Article

Not peer-reviewed version

An Improved Evaluation Metrics for Sentence Suggestions in Nursing and Elderly Care Record Applications

Defry Hamdhana * , Haru Kaneko , John Noel Victorino , Sozo Inoue

Posted Date: 18 December 2023

doi: 10.20944/preprints202312.1249.v1

Keywords: sentence suggestion; nursing care record; evaluation metrics; elderly care record

Article

An Improved Evaluation Metrics for Sentence Suggestions in Nursing and Elderly Care Record Applications

Defry Hamdhana [1,2,*], Haru Kaneko [1], John Noel Victorino [1] and Sozo Inoue [1]

[1] Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu 808-0196, Japan;
[2] Department of Informatics, Universitas Malikussaleh, Aceh Utara 24355, Indonesia
[*] Correspondence: defryhamdhana@unimal.ac.id

Abstract: In this paper, we propose a novel approach named EmbedHDP to enhance the evaluation models used to assess sentence suggestions within nursing care record applications. The key focus is determining whether these suggestions garner assessments that align with caregivers as human evaluators. It is crucial due to the direct relevance of the information provided to the health or condition of the elderly. The motivation behind this proposal stems from challenges observed in previous models, such as BERTScore, which encountered difficulties in effectively evaluating the nurse care record domain, consistently providing quality assessments of generated sentence suggestions above 60%. Additionally, while widely used, cosine similarity exhibits limitations concerning word order, leading to potential misjudgments of semantical differences within similar word sets. Similarly, relying on lexical overlap, ROUGE tends to overlook semantic accuracy. Furthermore, despite its utility, BLEU neglects semantic coherence in its evaluations. EmbedHDP excels in evaluating nurse care records by effectively handling a variety of sentence structures and medical terminology and providing differentiated and contextually relevant assessments. We used a dataset comprising 320 pairs of sentences with correspondingly equivalent lengths. The results revealed that EmbedHDP outperformed other evaluation models, achieving a coefficient score of 61%, followed by cosine similarity with a score of 59%, and BERTScore with 58%. This shows the effectiveness of our proposed approach in improving the evaluation of sentence suggestions in nursing care record applications.

Keywords: sentence suggestion; nursing care record; evaluation metrics

1. Introduction

The goal of sentence suggestion is to complete a sentence based on the context or words provided by the caregivers. In this task, the machine model needs to understand the word structure inputted by the caregiver and predict the right word to complete it into a sentence [1,2]. The expected result of sentence suggestion is to complete sentences similar to the sentences in the database as ground truth. This is used to increase communication speed and reduce the time required to write text [3]. We adapted the sentence suggestion in the nursing care record application to help caregivers report elderly patients' condition periodically quickly and efficiently [4]. Thus, caregivers can do other activities to help elders while written reports are more precise or typo-free. Furthermore, the generated sentence suggestions must exhibit information similar to their ground truths, considering that information within nursing care records is crucially linked to the condition and health of the elders. A nursing care record is a collection of electronic health records documenting an elder's health care services and events. The information in elderly care includes diagnoses, examination results, care plans, prescribed medications, and performed medical procedures [5]. Nursing care records additionally incorporate meticulous documentation of care provided, ranging from caregivers to physicians, and encompassing 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.

The study encounters challenges in devising appropriate evaluation metrics for assessing the quality of sentence suggestions generated within nursing care records. These challenges include diverse sentence structures and specialized medical terminology, which pose difficulties for conventional metrics designed for well-formed sentences and general language understanding. The subjectivity in human evaluation, rooted in caregivers' contextual understanding of elderly care [6], adds complexity, making it challenging to create automated metrics that authentically mimic human judgment. Additionally, the context sensitivity of nursing care records, where information is highly context-dependent [7], requires nuanced evaluation methods that existing metrics may struggle to provide. Another challenge arises from using Japanese in nursing care record applications, which possess non-conventional language structures. The intricacies of the Japanese language, characterized by unique structures and grammar [8], present additional hurdles in generating and evaluating sentence suggestions. The non-standard nature of Japanese sentence structures requires specialized adaptation in language models and evaluation metrics. Addressing these challenges is crucial to improving the accuracy of evaluation metrics tailored for the distinct characteristics of nursing care record sentences and content.

We delve into the critical need to evaluate the models generated by our system, with the primary objective of ensuring the reliability of the sentences produced by these models in reporting elderly conditions and exhibiting similar information to a set of baseline sentences previously reported by caregivers. Our analysis examines the mechanisms of current evaluation metrics such as BERTScore, Cosine Similarity, ROUGE, and BLEU to achieve robust metrics evaluation. Utilizing sample data assessed by three caregivers working in elderly care facilities in Japan, we identify certain shortcomings in current evaluation metrics when measuring similar information between sentence suggestions and ground truth. BERTScore uses pre-trained BERT models. It may not effectively capture domain-specific nuances such as nurse care records. Cosine similarity fails to consider word order, potentially causing errors in assessing semantic differences in groups of similar words. ROUGE may ignore semantic accuracy because it bases its evaluation on lexical overlap, ignoring context and word meaning. In addition, BLEU ignores semantic coherence. This exploration underscores the importance of refining existing evaluation metrics to effectively gauge the nuanced nature of sentence suggestions within the context of elderly care facilities.

We propose EmbedHDP, a hybrid topic model integrated with word embedding, intending to analyze the quality of sentence suggestion generation in nursing care record applications. The Hierarchical Dirichlet Process (HDP) is a sensible approach for training models on nursing care records with incomplete or fragmented sentences. HDP's flexibility, devoid of prior specifications on the number of groups or topics, makes it adept at handling non-standard sentence structures. Its hierarchical nature enhances adaptability for modeling complex and uncertain data structures. For care records containing occasional medical terminology, EmbedHDP is effective. Utilizing a dictionary with words common in care records, including medical terms, EmbedHDP calculates topic distributions during model training. In addressing the context sensitivity of care records, Bag-of-Words (BoW) falls short. Unlike BoW, Word Embedding creates vectors capturing the semantic meaning of words, allowing for a nuanced understanding of context. Word Embedding is more suitable for representing the intricate context within care records.

We used a dataset comprising 320 pairs of sentences with correspondingly equivalent lengths and evaluated them using both EmbedHDP and the current evaluation metrics. By using caregiver assessments as a benchmark, we calculated the coefficient score for each evaluation model. The results indicate that our proposed evaluation model, EmbedHDP, outperformed other evaluation models, achieving a score of 61%, followed by Cosine Similarity with 59%, and BERTScore with 58%.

The contributions we have made are as follows:

1. Conducted the collection of a dataset comprising 390 pairs of sentences, consisting of sentence suggestions and their corresponding ground truth. Three caregivers meticulously evaluated this dataset and subsequently utilized it as the benchmark ground truth.

2. Perform testing and analysis on current evaluation metrics such as BERTScore, Cosine Similarity, ROUGE, and BLEU to assess the quality of sentence suggestions in nursing care applications.
3. Introduced an innovative evaluation model featuring a novel approach that significantly improved the performance compared to existing evaluation models, aligning more closely with caregiver assessments.

Collectively, these contributions represent our efforts to enhance the methodology and accuracy of evaluating sentence suggestions within nursing care record applications.

2. Related Works

In this chapter, we will discuss several research studies conducted by experts in the field. These studies have inspired us to propose a new model for evaluating the quality of sentence suggestions, particularly in the context of care records. We will also explore the existing evaluation metrics used to assess sentence suggestions.

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. Elderly care can be applied in nursing because it addresses the unique healthcare needs and considerations associated with the aging population, encompassing a holistic approach that involves physical, mental, and emotional well-being. 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 [9] 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 [10]. Initially, nursing care records only consisted of digital medical documents that were easily searchable and accessible to health professionals. 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. However, with technological developments, care records are now more complex and capable of collecting, processing, and analyzing patient health data on a large scale. However, care records also have challenges related to their use, such as data security, interoperability, and proper use by health professionals.

In this study, we used FonLog as a nursing care record application installed in more than 30 healthcare facilities in Japan. FonLog[11] 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 1.

Table 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), 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)

FonLog's Notices 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 1 shows the notices input for the vital activity type, which records extra information about the patient's vital activity.

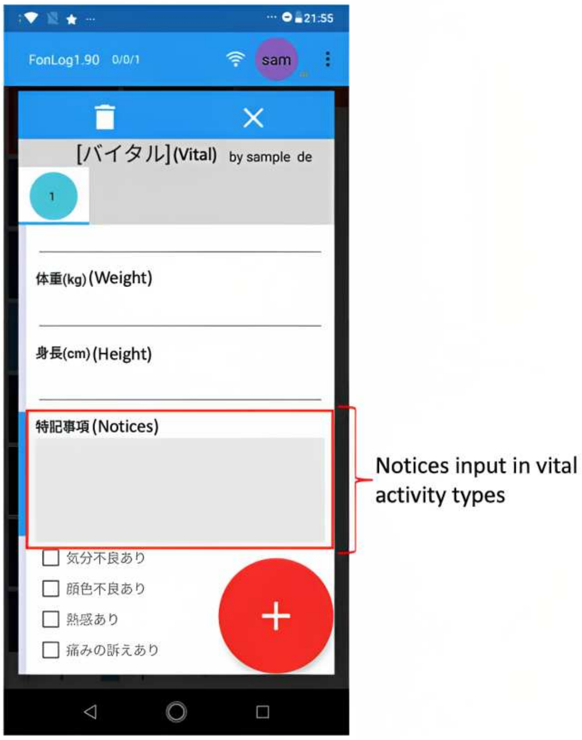

Figure 1. Notices input in FonLog Application

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[12]. 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.

To effectively address the challenges posed by diverse sentence structures and specialized medical terminology in care records, we require an evaluation metric that mirrors the discernment of caregivers. This metric should possess the ability to assess the generated sentence suggestions for high information similarity with sentences in the care record database, serving as the ground truth. It is crucial that the evaluation metric not only evaluates grammatical accuracy but also recognizes nuanced language structures and specialized medical terminology, aligning closely with the expertise of healthcare practitioners.

The current evaluation metrics, however, fall short of addressing these two challenges due to their limitations in recognizing diverse language structures and specialized medical terminology. The existing metrics struggle to navigate the intricacies of varied sentence constructions and accurately evaluate the appropriateness of language, hindering their effectiveness in the context of nursing care records. Additionally, their inadequacy in deciphering the nuanced medical terminology further underscores the need for a more sophisticated evaluation approach. Developing metrics to overcome these limitations is essential for ensuring accurate assessments of generated content, especially in the intricate landscape of care record sentences.

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. [3] 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 [13] 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. [14] 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 computation time. Therefore, this study concludes that combining forward, backwards, and bi-directional LSTM-RNNs is the most effective method for estimating sentence probabilities. Rakib et al. [15] 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.

2.3. Current Evaluation Metrics

Numerous evaluation metrics are commonly employed for various general tasks, including assessing semantic or syntactic similarity, conducting evaluations in summarizing 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 below, each current evaluation metric is described along with its limitations.

Table 2. Comparison of Evaluation Metrics for Sentence Suggestion in Care Records

Evaluation Metrics	Mechanism	Limitation
BERTScore [16]	Comparing contextual embedding 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 [17]	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 semantically difference with similar word sets.
ROUGE [18]	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 [19]	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 4, 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 embedding 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.

Based on the related works discussed earlier, our research motivation is to propose an evaluation metric that aligns effectively with caregivers' opinions. Recognizing the limitations of existing metrics in capturing the nuances of diverse sentence structures and specialized medical terminology, our objective is to develop a metric specifically relevant to the complexities of sentence suggestion generation in the context of nursing care records. This research endeavor seeks to bridge the distance between conventional evaluation metrics and the nuanced expectations of healthcare professionals.

3. Proposed Method for Evaluation Metrics

In this paper, we propose EmbedHDP to address several challenges inherent in nursing care record sentences, such as diverse sentence structures and medical terminology. The utilization of the Hierarchical Dirichlet Process (HDP) for training represents a judicious approach. HDP offers flexibility by eliminating the need for prior specifications regarding the number of groups or topics, which is particularly crucial in the case of non-standard sentence structures. The hierarchical nature of HDP enhances its adaptability, rendering it a potent tool for modeling data with intricate and uncertain structures. Meanwhile, word embedding plays a crucial role in mapping vector distances between words in nursing care records, subsequently undergoing unsupervised training with a dictionary comprising a collection of words found in nursing 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 [20]. The essence of topic modeling is to identify hidden patterns in the text and discover topics that are interconnected 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 [21] 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 the need for users 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:

Input: sentence1 = sentence similarity, sentence2 = ground truth
Output: similarity score between sentence1 and sentence2

1. S_1 be the set of unique tokens in sentence1,
2. S_2 be the set of unique tokens in sentence2,
3. D be the dictionary formed by combining S_1 and S_2,
4. C_1 be the Bag of Words (BoW) vector representing sentence1 in the corpus,
5. C_2 be the BoW vector representing sentence2 in the corpus,
6. $\text{HDP}(D, [C_1, C_2])$ be the trained Hierarchical Dirichlet Process model with dictionary D and corpus $[C_1, C_2]$,
7. T_1 be the topic distribution vector for sentence1 obtained from the trained HDP model,
8. 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:

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

Due to the challenges posed by incomplete or fragmented sentences, the use of the Hierarchical Dirichlet Process (HDP) for training represents a sensible approach. HDP 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 within HDP provides flexibility and adaptability, making it an effective tool for modeling data with complex and uncertain structures. Here is an example of the utilization of HDP that yields favorable assessments in incomplete sentences.

Table 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.

doi:10.20944/preprints202312.1249.v1

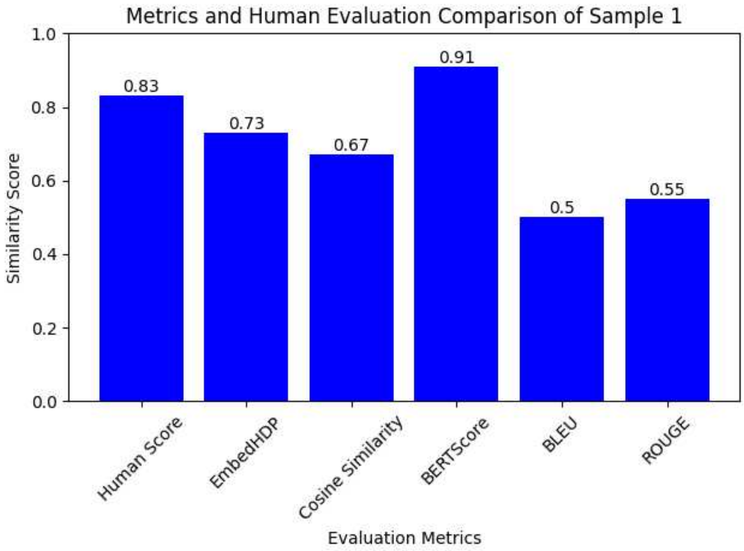

Figure 2. Metrics and human opinion assessment of Sample 1

The scores assigned by humans serve as 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. The incorporation of 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.

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 [22], GloVe [23], and FastText [24], transforms words into dense numerical vectors, capturing intricate semantic relationships and nuanced contextual information. The mechanism behind 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.

To examine the strengths and weaknesses of word embedding models, we can refer to the table presented below:

Table 4. Strength and Weakness of Word Embedding Models

Word Embedding Model	Strength	Weakness
Word2Vec	Semantic similarity, efficient, contextual understanding	Out-of-Vocabulary words, limited word level, insensitive to word order
GloVe	Global context, effective for common words, linear structure	Less effective for Rare words, limited contextual understanding
fasText	Sub-word information, better representation for rare words, efficiency	Computationally more demanding, memory usage

doi:10.20944/preprints202312.1249.v1

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[25]. 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 been conducted, spanning diverse linguistic contexts and applications, to assess the efficacy of various word embedding models. Investigations have ranged from comparing pre-trained word embedding vectors for word-level semantic text similarity in Turkish [26] to evaluating Neural Machine Translation (NMT) for languages such as English and Hindi [27]. Additionally, an exploration of the accuracy of three prominent word embedding models within the context of Convolutional Neural Network (CNN) text classification [28] has been undertaken. 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.

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 5. 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 canceled 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.

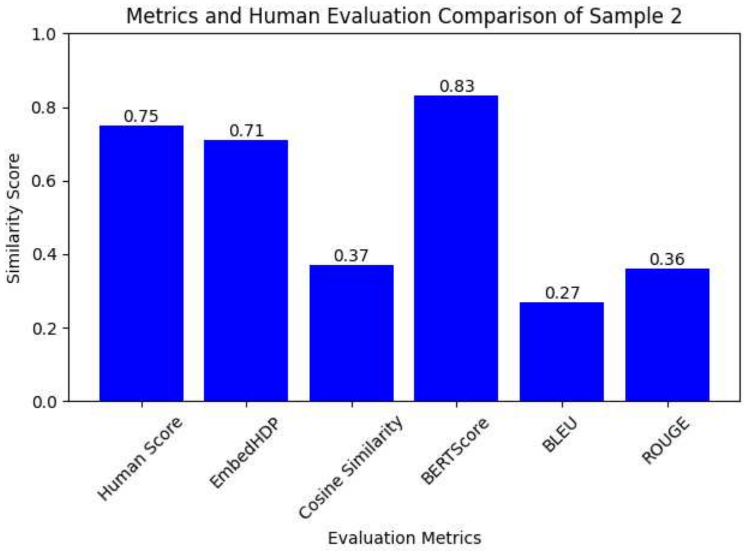

Figure 3. Metrics and human opinion assessment of Sample 2

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

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 6. 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 to have 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 363 several other current evaluation metrics for Sample 3.

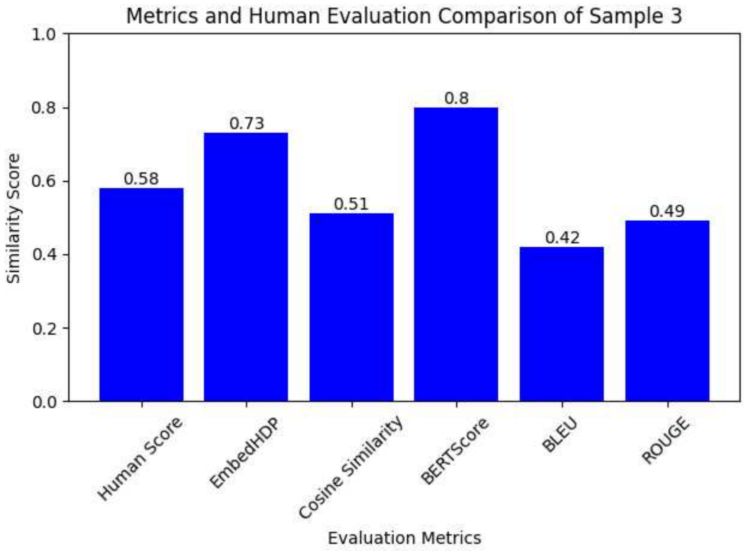

Figure 4. Metrics and human opinion assessment of Sample 3

3.3. *EmbedHDP*

In an effort to improve accuracy and diversify representations, we introduce the EmbedHDP model, which combines HDP with word embedding techniques, overcoming the limitations of the BoW model and improving contextual understanding between words in a sentence. As a result, we attempt to convert the tokens in both sentences into vectors that are generated by the word embedding, fastText in this case, by applying several parameters to each token in each sentence. The following is the mathematical equation for EmbedHDP:

Input: sentence1 = sentence suggestion, sentence2 = ground truth

Output: similarity score between sentence1 and sentence2

1. S_1 be the set of unique tokens in sentence1,
2. S_2 be the set of unique tokens in sentence2,
3. D be the dictionary formed by a unique set of words in the care record dataset,
4. T_1 be the tokenization by using Mecab and particles from S_1,
5. T_2 be the tokenization by using Mecab and particles from S_2,
6. C_1 be the fastText vector from T_1,
7. C_2 be the fastText vector from T_2,
8. HDP$(D, [C_1, C_2])$ be the trained Hierarchical Dirichlet Process model with dictionary D and corpus $[C_1, C_2]$,
9. T_1 be the topic distribution vector for sentence1 obtained from the trained HDP model,
10. 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:

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

3.3.1. Tokenization

The Japanese language captivates attention with its unique linguistic structure, where verbs occupy the final position in sentences [29]. 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) [30]. 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 [8]. 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 7. 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."
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 stands as the most crucial element in training HDP to derive topics. By default, the process of 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 lies in the fact 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.

4. Evaluation

The objective of EmbedHDP is to enhance the performance of evaluation metrics when assessing sentence suggestions generated by the model in comparison with their respective ground truths within the scope of care records. In this section, we will present evidence that EmbedHDP yields promising results in evaluating 228 data samples. This is evidenced by the correlation coefficient score of EmbedHDP, which consistently outperforms the correlation coefficients of current evaluation metrics. This indicates that EmbedHDP consistently produces similarity scores between sentence suggestions and ground truths that closely align with the scores provided by human evaluation. In other words, EmbedHDP can effectively measure the similarity between the two sentences, by capturing topics that are unearthed between them.

4.1. Filtering Data Sample

In Section 3, we have discussed the limitations associated with 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. In response to this consideration, we have implemented a data filtering criterion for testing our proposed evaluation model. 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 aims to facilitate 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:

```
Input: sentence1 (sentence suggestion), sentence2 (ground truth)
if len(sentence1) > 13 or len(sentence2) > 13:

# If either sentence is longer than 13 words, eliminate both sentences
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 the majority 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.

4.2. Evaluation Metrics Utilized

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 8 below shows that EmbedHDP outperforms current evaluation metrics in coefficient score on 320 test data.

Table 8. Comparison of EmbedHDP with other current evaluation metrics

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

EmbedHDP has demonstrated its success in surpassing 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 60%, 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. The presence of outliers can pose a significant challenge in data evaluation and analysis, including within the context of utilizing EmbedHDP and current evaluation metrics. Outliers have the potential to 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 9 shows the correlation coefficient for 70 data as outliers.

Table 9. Comparison of EmbedHDP with other current evaluation metrics

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

4.3. 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 10. 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.)

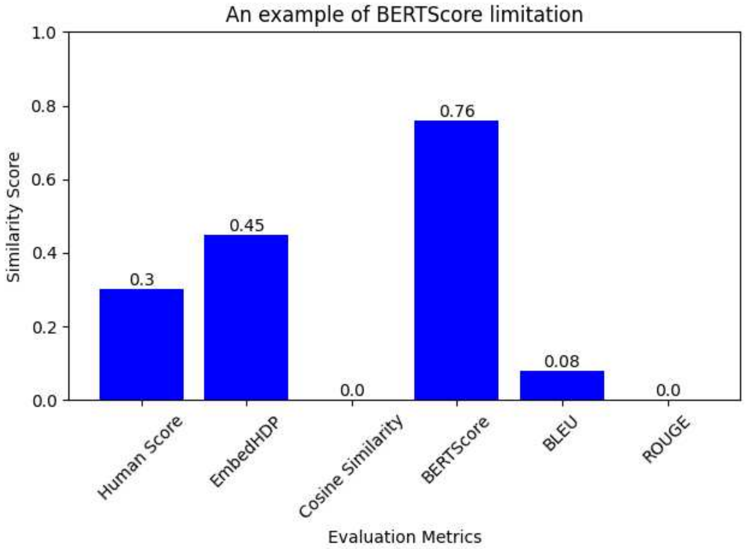

Figure 5. Metrics and human evaluation assessment of BERTScore limitation sentence

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

Table 11. 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)

doi:10.20944/preprints202312.1249.v1

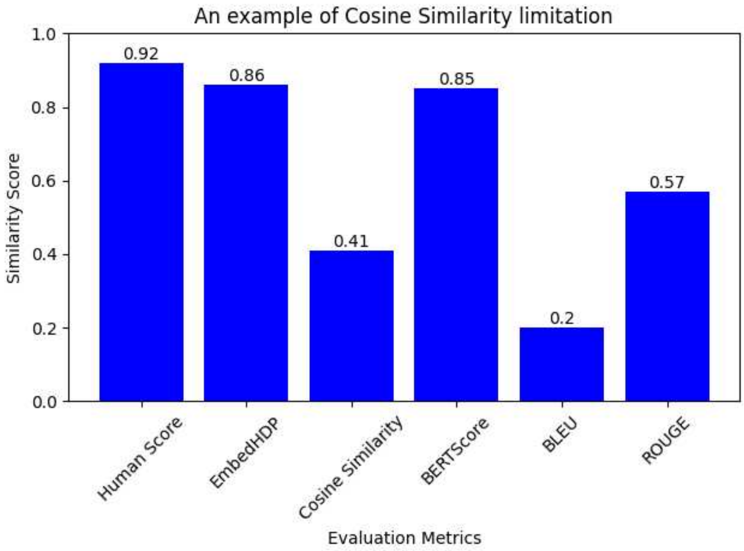

Figure 6. Metrics and human evaluation assessment of Cosine Similarity limitation sentence

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

Table 12. 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)

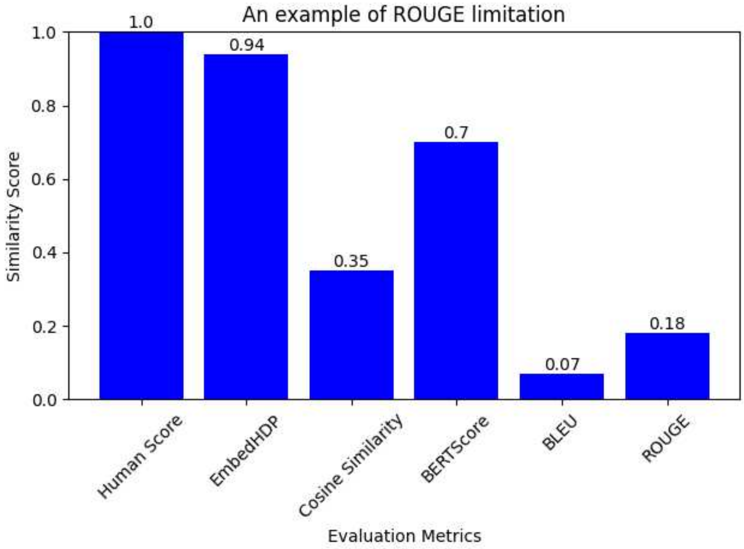

Figure 7. Metrics and human evaluation assessment of ROUGE limitation sentence

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

Table 13. 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。	熱発の為、ナースに報告し中止。
(I had a fever, so I informed the nurse and cancelled the session)	(Due to fever, the nurse was informed and the procedure was discontinued)

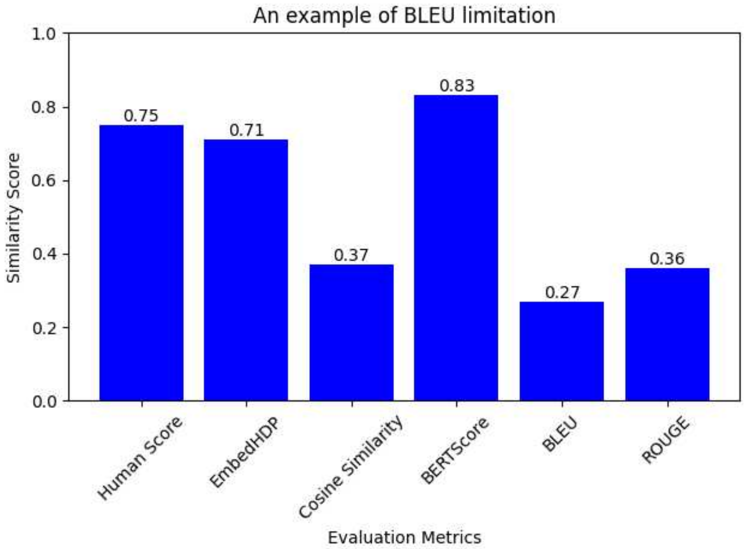

Figure 8. Metrics and human evaluation assessment of ROUGE limitation sentence

5. Discussion

This research aims to provide an evaluation metric that can better capture the semantic information in advice sentences in care notes and produce assessments that reflect human evaluations. The goal is to improve the evaluation process by developing metrics that not only consider syntactic correctness but also investigate the semantic richness and relevance of the generated content. Alignment with human evaluation emphasizes the importance of creating metrics that go beyond surface-level linguistic features and seek to measure deeper understanding such as capturing the topic and contextuality of suggested sentences. Ultimately, this effort contributes to a more accurate and meaningful evaluation in the domain of making sentence suggestions in care record applications.

Nevertheless, it must be acknowledged that EmbedHDP still possesses limitations that necessitate careful analysis for future development. One significant challenge lies in EmbedHDP's vulnerability when analyzing the similarity between two care record sentences, particularly when these sentences are sufficiently lengthy, containing 14 words or more. The model encounters difficulties in effectively capturing the nuances of the relationship between sentences in such scenarios.

Addressing this limitation becomes crucial for refining EmbedHDP to be applicable more comprehensively, especially in situations where significant variations in sentence length and hierarchical relationships between sentences are common. Overcoming these challenges is imperative to enhance EmbedHDP's capabilities, ensuring its effectiveness in a broader range of applications. This highlights the ongoing need for advancements in natural language processing models, emphasizing the importance of adapting to diverse linguistic contexts and complexities inherent in real-world datasets.

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 the 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 the domain of care records, where language is highly specialized and context-dependent, adapting evaluation metrics like BERTScore through fine-tuning is a strategic move.

6. Conclusions

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 pose specific challenges, including diverse sentence structures and specialized medical terminology. We introduce evaluation metrics that address the two main challenges in care records by employing a methodology that calculates topic similarity with the assistance of word embedding vectors. This innovative approach aims to overcome the challenges of diverse sentence structures and specialized medical terminology in care records. By leveraging 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, providing a more accurate and contextually relevant assessment of sentence suggestions in the domain of care records. This approach reflects a commitment to advancing evaluation techniques tailored to the unique linguistic characteristics of healthcare data.

The main contribution of this paper lies in the success of our proposed evaluation model, which has outperformed current evaluation metrics using a coefficient correlation approach as a measure. This novel methodology, particularly applied in EmbedHDP, demonstrates the model's capability to capture both the strength and direction of the relationship with human evaluation.

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.

References

1. Yang, T.; Deng, H. Intelligent sentence completion based on global context dependent recurrent neural network language model. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing, 2019, pp. 1–5.
2. Park, H.; Park, J. Assessment of word-level neural language models for sentence completion. *Applied Sciences* **2020**, *10*, 1340.
3. Asnani, K.; Vaz, D.; PrabhuDesai, T.; Borgikar, S.; Bisht, M.; Bhosale, S.; Balaji, N. Sentence completion using text prediction systems. Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014: Volume 1. Springer, 2015, pp. 397–404.
4. Mamom, J. Digital technology: innovation for malnutrition prevention among bedridden elderly patients receiving home-based palliative care. *Journal of Hunan University Natural Sciences* **2020**, *47*.
5. Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; McLachlan, S. 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* **2018**, *25*, 230–238.

6. Churruca, K.; Ludlow, K.; Wu, W.; Gibbons, K.; Nguyen, H.M.; Ellis, L.A.; Braithwaite, J. A scoping review of Q-methodology in healthcare research. *BMC medical research methodology* **2021**, *21*, 125.
7. of General Practicioners, R.A.C.; others. Privacy and managing health information in general practice, 2021.
8. Shibatani, M.; Miyagawa, S.; Noda, H. *Handbook of Japanese syntax*; Vol. 4, Walter de Gruyter GmbH & Co KG, 2017.
9. Caballero Barajas, K.L.; Akella, R. Dynamically modeling patient's health state from electronic medical records: A time series approach. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 69–78.
10. Evans, R.S. Electronic health records: then, now, and in the future. *Yearbook of medical informatics* **2016**, *25*, S48–S61.
11. Mairittha, N.; Mairittha, T.; Inoue, S. A mobile app for nursing activity recognition. Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers, 2018, pp. 400–403.
12. Stevens, S.; Pickering, D. Keeping good nursing records: a guide. *Community eye health* **2010**, *23*, 44.
13. Mirowski, P.; Vlachos, A. Dependency recurrent neural language models for sentence completion. *arXiv preprint arXiv:1507.01193* **2015**.
14. Irie, K.; Lei, Z.; Deng, L.; Schlüter, R.; Ney, H. Investigation on estimation of sentence probability by combining forward, backward and bi-directional LSTM-RNNs. INTERSPEECH, 2018, pp. 392–395.
15. Rakib, O.F.; Akter, S.; Khan, M.A.; Das, A.K.; Habibullah, K.M. Bangla word prediction and sentence completion using GRU: an extended version of RNN on N-gram language model. 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2019, pp. 1–6.
16. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* **2019**.
17. Rahutomo, F.; Kitasuka, T.; Aritsugi, M. Semantic cosine similarity. The 7th international student conference on advanced science and technology ICAST, 2012, Vol. 4, p. 1.
18. Schluter, N. The limits of automatic summarisation according to rouge. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2017, pp. 41–45.
19. Reiter, E. A structured review of the validity of BLEU. *Computational Linguistics* **2018**, *44*, 393–401.
20. Kherwa, P.; Bansal, P. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems* **2019**, *7*.
21. Kingman, J. Theory of Probability, a Critical Introductory Treatment, 1975.
22. Goldberg, Y.; Levy, O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* **2014**.
23. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
24. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* **2016**.
25. Hill, F.; Cho, K.; Jean, S.; Devin, C.; Bengio, Y. Embedding word similarity with neural machine translation. *arXiv preprint arXiv:1412.6448* **2014**.
26. Tulu, C.N. 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* **2022**, *16*.
27. Sitender.; Sangeeta.; Sushma, N.S.; Sharma, S.K. Effect of GloVe, Word2Vec and FastText Embedding on English and Hindi Neural Machine Translation Systems. In *Proceedings of Data Analytics and Management: ICDAM 2022*; Springer, 2023; pp. 433–447.
28. Dharma, E.M.; Gaol, F.L.; Warnars, H.; Soewito, B. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *J Theor Appl Inf Technol* **2022**, *100*, 31.
29. Kubo, M. Japanese syntactic structures and their constructional meanings. PhD thesis, Massachusetts Institute of Technology, 1992.

30. Shimomura, Y.; Kawabe, H.; Nambo, H.; Seto, S. The translation system from Japanese into braille by using MeCab. Proceedings of the Twelfth International Conference on Management Science and Engineering Management. Springer, 2019, pp. 1125–1134.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.