	0.77
162	アバ整形外科定期受診	アバ先生、来週中にレントゲンを撮り直すのと、血栓疑いの採血しましょうか、と	0.67	0.08	0.60	0.13	0.01	0.89
163	ただが悪くなっては弱い痛み止めを屯用で出しておくので飲んで心臓の状態が良いとの連絡あり	ただ、ああ医院には血栓をみる特別なスピッツが置いていない、と	0.42	0.31	0.70	0.11	0.20	0.86
164	とりあえず剤処方されることもなく経過は良好との連絡あり	とりあえず採血をまずはしてみましょうと話される	0.42	0.30	0.72	0.18	0.21	0.47
165	アバに上記話すと水曜日に予定入れてさしあげる	アバ CM に上記話すと水曜日に予定入れてくれる	0.75	0.86	0.92	0.78	0.80	0.79

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
166	受診入院との連絡あり	受診後入院との連絡あり	0.83	1.00	0.94	0.87	0.83	0.77
167	左の包帯を外しており顔色も悪くように見える	左肘の皮下出血部の皮膚が浮いてきている	0.25	0.17	0.73	0.00	0.15	0.54
168	がゼと包帯保護する	が－ゼと包帯保護する	1.00	1.00	0.95	0.87	0.83	1.00
169	トレシーバ単位皮下注射施行確認する	トレシーバ 4 単位施行する	0.67	0.67	0.86	0.61	0.37	0.81
170	ガーゼ保護テープついてるので飲んで心臓の状態が良いとの連絡あり	ガーゼ　保護テープついてるので　様子観察している	0.50	0.38	0.88	0.42	0.41	0.98
171	排尿後、手順に従ってパット交換を行います。	排尿あり パット交換する	0.67	0.46	0.81	0.65	0.22	0.69
172	風邪あり	風邪症状あり	0.67	0.80	0.85	0.71	0.54	0.70
173	左がとの連絡あり	左足首に　湿布貼る	0.50	0.00	0.69	0.32	0.16	0.58
174	御されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	御来館され、利用者様とお話される	0.42	0.26	0.67	0.00	0.10	0.85
175	11 月のカレンダー, 熱心に取り組まれる	11 月のカレンダー, 熱心に取り組まれる	0.92	1.00	1.00	1.00	1.00	1.00
176	熱心取り組まれる	熱心に取り組まれる, 他者様と交流される	0.67	0.50	0.76	0.71	0.09	0.69

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
177	両手爪切り介助	両手背に軟膏塗布	0.25	0.22	0.73	0.00	0.00	0.77
178	あまりされることもなく経過は良好との連絡あり	あまり改善されておらず	0.25	0.21	0.73	0.00	0.19	0.35
179	左下部の褥瘡が水疱部より浸出液疼痛あるフシジンレオ塗布する経過は良好との連絡あり	左下腿裂傷部にゲンタシン＋ガーゼ＋防水テープ保護する	0.67	0.07	0.71	0.14	0.12	0.54
180	餃子参加される	餃子召し上がる	0.83	0.33	0.82	0.50	0.18	0.84
181	コロナにて入浴中止する	コロナ抗原検査実施し陰性	0.50	0.18	0.74	0.26	0.21	0.64
182	屋上散歩, 熱心に取り組まれる	屋上に散歩, 他者様と交流される	0.80	0.57	0.77	0.40	0.36	0.75
183	両に皮疹有り	両下肢　ヘパリンクリーム塗布	0.80	0.22	0.67	0.24	0.11	0.67
184	アバ病院受診に行かれる	アバ PT 来館, 熱心に取り組まれる	0.40	0.46	0.72	0.20	0.18	0.38
185	セレコキシブあり	セレコキシブ処方あり	0.80	0.80	0.86	0.71	0.54	1.00
186	テリボン注射施行確認する	テリボン自己注射施行確認する	0.80	0.91	0.94	0.89	0.64	0.64
187	コロナ検査実施し陰性	コロナの為中止	0.50	0.22	0.80	0.29	0.17	0.42
188	御希望より眠前薬飲むかどうかわからないので飲んで心臓の状態が良いとの連絡あり	御本人希望より眠前薬を拒否される	0.42	0.30	0.81	0.45	0.15	0.69
189	PCR 検査結果は陽性ではなく、陰性です。	PCR 検査結果は陰性	1.00	0.67	0.93	0.89	0.39	1.00

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
190	アバ診察	アバ先生診察	1.00	0.80	0.90	0.82	0.54	1.00
191	血液結果炎症反応高ければ明日連絡しますと言われる	血液検査結果炎症反応あるが他はそこまで悪くない	0.25	0.33	0.76	0.53	0.28	0.68
192	肺水も溜まっている姿みる	肺に水も溜まっていない	0.50	0.57	0.87	0.77	0.48	0.61
193	ただが悪くなっては弱い痛み止めを屯用で出しておくので飲んで心臓の状態が良いとの連絡あり	ただ飲み込みが悪く、弱い咳払いがあるので食べ物が気道に入りやすい状態	0.42	0.33	0.75	0.37	0.26	0.68
194	尿路症を起こしている姿みる	尿路感染症を起こしているのだと思う	0.67	0.67	0.84	0.60	0.49	0.64
195	食事入っている姿みる	食事が入っていないみたいなので入院をして点滴治療をし	0.58	0.27	0.70	0.41	0.09	0.77
196	パーキンソン薬飲んで心臓の状態が良いとの連絡あり	パーキンソンの薬も少し入っているが、その薬も調整していきます	0.25	0.27	0.73	0.38	0.12	0.30
197	本日入院との連絡あり	本日より入院となる	0.92	0.55	0.86	0.82	0.44	0.92
198	風船行う, 熱心に取り組まれる	風船バレー行う, 熱心に取り組まれる	0.83	0.92	0.94	0.91	0.86	0.64
199	排便少量 2 回目行なう	排便がないと本人より浣腸希望あり	0.42	0.15	0.70	0.20	0.09	0.36
200	作品, 他者様と交流される	作品見学, 他者様と交流される	0.50	0.92	0.92	0.89	0.87	0.53

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
201	眠剤する清拭後右腋窩と顔に軟膏塗布防水テープ保護する	眠剤注入する清拭後右腋窩と顔に軟膏塗布する左耳にアズノール塗布	0.75	0.75	0.86	0.73	0.65	0.86
202	湿布鎮痛剤処方される	湿布貼付	0.83	0.25	0.84	0.35	0.12	0.99
203	発疹ミックス軟膏塗布防水テープ保護する	発疹増加ミックス軟膏塗布	0.67	0.62	0.89	0.68	0.47	0.43
204	指示軟膏塗布防水テープ保護する	指示の軟膏塗布する, 出血, 膿あり, ガーゼ交換実施	0.42	0.44	0.72	0.43	0.20	0.46
205	次回採血予定	次回採血予定	0.83	1.00	1.00	1.00	1.00	1.00
206	両手爪切り介助	両手背	0.42	0.33	0.86	0.00	0.00	0.99
207	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後熱傷部にエキザルベ塗布する	0.58	0.67	0.80	0.58	0.58	0.59
208	マーズレン　ランソプラゾール処方あるので飲んで心臓の状態が良いとの連絡あり	マーズレン中止　ランソプラゾール処方あるので　次回定期薬より中止とする	0.50	0.41	0.82	0.45	0.40	0.75
209	餃子参加されることもなく経過は良好との連絡あり	餃子レク参加される	0.50	0.44	0.83	0.52	0.28	0.51
210	リドメックスローション	リドメックスローション塗布	0.92	0.67	0.88	0.71	0.00	1.00

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
211	は吐血の可能性が有る事を報告する	「これは吐血の可能性が有るね」と言われる	0.75	0.63	0.81	0.73	0.49	0.71
212	明朝絶食	明朝まで絶食	0.58	0.80	0.89	1.00	0.58	0.57
213	薬飲まれず吐き出しあり	薬も中止する	0.50	0.20	0.76	0.35	0.11	0.60
214	明日ああ病院受診に行かれる	明日、ああ病院受診となる	0.83	0.67	0.89	0.91	0.56	0.84
215	ソルデムの組み合わせによる新たな治療法が 500ml のソルアセト D とともに開始されました。	ソルデム 3AG 500ml ＋ソルアセト D 500ml 施行すされる	0.50	0.40	0.77	0.54	0.29	0.77
216	プリンペラン頓服処方有る	プリンペランの頓服処方有る	0.83	0.89	0.92	1.00	0.78	0.89
217	ケーキフルーツを召し上がる	ケーキとフルーツを召し上がる	0.75	0.89	0.95	1.00	0.77	0.71
218	BS:148/dl ヒューマログ 4 単位皮下注射施行	BS:148/dl ヒューマログ 4 単位皮下注射施行	0.92	1.00	0.97	1.00	1.00	1.00
219	右頬処置施行するも異常なし	右頬の処置する, ガーゼ交換実施	0.50	0.43	0.76	0.45	0.30	0.60
220	発疹に軟膏塗布防水テープ保護する	発疹部にリンデロン VG 軟膏を塗りました。	0.67	0.38	0.74	0.31	0.20	0.62
221	右部後面ゲンタシン軟膏塗布防水テープ保護する	右踵部、浸出液有り中央角質部脱落し陥没有る	0.58	0.00	0.70	0.18	0.12	0.55

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
222	排便が遅れているため、4 日目にして管腸を 30ml 投与しましたが、硬い小石のような便が 6〜7 個しか出ませんでした。	排便なし 4 日目のためかん腸 30ml 施行するも、小石のような硬便が 6〜7 個のみ	0.50	0.43	0.81	0.63	0.37	0.64
223	テレミンソフトウェアは、新しいバージョンにアップデートされました。	テレミンソフト座薬挿肛する	0.25	0.00	0.73	0.18	0.09	0.82
224	リハビリ中止して終了する	リハビリは中止した	0.92	0.55	0.89	0.82	0.45	0.81
225	奥様今後対応主治医との連絡あり	奥様の話し合いのみ実施した	0.58	0.27	0.73	0.24	0.25	0.52
226	排尿が終わったら、トイレを使用した後にパッド交換を行います。	排尿あり、パッド交換, 車椅子にてフロアへ誘導する	0.67	0.24	0.75	0.40	0.20	0.64
227	臀に発赤腫脹浸出液疼痛あるフシジンレオ塗布する経過は良好との連絡あり	臀裂部に発赤剥離あり	0.67	0.35	0.78	0.30	0.18	1.00
228	洗浄綿球つめ、本人が触らないように見える	洗浄しワセリン塗布	0.08	0.14	0.75	0.24	0.08	0.32
229	背部痒いと訴えられるため痒がられるところに保湿剤塗布する	背部乾燥で痒いと言われるためワセリン塗布	0.83	0.45	0.82	0.52	0.34	0.82
230	ただが悪くなっては弱い痛み止めを頼用で出す。	ただれ消失している	0.25	0.11	0.69	0.00	0.05	0.15

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
231	白湯ミリリットル +nacl4 g注入し終了する	白湯 400ml 注入しフラッシュして終了する	0.42	0.53	0.83	0.43	0.52	0.61
232	ご様面会、御来館されなかったら降圧剤	ご家族様に、眼科受診結果電話する	0.67	0.11	0.70	0.17	0.12	0.74
233	両眼内障であること、右眼は緑内障疑いで眼圧が高くなっては弱い痛み止めを屯用で出しておくので飲んで心臓の状態が良いとの連絡あり	両眼白内障であること、右眼は緑内障疑いで眼圧が高くなっていること、だから目が見えにくくなっている、と説明を受けられ、眼圧を下げる点眼薬を処方されたこと	0.58	0.49	0.80	0.51	0.42	0.73
234	口腔と陰部洗浄し綿球つめ	口腔ケアと陰部洗浄しワセリン塗布	0.42	0.63	0.84	0.55	0.59	0.44
235	インフルエンザの症状は高熱と咳です。	インフルエンザ予防接種	0.50	0.18	0.80	0.29	0.10	1.00
236	表皮ほぼ完成して終了する	表皮、ほぼ完成している	0.92	0.77	0.91	0.82	0.66	0.96
237	指示軟膏塗布防水テープ保護する	指示の軟膏塗布する, ガーゼ交換実施	0.67	0.53	0.76	0.50	0.41	0.75
238	リハビリ帰館されることもなく経過は良好との連絡あり	リハビリの為、アバ病院へ送迎を行う	0.50	0.18	0.74	0.18	0.11	0.68
239	両に軟膏塗布防水テープ保護する	両上下肢リンデロン臀部亜鉛華軟膏ミックス軟膏塗布	0.25	0.32	0.70	0.45	0.23	0.89

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
240	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後右大転子部に塗布する 2 回目行なう	0.33	0.47	0.72	0.32	0.33	0.41
241	右転子部に軟膏塗布防水テープ保護する	右第 1 趾爪　ガーゼ汚染あり	0.42	0.00	0.73	0.12	0.11	0.83
242	排尿後、パッド交換を行い、その後車椅子でフロアに誘導する予定です。	排尿あり、リハパンとパッド交換	0.58	0.32	0.79	0.47	0.17	0.41
243	車対応, 出発されることもなく経過は良好との連絡あり	車椅子対応, 出発される	0.67	0.42	0.83	0.61	0.38	0.72
244	ボル投げと輪投げふ, 熱心に取り組まれる	ボ一ル投げと輪投げふ, 熱心に取り組まれる, 他者様と交流される	0.75	0.67	0.84	0.68	0.38	0.79
245	車対応, 出発されることもなく経過は良好との連絡あり	車椅子にてフロアへ誘導する	0.33	0.00	0.67	0.20	0.09	0.69
246	左の包帯を外しており顔色も悪くように見える	左下肢　洗浄し　ゲーベンガーゼ保護する	0.58	0.00	0.67	0.00	0.00	0.62
247	アバ病院整形外科定期受診	アバ病院整形外科定期受診	1.00	1.00	1.00	1.00	1.00	1.00
248	左肩疼痛あり	左肩が痛いと本人の訴えあり	0.83	0.36	0.71	0.35	0.11	0.99
249	昨日痛くて眠られなかったら降圧剤（アムロジピン・アダラート）服用して終了する	昨日は痛くて眠られなかった	0.50	0.50	0.83	0.58	0.36	0.70

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
250	レントゲンあり	レントゲン検査行う	0.92	0.40	0.81	0.58	0.22	0.99
251	骨は低くないので飲んで心臓の状態が良いとの連絡あり	骨折はしていないがひびが入っているかもしれない	0.50	0.19	0.71	0.00	0.10	0.53
252	日々映らない	日々は映らない	0.83	0.80	0.95	1.00	0.73	0.70
253	痛は腿の筋肉を鍛える事で改善されることもなく経過は良好との連絡あり	痛み止めを屯用で出しておくので飲んで下さい	0.40	0.13	0.66	0.00	0.08	0.68
254	痛は腿の筋肉を鍛える事で改善されることもなく経過は良好との連絡あり	痛み止めの塗り薬も処方される	0.33	0.29	0.69	0.00	0.17	0.58
255	処方骨粗鬆症の薬飲んで心臓の状態が良いとの連絡あり	処方薬骨粗鬆症の薬追加して処方される	0.50	0.31	0.81	0.64	0.31	0.57
256	アズノール塗布防水テープ保護する	アズノール軟膏塗布する	0.67	0.60	0.91	0.52	0.55	0.64
257	恥垢ており顔色も悪くように見える	恥垢たまっており、陰臀部洗浄し、アズノール軟膏塗布する	0.42	0.36	0.70	0.18	0.13	0.82
258	ワセリンし着圧ソックス装着する	ワセリン塗布	0.83	0.22	0.84	0.35	0.12	0.80
259	左下部の褥瘡が水疱部より浸出液疼痛あるフシジンレオ塗布する経過は良好との連絡あり	左下腿は足背まで乾燥しているが右下腿は全くなし	0.25	0.12	0.74	0.10	0.13	0.77

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
260	腰痛でお悩みの方には	腰痛訴えあり	1.00	0.20	0.81	0.41	0.12	1.00
261	頓服のロキソプロフェン１錠追加、内服されることもなく経過は良好との連絡あり	頓服薬のロキソプロフェン 1 錠、レバミピド１錠追加、内服される	0.42	0.58	0.87	0.71	0.48	0.55
262	胸部背部　レスタミンミックス軟膏塗布防水テープ保護する	胸部　背部　アズノール　大腿部　リンデロン塗布	0.50	0.40	0.81	0.46	0.25	0.69
263	整形定期受診	整形外科受診, 出発される, 帰館される	0.67	0.33	0.70	0.52	0.03	0.80
264	背部で痒いと訴えられるため痒がられるところに保湿剤塗布する経過は良好との連絡あり	背部、頭部、左上腕内側にミックス軟膏塗布する	0.67	0.26	0.75	0.20	0.14	0.92
265	入浴熱傷部に軟膏塗布防水テープ保護する	入浴されていない	0.42	0.13	0.80	0.35	0.12	0.62
266	排尿の後、床で眠ってしまうことがあります。	排尿あり、オムツ装着、パジャマに更衣介助	0.42	0.22	0.63	0.20	0.18	0.49
267	フルーツ召し上がる	フルーツ召し上がる	1.00	1.00	1.00	1.00	1.00	1.00
268	ソケイに軟膏塗布防水テープ保護する	ソケイ部　皮膚科処方の軟膏塗布する	0.58	0.50	0.82	0.43	0.47	0.48
269	コップ白湯	コップお白湯	1.00	0.80	0.92	1.00	0.58	1.00

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
270	排便前に軽い運動をすると、消化が改善されることがあります。	排便中量少量２回あるため眠前センノシド１錠中止する	0.42	0.15	0.67	0.13	0.14	0.68
271	陰部臀部褥瘡跡変化なし	陰部、臀部、腰部の剥離部の表皮形成中	0.50	0.27	0.72	0.32	0.14	0.68
272	ハロウィンお菓子を召し上がる	ハロウィンお菓子を召し上がる,熱心に取り組まれる	0.67	0.67	0.85	0.71	0.31	0.70
273	排便していない間に、便が軟らかくなることがあります。	排便困難あり摘便する硬便中量ある出血少量有る	0.42	0.24	0.66	0.33	0.20	0.73
274	右足する経過は良好との連絡あり	右足塗布する	0.17	0.33	0.78	0.35	0.18	0.75
275	両手爪切り介助	両手の爪切り介助	1.00	0.89	0.93	1.00	0.70	1.00
276	頭部塗布乾燥して終了する	頭部に塗布	0.58	0.40	0.85	0.71	0.20	0.68
277	腹部部に軟膏塗布防水テープ保護する	腹部に塗布する痔に軟膏塗布する	0.58	0.59	0.83	0.57	0.54	0.24
278	少しも触ると痛いよう	少し、両下腿の発赤が大きくなっていることを報告	0.42	0.12	0.65	0.24	0.05	0.41
279	悪くばまた皮膚科受診しますと言われるところに保湿剤塗布する経過は良好との連絡あり	悪くなればまた皮膚科受診しますとお伝えする	0.75	0.56	0.86	0.49	0.40	0.59
280	右入浴後熱傷部に軟膏塗布防水テープ保護する	右足首の貼る	0.50	0.00	0.73	0.18	0.08	0.99

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
281	ラコール注入 + 食後薬＋白湯 50ml 注入開始する	ラコール 1P 注入 + 夕食後薬注入＋白湯フラッシュ施行し終了	0.50	0.45	0.82	0.67	0.36	0.41
282	リンデロン塗布	リンデロンローション塗布	0.92	0.50	0.92	0.82	0.61	1.00
283	変わり腰や下肢の痛みは腿の筋肉を鍛える事で改善されることもなく経過は良好との連絡あり	変わりなし 腰や下肢の痛みが強くなるようならレントゲンを撮りに来てくだいといわれる	0.92	0.39	0.74	0.38	0.30	0.80
284	脱肛なし	脱肛改善なし	0.83	0.80	0.89	0.71	0.61	1.00
285	リハパン淡血性少量付着有る	リハパンに淡血性少量付着有る	0.67	0.92	0.95	1.00	0.87	0.55
286	バケツに黄土色粘土便、中量有る	バケツ内に黄土色普通便少量付着有る	0.67	0.71	0.84	0.63	0.52	0.97
287	アバ来館, 熱心に取り組まれる	アバ中央病院受診に家族対応にて行かれる	0.67	0.37	0.70	0.17	0.20	0.94
288	左部、自身で軟膏塗って下さいと言われるところに保湿剤塗布する経過は良好との連絡あり	左上肢、頭部の皮疹に指示の軟膏塗布する	0.25	0.30	0.76	0.22	0.20	0.90
289	入浴熱傷部に軟膏塗布防水テープ保護する	入浴の声掛けされるが拒否される	0.42	0.11	0.69	0.20	0.14	0.40

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
290	糖尿骨のデーター良くなっては弱い痛み止めを屯用で出しておくので飲んで心臓の状態が良いとの連絡あり	糖尿内科 骨のデーター良くなっている	0.58	0.37	0.86	0.56	0.23	0.90
291	半年一回の注射で効果あり	半年に一回の注射で効果あり	1.00	0.92	0.96	1.00	0.87	1.00
292	この様子観察して終了する	このまま様子観察	0.83	0.60	0.82	0.82	0.34	0.78
293	また両足趾趾間にニゾラールクリーム塗布	また急激に体重が落ちているため気になるとの事	0.42	0.22	0.67	0.00	0.08	0.38
294	トレシーバ単位皮下注射施行確認する	トレシーバ 4 単位皮下注射施行	0.67	0.77	0.93	0.83	0.48	0.70
295	便ゆるく　パウチ　剥がれかけて浮腫あり	便がゆるく　パウチ　剥がれかけており　交換する	0.75	0.67	0.88	0.89	0.56	0.66
296	剥離は濡らさないように見える	剥離部は濡らさないようにとのこと, 出発される, 帰館される	0.58	0.53	0.74	0.52	0.20	0.50
297	白湯ミリリットル +nacl4 ｇ注入し終了する	白湯 400ml+ 昼食後薬注入しフラッシュして終了する	0.58	0.48	0.80	0.45	0.39	0.70
298	両耳耳鼻科処方の軟膏塗布防水テープ保護する	両耳　耳鼻科処方の点耳　施行する	0.50	0.67	0.88	0.42	0.51	0.80

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
299	4 日前から腰の痛みは腿の筋肉を鍛える事で改善されることもなく経過は良好との連絡あり	４日排便なしグリセリン浣腸施行する	0.67	0.00	0.69	0.23	0.06	0.43
300	腫脹するが痛み無し	腫脹持続するが痛み無し	0.75	0.91	0.92	0.87	0.83	0.80
301	ソケイに軟膏塗布防水テープ保護する	ソケイ部　亜鉛華軟膏塗布	0.50	0.46	0.84	0.46	0.31	0.43
302	アバ病院受診に行かれる	アバ中央病院受診に行かれる	0.67	0.92	0.94	0.89	0.82	0.69
303	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後に臀部褥瘡クロマイ P+ガーゼ保護する	0.42	0.40	0.74	0.24	0.36	0.43
304	また両足趾趾間にニゾラールクリーム塗布	また、両下腿ワセリン塗布しエラスコット巻き直す	0.42	0.25	0.77	0.34	0.22	0.69
305	左部後面ゲンタシン軟膏塗布防水テープ保護する	左大腿部後面ゲンタシン軟膏塗布防水テープ使用	0.58	0.71	0.92	0.67	0.77	0.56
306	バルーンの為中止	バルーン交換の為、アバ泌尿器科へ送迎を行う	0.33	0.40	0.69	0.38	0.09	0.36
307	別異常なし	別に異常なし	1.00	0.80	0.90	1.00	0.63	1.00
308	排尿が頻繁で、軟便少量あり、清拭後にパッドを交換しました。	排尿なし、軟便少量あり、清拭しパッド交換	0.67	0.64	0.88	0.82	0.48	0.67

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine similarity	BLEU	Embed HDP
309	往診れなかったら降圧剤（アムロジピン・アダラート）服用して終了する	往診時､痰が出るようでティッシュに白色痰を少量ずつ出されているが、先生特に何も言われず	0.25	0.17	0.64	0.09	0.08	0.71
310	ロコイド軟膏塗布防水テープ保護する	ロコイドミックス軟膏塗布 左眼オフロキサシン軟膏塗布する	0.58	0.40	0.80	0.57	0.43	0.60
311	花見参加	花見に参加	0.83	0.80	0.87	1.00	0.58	0.87
312	アバ来館, 熱心に取り組まれる	アバ飾りを作られる, 熱心に取り組まれる	0.25	0.67	0.84	0.73	0.55	0.49
313	4 日前から腰の痛みは腿の筋肉を鍛える事で改善されることもなく経過は良好との連絡あり	４日排便無く指示にて施行する	0.25	0.00	0.69	0.23	0.09	0.51
314	皮膚受診しますと言われるところに保湿剤塗布する経過は良好との連絡あり	皮膚科処方の軟膏塗布する	0.50	0.30	0.80	0.30	0.16	0.60
315	右足する経過は良好との連絡あり	右足首入浴後エスパ帯を巻き直す	0.25	0.00	0.68	0.00	0.00	0.58
316	今富来館, 熱心に取り組まれる	今富 PTA 来館, 熱心に取り組まれる	0.58	0.92	0.89	0.91	0.87	0.53
317	白湯ミリリットル +nacl4 ｇ注入し終了する	白湯 500ml 注入開始する	0.42	0.40	0.85	0.32	0.34	0.44

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
318	血圧高い為中止	血圧高いも、労作時だからでしょうとのこと	0.58	0.31	0.67	0.45	0.10	0.80
319	本日採血施行するも異常なし	本日、採血施行する	0.83	0.73	0.88	0.87	0.54	1.00
320	45 日前から腰の痛みは腿の筋肉を鍛える事で改善されることもなく経過は良好との連絡あり	45 日前から腰の痛みがあった、同じようなことが 10 年前にもあった	0.60	0.21	0.76	0.40	0.28	0.81
321	レントゲン骨密度は低くないので飲んで心臓の状態が良いとの連絡あり	レントゲン MRI 骨密度の検査を行われる	0.25	0.30	0.81	0.41	0.13	0.85
322	結果骨には血栓をみる特別なスピッツが置いておいてくださいと指示	結果、腰椎 3 番目の圧迫骨折は随分前のもので今回腰椎 2 番目の圧迫骨折が見られた	0.25	0.16	0.66	0.06	0.11	0.64
323	治る 1 ヶ月はかかる、カロナールと湿布処方されることもなく経過は良好との連絡あり	治るまで 1 ヶ月はかかる、カロナールと湿布処方される	0.58	0.67	0.94	0.84	0.53	0.50
324	骨は低くないので飲んで心臓の状態が良いとの連絡あり	骨密度は低くないのでその分は処方されず	0.25	0.36	0.79	0.34	0.33	0.59
325	コルセット作ることを報告する	コルセットを作ることを勧められる	0.83	0.55	0.91	0.67	0.50	0.73

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
326	居室転倒されなかったら降圧剤(アムロジピン・アダラート）服用して終了する	居室にて転倒された様子報告あり	0.50	0.33	0.75	0.35	0.26	0.80
327	口中はワセリン塗布しソックス装着する	口周囲腫脹、内出血あり	0.25	0.00	0.65	0.00	0.00	0.76
328	ご様面会, 御来館されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	ご家族様要望にてああ病院受診する	0.33	0.15	0.67	0.12	0.12	0.81
329	頭部検査するも異常なし	頭部 MR 検査するも異常なし	0.83	0.92	0.93	0.87	0.82	0.94
330	頭痛吐き気等あれば再度受診するように見える	頭痛や吐き気等あれば再度受診するようにとのこと	0.92	0.78	0.90	0.91	0.72	1.00
331	排尿後、新しいパッドと交換して、水様便のわずかな量をパウチに捨てました。	排尿あり、パッド交換、パウチの水様便少量破棄する	0.50	0.44	0.78	0.61	0.29	0.70
332	本人浣腸希望あり	本人確認にて拒否される	0.33	0.20	0.78	0.33	0.13	0.69
333	排尿後、軽度の下腹部不快感を感じましたが、下痢便少量ありませんでした。	排尿あり、下痢便少量あり、清拭しパッド交換	0.33	0.37	0.74	0.38	0.26	0.87
334	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後体全体に軟膏塗布する	0.50	0.67	0.84	0.43	0.49	0.45

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
335	ビーソフテン塗布	ビーソフテンローション塗布	0.83	0.50	0.91	0.82	0.54	1.00
336	ラコール注入 + 食後薬＋白湯 50ml 注入開始する	ラコール 1P 注入 + 朝食後薬注入＋白湯フラッシュ施行し終了	0.50	0.45	0.83	0.73	0.44	0.50
337	左がとの連絡あり	左かがとの処置する, 創部洗浄	0.33	0.46	0.71	0.35	0.28	0.75
338	FAX で、お知らせいたします。	FAX にて、上記内容伝える	0.58	0.22	0.71	0.29	0.11	0.48
339	とりあえず剤処方されることもなく経過は良好との連絡あり	とりあえず抗生剤を続けて 5 日分くらいだしておくので様子見て	0.33	0.14	0.70	0.24	0.09	0.70
340	ご様面会, 御来館されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	ご家族様面会, 御来館され、利用者様とお話される	0.33	0.41	0.75	0.48	0.33	0.83
341	発疹に軟膏塗布防水テープ保護する	発疹部にミックス軟膏塗布	0.58	0.62	0.87	0.55	0.39	0.59
342	白湯 +nacl4 g注入し終了する	白湯 400 ミリリットル +NACL 4 G+ 抑肝散注入しフラッシュにて終了する	0.58	0.61	0.71	0.63	0.33	0.53
343	両に軟膏塗布防水テープ保護する	両下腿　ワセリン　エラスコット保護する	0.58	0.40	0.76	0.27	0.35	0.76
344	両手爪切り介助	両手背に指示の軟膏塗布する	0.42	0.17	0.69	0.00	0.00	0.81
345	また熱いお茶と飲ませてくださいと指示	また　熱いお茶と言われるため入れてさしあげる	0.83	0.53	0.79	0.28	0.43	0.99

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
346	ポピドン点耳後ベタメタゾン点耳　施行するも異常なし	ポピドンヨード点耳後ベタメタゾン点耳　処置する	0.50	0.53	0.91	0.80	0.59	0.37
347	ああ咽喉科受診しますと言われるところに保湿剤塗布する経過は良好との連絡あり	ああ耳鼻咽喉科受診	0.33	0.16	0.81	0.54	0.16	0.81
348	体の発疹あるところに保湿剤塗布する経過は良好との連絡あり	体全体に軟膏塗布する	0.50	0.29	0.81	0.35	0.16	0.57
349	右足二指の循環不良有り	右足首　湿布貼り　シーネ　包帯交換する	0.33	0.00	0.70	0.00	0.00	0.31
350	朝食のため血液サラサラ度も見ますね〜と	朝食配膳のため、訪室	0.33	0.33	0.76	0.24	0.23	0.97
351	自分酸素カテーテルを外されておらず	自分で酸素カテーテルを外されていた	0.75	0.71	0.91	1.00	0.66	0.77
352	胸部背部　レスタミンミックス軟膏塗布防水テープ保護する	胸部　背部　レスタミンミックス軟膏	0.92	0.62	0.95	0.74	0.45	1.00
353	掻かれているところで痛みが和らぐまで、オイラックスクリームを塗る必要があります。	掻かれているところはオイラックスクリーム塗布する	0.92	0.42	0.87	0.47	0.27	0.87
354	白湯 +nacl4 g注入し終了する	白湯 400ml 滴下後、昼食後薬注入し、20ml フラッシュし終了する	0.42	0.43	0.74	0.29	0.25	0.47

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
355	本日採血施行するも異常なし	本日､コロナウイルスのクラスターにより電話受診となる	0.25	0.11	0.65	0.20	0.08	0.51
356	主治との連絡あり	主治医､アバ先生と直接電話でナースがお話	0.50	0.13	0.71	0.25	0.08	0.68
357	とくにか質問されることもなく経過は良好との連絡あり	とくに何か質問されることもなく4 週間分処方しておきますと、言われる	0.33	0.52	0.73	0.30	0.37	0.35
358	花壇手入れ参加, 熱心に取り組まれる	花壇の手入れ参加, 熱心に取り組まれる	1.00	0.93	0.97	1.00	0.88	1.00
359	ミックス塗布防水テープ保護する	ミックス軟膏　ガーゼ保護する	0.58	0.55	0.87	0.40	0.54	0.67
360	背部箇所あり	背部発赤箇所あり	0.67	0.80	0.87	0.82	0.66	0.70
361	うちわ, 熱心に取り組まれる	うちわ作, 熱心に取り組まれる, 他者様と交流される	0.75	0.67	0.83	0.71	0.30	1.00
362	頭めまいがする為入浴拒否されることもなく経過は良好との連絡あり	頭とめまいがする為入浴拒否される	0.75	0.71	0.92	0.79	0.50	0.91
363	左目腫脹部に軟膏塗布防水テープ保護する	左目の腫脹部にタリビット眼軟膏塗布	0.58	0.59	0.85	0.67	0.51	0.45

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
364	眠前飲むかどうかわからないので飲んで心臓の状態が良いとの連絡あり	眠前薬内確認する	0.30	0.09	0.77	0.16	0.05	0.41
365	体重ている姿みる	体重増えているが、浮腫なし	0.50	0.50	0.70	0.33	0.31	0.32
366	部屋歩いても苦しくない	部屋を歩いても苦しくない	0.92	0.92	0.94	1.00	0.85	0.90
367	時々するが痛み無し	時々下痢するが便秘より良いので、このままの処方	0.42	0.29	0.69	0.26	0.12	0.64
368	貧血薬飲んで心臓の状態が良いとの連絡あり	貧血の薬飲んで心臓の状態が良い	0.83	0.87	0.92	0.93	0.67	0.86
369	利尿追加	利尿剤やめられないのでしっかり飲んで出して	0.58	0.15	0.68	0.29	0.00	0.70
370	仙骨洗浄し綿球つめ、本人が触らないように見える	仙骨部に小指爪大の剥離 (褥瘡) あり	0.50	0.18	0.70	0.14	0.08	0.68
371	洗浄綿球つめ、本人が触らないように見える	洗浄しデュオアクティブドレッシング貼付する	0.33	0.13	0.65	0.18	0.08	0.56
372	アズノール陰部　臀部　亜鉛華軟膏塗布防水テープ保護する	アズノール塗布 陰部　臀部　亜鉛華軟膏ミックス軟膏塗布	0.75	0.67	0.89	0.82	0.65	0.93
373	ラコール注入 + 食後薬＋白湯 50ml 注入開始する	ラコール 1P 注入 + 夕食後薬注入＋白湯フラッシュ施行	0.58	0.50	0.83	0.69	0.41	0.53

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
374	眠前飲むかどうかわからないので飲んで心臓の状態が良いとの連絡あり	眠前薬 + 白湯 50ml フラッシュする	0.33	0.08	0.68	0.12	0.05	0.80
375	排便が困難であるため、ピコスルファートナトリウム内容液を 10 滴追加して内服する予定です。	排便なし 3 日目のためピコスルファートナトリウム内容液 10 滴追加内服する	0.75	0.60	0.88	0.78	0.45	0.69
376	ああ循環器内科定期受診	ああ病院循環器内科定期受診	0.75	0.91	0.91	0.93	0.81	0.71
377	クレアチニン数値が悪いが入院時と比べても苦しくない	クレアチニンの数値が悪いが入院時と変わりない	0.58	0.73	0.94	0.77	0.64	0.91
378	貧血ある打撲後ない湿布貼り様子みる	貧血があるので薬を継続で服用の事	0.33	0.22	0.68	0.17	0.18	0.30
379	血圧も、意味がなく本人に負担をかけるだけとの連絡あり	血圧が高いので降圧剤（アムロジピン・アダラート）服用している	0.33	0.15	0.65	0.15	0.13	0.46
380	降圧（アムロジピン・アダラート）服用して終了する	降圧剤は副作用で浮腫みが出る事がある	0.58	0.11	0.70	0.17	0.08	0.69
381	両浮腫が軽減されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	両下肢浮腫が強く出ているので利尿剤を出すが、浮腫が軽減されなかったら降圧剤を減らします	0.40	0.51	0.79	0.57	0.35	0.78

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
382	左部、自身で軟膏塗って下さいと言われるところに保湿剤塗布する経過は良好との連絡あり	左踵部、石けん洗浄しワセリン＋ガーゼ保護する	0.25	0.07	0.72	0.00	0.15	0.67
383	本日便１回測ってくださいと指示	本日多量便１回少量便１回のため中止する	0.50	0.33	0.74	0.66	0.27	0.67
384	突然排便する	自然排便少量あり	0.58	0.29	0.77	0.41	0.24	0.42
385	そので利尿薬処方するか考えますと言われるところに保湿剤塗布する経過は良好との連絡あり	その後もダラダラと続きそうなため GE60ml 施行する	0.25	0.11	0.65	0.00	0.10	0.53
386	入浴熱傷部に軟膏塗布防水テープ保護する	入浴後ふくらはぎ部の皮膚乾燥部に軟膏塗布する	0.42	0.53	0.80	0.53	0.51	0.45
387	右足二指の循環不良有り	右足第二指の循環不良有り	0.75	1.00	0.96	0.93	0.85	0.79
388	指の変色が気になるため、次回の診察で専門家の意見を聞いてみる予定です。	指の変形で重なる部分の為、次回外科受診時診てもらう予定入れる	0.58	0.40	0.75	0.25	0.33	0.70
389	両浮腫が軽減されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	両下肢、体幹、痒いと言われるところ、乾燥しているところにワセリン塗布	0.25	0.19	0.67	0.12	0.19	0.78

No	Sentence Suggestion	Ground Truth	Expert Score	ROUGE	BERT Score	cosine simi-larity	BLEU	Embed HDP
390	トイレ洗面所のところで転倒されなかったら降圧剤（アムロジピン・アダラート）服用して終了する	トイレと洗面所のところで転倒されていた	0.50	0.55	0.88	0.63	0.43	0.84

Bibliography

[1] Kavita Asnani, Douglas Vaz, Tanay PrabhuDesai, Surabhi Borgikar, Megha Bisht, Sharvari Bhosale, and Nikhil Balaji. Sentence completion using text prediction systems. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014: Volume 1*, pages 397–404. Springer, 2015.

[2] Agustin Garcia Asuero, Ana Sayago, and AG González. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59, 2006.

[3] Davide Ausili, Cecilia Sironi, Laura Rasero, and Amy Coenen. Measuring elderly care through the use of a nursing conceptual model and the international classification for nursing practice®. *International journal of nursing knowledge*, 23(3):146–152, 2012.

[4] Srinivas Bangalore, Owen Rambow, and Steve Whittaker. Evaluation metrics for generation. In *INLG' 2000 Proceedings of the First International Conference on Natural Language Generation*, pages 1–8, 2000.

[5] Roberto J Bayardo, Yiming Ma, and Ramakrishnan Srikant. Scaling up all pairs similarity search. In *Proceedings of the 16th international conference on World Wide Web*, pages 131–140, 2007.

[6] Nicola Brooks. How to undertake effective record keeping and documentation. 2021.

[7] Karla L Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78, 2015.

[8] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256, 2006.

[9] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[10] Eddy Muntina Dharma, F Lumban Gaol, HLHS Warnars, and Benfano Soewito. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *J Theor Appl Inf Technol*, 100(2):31, 2022.

[11] Michele T Di Palo. Rating satisfaction research: is it poor, fair, good, very good, or excellent? *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 10(6):422–430, 1997.

[12] Bahaa Eddine Elbaghazaoui, Mohamed Amnai, and Youssef Fakhri. Predicting the next word using the markov chain model according to profiling personality. *The Journal of Supercomputing*, pages 1–16, 2023.

[13] R Scott Evans. Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, 25(S 01):S48–S61, 2016.

[14] Aejaz Farooq Ganai and Farida Khursheed. Predicting next word using rnn and lstm cells: Stastical language modeling. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 469–474. IEEE, 2019.

[15] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[16] Felix Hill, Kyunghyun Cho, Sebastien Jean, Coline Devin, and Yoshua Bengio. Embedding word similarity with neural machine translation. *arXiv preprint arXiv:1412.6448*, 2014.

[17] Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides GM Petrakis, and Evangelos Milios. Information retrieval by semantic similarity. *International journal on semantic Web and information systems (IJSWIS)*, 2(3):55–73, 2006.

[18] Anna Huang et al. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[19] Kai Huter, Tobias Krick, Dominik Domhoff, Kathrin Seibert, Karin Wolf-Ostermann, and Heinz Rothgang. Effectiveness of digital technologies to support nursing care: results of a scoping review. *Journal of multidisciplinary healthcare*, pages 1905–1926, 2020.

[20] Kazuki Irie, Zhihong Lei, Liuhui Deng, Ralf Schlüter, and Hermann Ney. Investigation on estimation of sentence probability by combining forward, backward and bi-directional lstm-rnns. In *INTERSPEECH*, pages 392–395, 2018.

[21] O ITU. Series p: Telephone transmission quality, telephone installations, local line

networks-subjective video quality assessment methods for multimedia applications, 2008.

[22] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[23] Yasser Khan, Aminy E Ostfeld, Claire M Lochner, Adrien Pierre, and Ana C Arias. Monitoring of vital signs with flexible and wearable medical devices. *Advanced materials*, 28(22):4373–4395, 2016.

[24] Pooja Kherwa and Poonam Bansal. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24), 2019.

[25] JFC Kingman. Theory of probability, a critical introductory treatment, 1975.

[26] Anne Z Kisak and Kathryn J Conrad. Using technology to develop and distribute patient education storyboards across a health system. In *Oncology nursing forum*, volume 31, 2004.

[27] Rajiv Kohli and Sharon Swee-Lin Tan. Electronic health records. *Mis Quarterly*, 40(3):553–574, 2016.

[28] Miori Kubo. *Japanese syntactic structures and their constructional meanings*. PhD thesis, Massachusetts Institute of Technology, 1992.

[29] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*, 2023.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[31] Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop*, 2004.

[32] Nattaya Mairittha, Tittaya Mairittha, and Sozo Inoue. A mobile app for nursing activity recognition. In *Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers*, pages 400–403, 2018.

[33] Piotr Mirowski and Andreas Vlachos. Dependency recurrent neural language models for sentence completion. *arXiv preprint arXiv:1507.01193*, 2015.

[34] Lailil Muflikhah and Baharum Baharudin. Document clustering using concept space

and cosine similarity measurement. In *2009 International conference on computer technology and development*, volume 1, pages 58–62. IEEE, 2009.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[37] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1, 2012.

[38] Omor Faruk Rakib, Shahinur Akter, Md Azim Khan, Amit Kumar Das, and Khan Mohammad Habibullah. Bangla word prediction and sentence completion using gru: an extended version of rnn on n-gram language model. In *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–6. IEEE, 2019.

[39] Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.

[40] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[41] Natalie Schluter. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–45. Association for Computational Linguistics, 2017.

[42] Masayoshi Shibatani, Shigeru Miyagawa, and Hisashi Noda. *Handbook of Japanese syntax*, volume 4. Walter de Gruyter GmbH & Co KG, 2017.

[43] Yuko Shimomura, Hiroyuki Kawabe, Hidetaka Nambo, and Shuichi Seto. The translation system from japanese into braille by using mecab. In *Proceedings of the Twelfth International Conference on Management Science and Engineering Management*, pages 1125–1134. Springer, 2019.

[44] Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P Martin. Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *Journal of*

the American Medical Informatics Association, 28(1):155–166, 2021.

[45] Sitender, Sangeeta, N Sudha Sushma, and Saksham Kumar Sharma. Effect of glove, word2vec and fasttext embedding on english and hindi neural machine translation systems. In *Proceedings of Data Analytics and Management: ICDAM 2022*, pages 433–447. Springer, 2023.

[46] MD SOLOMON GARB, ELEANOR KRAKAUER, and CARSON JUSTICE. Abbreviations and acronyms in medicine and nursing.

[47] Sue Stevens and Dianne Pickering. Keeping good nursing records: a guide. *Community eye health*, 23(74):44, 2010.

[48] The Royal Children's Hospital Melbourne. Nursing documentation principles.

[49] Cagatay Neftali Tulu. Experimental comparison of pre-trained word embedding vectors of word2vec, glove, fasttext for word level semantic text similarity measurement in turkish. *Advances in Science and Technology. Research Journal*, 16(4), 2022.

[50] Noelia Vicente Oliveros, Teresa Gramage Caro, Covadonga Pérez Menendez-Conde, Ana María Álvarez-Díaz, Sagrario Martín-Aragón Álvarez, Teresa Bermejo Vicedo, and Eva Delgado Silveira. Effect of an electronic medication administration record application on patient safety. *Journal of Evaluation in Clinical Practice*, 23(4):888–894, 2017.

[51] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.

[52] Shirley V Wang, Sebastian Schneeweiss, Marc L Berger, Jeffrey Brown, Frank de Vries, Ian Douglas, Joshua J Gagne, Rosa Gini, Olaf Klungel, C Daniel Mullins, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies v1. 0. *Value in health*, 20(8):1009–1022, 2017.

[53] Patricia A Williams. *Fundamental Concepts and Skills for Nursing-E-Book*. Elsevier Health Sciences, 2021.

[54] Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information sciences*, 307:39–52, 2015.

[55] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*,

2019.