Article published online: 2021-07-08



# Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT

Faith Wavinya Mutinda[1] Shuntaro Yada[1] Shoko Wakamiya[1] Eiji Aramaki[1]

[1] Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara, Japan

Methods Inf Med 2021;60:e56–e64.

Address for correspondence Eiji Aramaki, PhD, Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan (e-mail: aramaki@is.naist.jp).

## Abstract

**Background** Semantic textual similarity (STS) captures the degree of semantic similarity between texts. It plays an important role in many natural language processing applications such as text summarization, question answering, machine translation, information retrieval, dialog systems, plagiarism detection, and query ranking. STS has been widely studied in the general English domain. However, there exists few resources for STS tasks in the clinical domain and in languages other than English, such as Japanese.

**Objective** The objective of this study is to capture semantic similarity between Japanese clinical texts (Japanese clinical STS) by creating a Japanese dataset that is publicly available.

**Materials** We created two datasets for Japanese clinical STS: (1) Japanese case reports (CR dataset) and (2) Japanese electronic medical records (EMR dataset). The CR dataset was created from publicly available case reports extracted from the CiNii database. The EMR dataset was created from Japanese electronic medical records.

**Methods** We used an approach based on bidirectional encoder representations from transformers (BERT) to capture the semantic similarity between the clinical domain texts. BERT is a popular approach for transfer learning and has been proven to be effective in achieving high accuracy for small datasets. We implemented two Japanese pretrained BERT models: a general Japanese BERT and a clinical Japanese BERT. The general Japanese BERT is pretrained on Japanese Wikipedia texts while the clinical Japanese BERT is pretrained on Japanese clinical texts.

**Results** The BERT models performed well in capturing semantic similarity in our datasets. The general Japanese BERT outperformed the clinical Japanese BERT and achieved a high correlation with human score (0.904 in the CR dataset and 0.875 in the EMR dataset). It was unexpected that the general Japanese BERT outperformed the clinical Japanese BERT on clinical domain dataset. This could be due to the fact that the general Japanese BERT is pretrained on a wide range of texts compared with the clinical Japanese BERT.

**Keywords**

- natural language processing
- semantic textual similarity
- clinical semantic textual similarity
- bidirectional encoder representations from transformers

**received**
February 2, 2021
**accepted after revision**
May 18, 2021
**published online**
July 8, 2021

DOI https://doi.org/10.1055/s-0041-1731390.
ISSN 0026-1270.



Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

## Introduction

Semantic textual similarity (STS) aims to compute the degree of semantic equivalence between texts based on the semantic content and meanings. It is common in many general English domain tasks such as text summarization, question answering, machine translation, information retrieval, dialog systems, plagiarism detection, and query ranking.[1] Although STS is similar to plagiarism detection, there are two major differences. First, plagiarism detection finds whether texts are similar, whereas STS finds the degree of similarity. Second, plagiarism detection uses texts from the internet for comparison, whereas STS uses texts depending on the research interests. STS is also related to paraphrase detection and textual entailment. There is a difference in that STS aims to capture the level of semantic equivalence, whereas paraphrase detection and textual entailment are a binary yes/no decision.[1,2] ►Fig. 1 shows an example of semantic textual similarity.

Due to its application across diverse tasks, many approaches to compute semantic similarity have been proposed. Existing approaches include corpus-based and knowledge-based models,[3] machine learning-based models,[4–7] neural networks-based models,[8–13] and BERT-based models.[11,14] Corpus-based method measure the degree of similarity between texts by using information exclusively extracted from a large corpus. Knowledge-based method measure the semantic similarity based on information extracted from semantic networks or structured resources like dictionaries, encyclopedias, thesauruses, Wikipedia, or WordNet.

Chen et al.[5] achieved the best performance in the 2018 clinical STS shared task.[15] Their proposed model employed traditional machine learning and deep learning. They trained a model with 63 features which included string-based, entity-based, number, and deep learning-based similarity features. Moreover, Zhao et al[7] used latent semantic analysis to learn vector-space representations, together with handcrafted features. Although traditional NLP approaches such as designing handcrafted features achieve good performance, they suffer from sparsity due to lack of large annotated data and language ambiguity.[10]

Mueller and Thyagarajan[9] proposed Siamese long short-term memory (LSTM) network for labeled data consisting of sentence pairs with variable length. Their approach relies on pretrained word-embeddings[16] and synonym augmentation. Further, Tai et al[13] proposed Tree-LSTMs which use syntactic trees to construct sentence representations. The standard LSTM model determines the hidden state from the current time-step input and previous time-step's hidden state. However, the Tree-LSTM model determines its hidden state from an input vector and the hidden states of all child units. The basic idea is that, by reflecting the sentence syntactic properties, the tree network can efficiently propagate more information than the standard sequential architecture.

Recently bidirectional encoder representations from transformers (BERT)[14] has achieved state-of-the-art performance in more than 10 NLP tasks. It is a popular approach for transfer learning and has been proven to be effective in achieving good accuracy for small datasets.[14,17] It can be used for tasks whose input is a sentence pair, such as sentence pair regression, question answering, and natural language inference. It learns distinctive embedding for the sentences so as to help the model in differentiating the sentences.

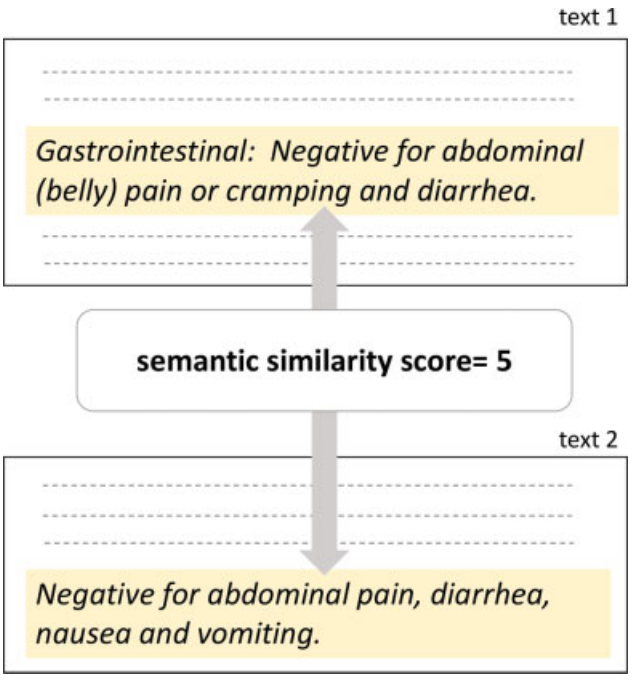


**Fig. 1** Semantic textual similarity example. Given a sentence-pair, a model computes semantic similarity score on a scale from 0 (low semantic similarity) to 5 (high semantic similarity).

SemEval (semantic evaluation) shared tasks have been held since 2012 to encourage the development of automated methods for STS tasks.[1,2,18–21] English STS has been widely studied with proposed state-of-the-art systems achieving high correlation (Pearson correlation score >80%) with human judgment.[2] However, these previous tasks focus on the general English domain. There exist very few resources for STS tasks in the clinical domain due to restricted access to clinical data because of patient privacy and confidentiality.[15,22] Wang et al[15,22] created an English clinical STS dataset from actual notes at Mayo clinic and organized shared tasks in 2018 and 2019. In their dataset, they removed all the protected health information, and the dataset can be accessed by signing a Data Use Agreement.

In this study we created two datasets for Japanese clinical STS: (1) Japanese case reports (CR dataset) and (2) Japanese electronic medical records (EMR dataset). As previously mentioned, the reason for few resources in the clinical domain is due to data privacy, which prohibits public sharing of medical data. To overcome this challenge, we created one dataset from a public resource, and made this dataset publicly available.[a] Specifically, the CR dataset was created by extracting case reports from CiNii,[b] a Japanese database

[a] The dataset is available at https://github.com/sociocom/Japanese-Clinical-STS.
[b] https://ci.nii.ac.jp/.

containing research publications in Japanese and English. Although research publications are different from real clinical texts, they have been widely used in various clinical natural language processing (NLP) researches to fill the gap for lack of publicly available real clinical texts. We also created a second dataset, the EMR dataset, from real clinical documents, however, this dataset is not publicly available.

Moreover, we used a BERT-based approach to capture the semantic similarity between texts. Recently many pretrained models, both general domain and domain-specific, have been developed. We investigate the performance of general and clinical domain pretrained Japanese BERT models on clinical domain datasets. Therefore, our contributions include:

1. Creating a publicly available dataset for Japanese sentence-level clinical STS from a public resource (CiNii) due to privacy issues associated with hospital clinical data.
2. Comparing the performance of the general and clinical Japanese BERT models.

## Methods

### Materials

This study used two Japanese datasets: case reports (CR dataset) and EMR documents (EMR dataset). We created the CR dataset from case reports which is publicly available.[c] By using the CR dataset, model performance can be measured with a publicly shareable dataset. In contrast, the EMR dataset was generated from medical documents and is not publicly available. The datasets consist of sentence pairs annotated on a scale from 0 (low semantic similarity) to 5 (high semantic similarity), where 0 means that the two-sentence pairs are completely dissimilar, i.e., their meanings do not overlap, and 5 means that the sentence pairs are completely similar semantically.

#### Japanese Case Reports (CR Dataset)

We created a publicly available dataset[c] to motivate research on Japanese clinical STS. There exist few resources for clinical STS tasks due to data privacy and confidentiality issues that prohibit public sharing of medical data. To overcome this challenge and create a publicly available dataset, we extract Japanese case reports from CiNii[d], which is a Japanese database containing articles published to Japanese journals and conferences. Japanese case reports were extracted from CiNii in PDF format (1,747 documents). The PDF documents were then converted to OCR format and split into sentences. Sentences that generally would not be found in real clinical documents such as references, author affiliations, and so on were removed.

After extracting all sentences, we created a dataset by using all possible combinations of sentence pairs. This resulted in a huge number of sentence pairs. Choosing sentence pairs randomly would have likely resulted in a dataset where the semantic similarity scores are highly imbalanced. Therefore, we adopted the approach used in previous tasks (SemEval[1–21] and MedSTS[22]). These previous studies use string similarity approaches to select sentence pairs for annotation. Although string similarity cannot entirely capture semantic similarity, they can capture some level of surface/syntactic similarity and hence significantly reduce the human effort required in selecting sentence pairs for annotation.

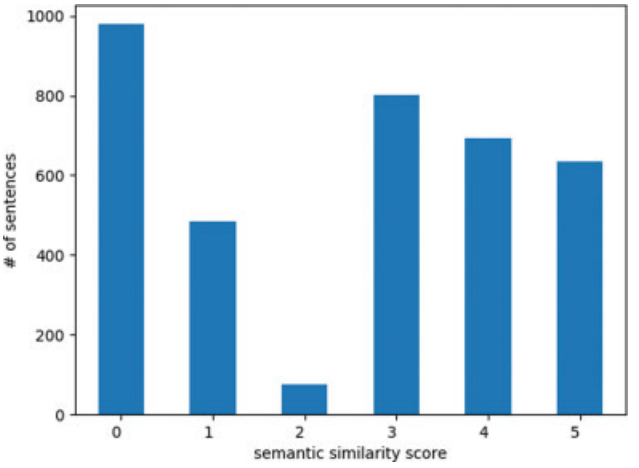


**Fig. 2** Distribution of semantic similarity scores in the Japanese case report dataset.

In this study, we used Python simstring library[e] to compute cosine similarity between the sentence pairs. Cosine similarity returns a score between 0 and 1. About 4,000 sentence pairs across all scores (0 to 1) were then selected for annotation by staff with medical background. The annotator assigned each sentence pair with a similarity score 0 (low semantic similarity) to 5 (high semantic similarity) depending on the semantic similarity. A second annotator annotated 10% of the data, and the annotators had a weighted Cohen Kappa agreement of 0.67, which can be regarded as acceptable for NLP tasks.[23] We used the same annotation guidelines as used in previous STS tasks[1,2,15,18–22] as shown in ►**Supplementary Appendix A**, available in the online version only. ►**Fig. 2** shows the distribution of semantic similarity scores in the CR dataset.

#### Japanese Electronic Medical Documents (EMR Dataset)

This dataset was created from actual Japanese medical documents consisting of radiography reports and electronic health record (EHR) notes. The EHR notes consist of progress notes. The radiography reports were provided by the National Cancer Center Japan, and the EHR notes were provided by Osaka University Hospital.[24] We filtered medical documents for patients with more than one entry and created document pairs based in chronological order as $[d_t, d_{t+1}]$. We asked the annotator to read sentences in $d_t$ and determine their semantic similarity with sentences in $d_{t+1}$. This dataset consists of approximately 2,000 sentence pairs annotated with semantic similarity scores from 0 to 5 similarly to the CR

[c] The dataset is available at https://github.com/sociocom/Japanese-Clinical-STS.
[d] https://ci.nii.ac.jp/.
[e] https://pypi.org/project/simstring-pure/

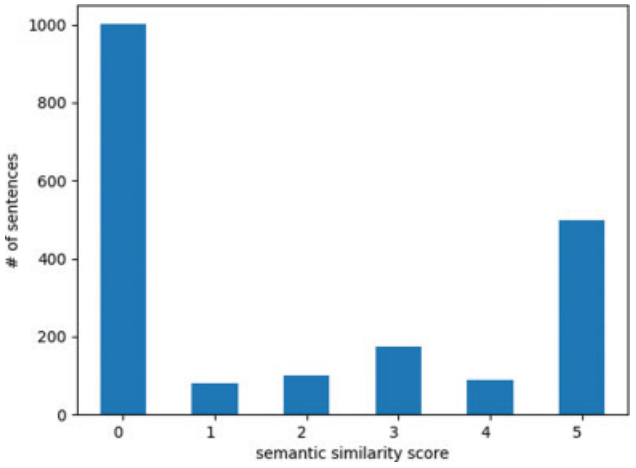


**Fig. 3** Distribution of semantic similarity scores in the EMR dataset. EMR, electronic medical record.

dataset. ►**Fig. 3** shows the distribution of the semantic similarity scores in the EMR dataset.

### Model

We adopted BERT[14] since it has been proven to be effective in achieving good accuracy for small datasets[14,17,25] like ours. Whereas data scarcity is one of the biggest challenges in NLP tasks, most NLP tasks require large amounts of training data so as to achieve reasonable accuracy. Dataset creation and annotation are expensive in terms of time and labor, and data are not available especially in the clinical domain. This challenge can be addressed by pretraining general domain language models using huge amounts of unlabeled data.[1] These pretrained models can be fine-tuned to specific tasks. It takes a lot of time to train the original BERT model from the beginning, and therefore training on a fine-tuned model reduces the time and memory usage.

Pretrained BERT models, both general domain and domain-specific, have been developed. General domain models are pretrained on cross-domain texts and therefore lack domain-related knowledge. Also, the linguistic characteristics of general domain texts and clinical domain texts are different hence creating the need for domain-specific BERT models.[26] In this study we investigate the performance of general Japanese BERT[f] and clinical Japanese BERT[27] models. The general Japanese BERT is pretrained on Japanese Wikipedia texts while the clinical Japanese BERT is pretrained on Japanese clinical texts (mainly notes by physicians and nurses) at University of Tokyo Hospital.[27]

The most common approach to use BERT is a feature-based approach where fine-tuning is not required, and instead the BERT vectors are used like word embeddings. The output of the BERT CLS token can also be used as a feature vector. The CLS (classification) token is a special BERT token added at the start of a sequence and represents the entire sequence.[14] Reimers and Gurevych[11] suggested that averaging the output of BERT or using the CLS token does not achieve good performance. They investigated different pooling methods for the BERT output such as mean and maximum pooling. However, the best strategy for extracting the feature vectors is still an open problem.

►**Fig. 4** shows the overview of our model. The input consists of a sequence of tokens of the two sentences concatenated by a special token, [SEP]. The input sequence also has the [SEP] token at the end to show the end of the input. The first token of the input sequence is the BERT special classification token, [CLS]. BERT encodes the sentence pair, and passes the final hidden state of the [CLS] as a representation of the input sequence. The output of the [CLS] token is passed to a fully connected linear output layer to calculate the semantic similarity score. The CR and EMR datasets are annotated on a discrete scale from 0 to 5 (i.e., 0, 1, 2, 3, 4, 5). We approached it as a classification problem and used a linear classifier with cross-entropy loss.[28]

## Results

### Experimental Settings

The CR and EMR datasets were split into 70% training set and 30% test set, respectively. Also, we prepared additional training set, n2c2, by translating the n2c2/OHNLP English dataset to Japanese using Googletrans, which is a python library that communicates with Google Translate API[g]. This n2c2/OHNLP dataset was provided in the 2019 n2c2/OHNLP Clinical Semantic Textual Similarity shared task, and is discussed in Wang et al.[15,22] In our experiments we do different combinations of the training data, to see how different datasets with different language variability affect the model performance.

We consider two experimental settings which we refer to as *strict* and *relaxed*. In the *strict* setting, we use the six scale semantic similarity scores (i.e., 0, 1, 2, 3, 4, 5) as discussed in the data annotation guidelines. In the *relaxed* setting, we consider a four scale where we combine scores 1 and 2, and 3 and 4, i.e., (0, [1, 2], [3, 4], 5). In the annotation guidelines (refer to ►**Supplementary Appendix A**, available in the online version only) the annotators stated that sometimes it was difficult to choose between semantic similarity scores 3 and 4. This is because, in some cases it is difficult to decide what constitutes "important" and "unimportant" information. Similarly, the same problem was experienced for semantic similarity scores 1 and 2. Therefore, we consider the relaxed setting for uniformity. We also expect that by using this kind of setting the classification performance can be improved.

### Performance of General and Clinical Japanese BERT Models

We evaluated the performance based on two evaluation metrics; the Pearson correlation coefficient (as in the previous STS shared tasks[1,2,18–22]) and classification accuracy between the predicted scores and gold scores. ►**Tables 1** and **2** show the results for the CR and EMR test sets, respectively. Both models, the general Japanese BERT and the clinical Japanese BERT, achieved a good performance. In

[f] https://github.com/cl-tohoku/bert-japanese

[g] https://pypi.org/project/googletrans/

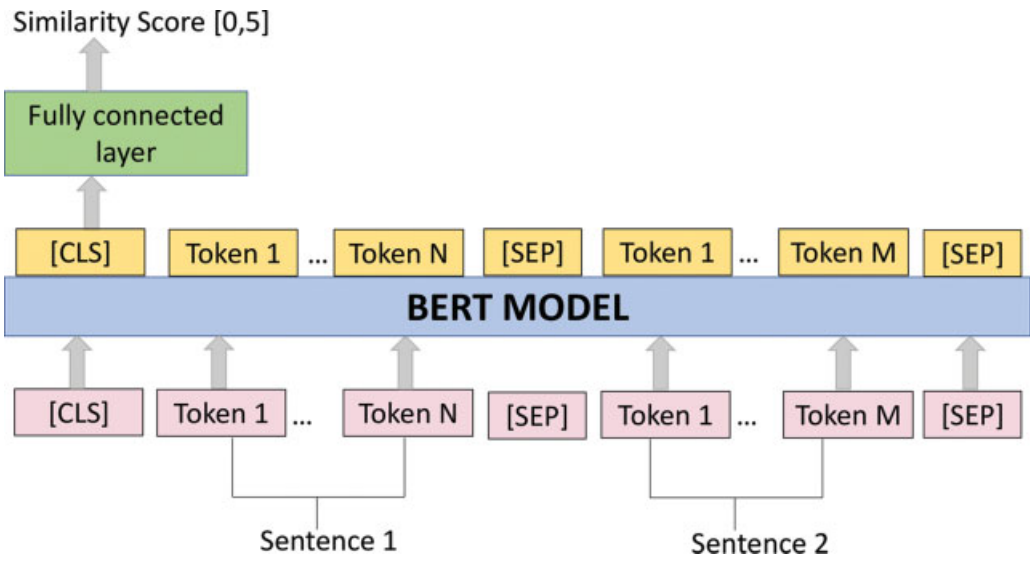


**Fig. 4** Overview of our model.

**Table 1** CR results based on Pearson correlation and classification accuracy

| | | Strict | | Relaxed | |
|---|---|---|---|---|---|
| **Model** | **Training data** | **Pearson** | **Accuracy** | **Pearson** | **Accuracy** |
| General Japanese BERT | CR | **0.904** | **72%** | 0.878 | **79%** |
| | EMR | 0.730 | 33% | 0.749 | 53% |
| | n2c2 | 0.716 | 35% | 0.705 | 55% |
| | CR + EMR | 0.897 | 71% | **0.882** | 78% |
| | CR + EMR+ n2c2 | 0.895 | 71% | 0.879 | 78% |
| Clinical Japanese BERT | CR | 0.890 | 67% | 0.854 | 75% |
| | EMR | 0.745 | 29% | 0.696 | 47% |
| | n2c2 | 0.656 | 25% | 0.613 | 39% |
| | CR + EMR | 0.885 | 68% | 0.862 | 76% |
| | CR + EMR+ n2c2 | 0.870 | 69% | 0.855 | 75% |

Abbreviations: BERT, bidirectional encoder representations from transformers; CR, case report; EMR, electronic medical record.

**Table 2** EMR results based on Pearson correlation and classification accuracy

| | | Strict | | Relaxed | |
|---|---|---|---|---|---|
| **Model** | **Training data** | **Pearson** | **Accuracy** | **Pearson** | **Accuracy** |
| General Japanese BERT | CR | 0.692 | 53% | 0.692 | 68% |
| | EMR | 0.864 | 79% | 0.860 | 84% |
| | n2c2 | 0.569 | 33% | 0.558 | 63% |
| | CR + EMR | 0.856 | 79% | 0.857 | 85% |
| | CR + EMR+ n2c2 | **0.875** | **81%** | **0.870** | **86%** |
| Clinical Japanese BERT | CR | 0.685 | 44% | 0.693 | 62% |
| | EMR | 0.845 | 76% | 0.824 | 82% |
| | n2c2 | 0.521 | 23% | 0.513 | 52% |
| | CR + EMR | 0.862 | 79% | 0.848 | 83% |
| | CR + EMR+ n2c2 | 0.848 | 78% | 0.833 | 82% |

Abbreviations: BERT, bidirectional encoder representations from transformers; CR, case report; EMR, electronic medical record.

the CR results, the general Japanese BERT achieved a Pearson correlation of 0.904 and 72% accuracy, whereas the clinical Japanese BERT best Pearson correlation and accuracy were 0.890 and 69%, respectively, in the strict setting. In the relaxed setting, the general Japanese BERT highest Pearson score and accuracy were 0.882 and 79%, respectively, while the clinical Japanese BERT achieved Pearson score of 0.862 and 75% accuracy.

In the EMR results, the general Japanese BERT best Pearson score and accuracy were 0.875 and 81%, respectively, while the clinical Japanese BERT achieved Pearson score of 0.862 and 79% accuracy in the strict setting. In the relaxed setting, the general Japanese BERT Pearson score and accuracy were 0.870 and 86%, respectively, whereas the clinical Japanese BERT were 0.848 and 83%, respectively. Although both BERT models performed well, in overall the general Japanese BERT model achieved the highest performance in both datasets.

## Discussion

### Effect of Training Data

In the CR results, training only on the CR dataset achieved the highest performance in the strict setting (Pearson correlation of 0.904 and accuracy of 72%). In the relaxed setting, training on the CR+ EMR achieved the best performance, Pearson correlation score of 0.882, but training on only the CR achieved the highest accuracy of 79%. We expected that training on more data would improve the performance, but training only on CR had best performance. Note that training only on n2c2 or EMR datasets achieved average performance in terms of Pearson correlation score (0.716 and 0.730 for the strict setting; 0.705 and 0.749 for the relaxed setting). Nevertheless, the classification accuracy is relatively low (35 and 33% for the strict setting; 55 and 53% for the relaxed setting). Although clinical Japanese BERT was trained on clinical texts, it achieved low performance in the CR test data. This could be attributed to the reason that case reports and real hospital text data are different in terms of vocabulary, abbreviations, linguistic patterns, and even sentence length.

In the EMR results, training on a combination of all the datasets achieved the highest performance (Pearson correlation of 0.875 and accuracy of 81% for the strict setting; Pearson correlation of 0.870 and accuracy of 86% in the relaxed setting). The EMR training set was small and therefore adding more data provided more training examples for our model hence improving the performance. Although both EMR and n2c2 datasets are created from real hospital documents, training on n2c2 dataset achieved the lowest performance (Pearson correlation of 0.569 and 33% accuracy for the strict setting; Pearson correlation 0.558 and 63% accuracy for the relaxed setting). This could be due to the reason that the n2c2 and our EMR datasets were created from different types of clinical notes. Our EMR dataset consisted of sentences from radiography notes and progress notes, while the n2c2 dataset consisted of sentences from other different types of clinical notes. Further, the n2c2 dataset was translated from English to Japanese using Google translate machine translation. The quality of machine translation was sufficient and most medical terms were translated efficiently. Although our preliminary manual check of the translated sentences looks sufficient, the performance of the proposed method could be improved by adopting better translation models. However, to compare the precise relation between the machine translation quality and STS performance is one of the future works.

In the EMR results, we expected the clinical Japanese BERT to achieve the best performance since it is trained on clinical texts, but the general Japanese BERT attained the highest performance. Although this could be surprising since domain specific pretraining is expected to perform better in general, the result suggests that semantic textual similarity relies more on fundamental linguistic features. This finding therefore encourages clinical applications based on semantic textual similarity, since widely available, general domain BERT models would work well. Moreover, the high performance of general Japanese BERT could also be due to the fact that it is trained on a wide range of texts and therefore it could generalize well.

### Error Analysis

►Tables 3 and 4 show error examples for the CR and EMR test sets, respectively. In the CR results, example (a) in ►Table 3 shows an example of abbreviation expansion problem. Abbreviation expansion is a major problem even in other NLP tasks, and in future there is a need for a precise method to handle this problem. Example (b) is a case of language variability, although the sentences are similar in meaning the choice of words varies greatly. In example (c), the model assigned a higher score to the sentence pair because actually the sentences are highly similar and have only a minor difference ("Yamada type 1 or type II" in the first sentence, and "Yamada type III" in the second sentence). Although this kind of difference is important in the clinical domain, in the general English domain this sentence pair can be treated as semantically equal. In example (d), the sentences are roughly equivalent, and although our model assigned a score of 4, the gold score should be 3.

In the EMR results, the sentence length varies greatly from very short sentences (1 or 2 words). Example (a) in ►Table 4 shows a typical example of short sentences found in EMR notes. The EMR dataset sentence lengths have a large difference, and our model was not able to correctly classify sentence pairs with very short sentences. In sentence pairs of examples (b) and (c), our model assigned a lower score because although the sentences have a high semantic similarity, the choice of words is quite different. For example in (c), "almost no change" and "slightly decreased" have close meaning semantically. It is easy for human beings to capture this kind of meaning but difficult for machines to capture this kind of similarity. Sentence pairs of examples (d) and (e) show a case of positive–negative relationship. Our model was not able to capture negation, and in future it is necessary to train our model to identify this kind of relationship.

**Table 3** Error analysis in the CR dataset

| | Examples | Gold score | System score |
|---|---|---|---|
| a | S1: 23 mmのSJM弁にてAVRを施行し,術後13日に軽快退院となった<br>AVR was performed using a 23 mm SJM valve, and she was discharged 13 d after surgery.<br>S2: 25 mmのSJM弁にて人動脈弁置換術(以下AVR)を施行し,術後12日に軽快退院となった<br>Human aortic valve replacement (AVR) was performed using a 25 mm SJM valve, and she was discharged 12 d after surgery. | 5 | 3 |
| b | S1: 術後化学療法は年齢と患者および家族の希望を考慮して施行しなかった<br>Postoperative chemotherapy was not given considering age and patient and family preferences.<br>S2: (5)化学療法は本腫瘍に対して効果が不確実であるため,施行しなかった<br>(5) Chemotherapy was not performed because of uncertain effect on this tumor. | 3 | 1 |
| c | S1: 1群;腫瘤表面は青白色調から赤紫色調を呈し,その形態が山田の1またはII型<br>Group 1; the surface of the tumor mass is bluish-white to magenta, and its morphology is Yamada type 1 or type II.<br>S2→121(1849)2群;腫瘤表面は青白色調から赤紫色調を呈し,その形態が山田のIII型<br>121(1849) 2 group; The surface of the tumor mass is bluish-white to magenta, and its morphology is Yamada type III. | 3 | 5 |
| d | S1: 術後3か月でほぼ骨架橋は完成し,骨梁構造も周囲とほぼ同じであった<br>Three months after the operation, the bone bridge was almost completed, and the trabecular structure was almost the same as the surrounding area.<br>S2: X線写真上では,骨架橋形成は術後3か月でほぼ完了し,移植骨部は周囲と同じ<br>On the radiograph, bone bridge formation was almost completed 3 mo after the operation, and the bone graft had the same trabecular structure as the surrounding area. | 3 | 4 |

## Conclusion

STS tasks have been widely studied especially in the general English domain. However, only a few resources exist for STS tasks in the clinical domain and languages other than English such as Japanese. To bridge this gap, we created a publicly available dataset for Japanese clinical STS. The dataset consists of approximately 4,000 sentence pairs extracted from Japanese

**Table 4** Error analysis in the EMR dataset

| | Examples | Gold score | System score |
|---|---|---|---|
| a. | S1: BForder<br>S2: 2014/5/26BF | 4 | 1 |
| b | S1: 左副腎の腫大を認めますが前回とサイズに著変を認めません。<br>Swelling of the left adrenal gland is observed, but there is no significant change in size from the previous visit.<br>S2: 左副腎に認められる低吸収結節のサイズや性状は前回よりほぼ変化ありません。<br>The size and properties of the hypodense nodule found in the left adrenal gland are almost unchanged from the previous visit. | 5 | 3 |
| c | S1: 前回CT(12/03/30)と比較すると陰影は軽度軽快しています。<br>Compared with the previous CT (December 30, 2012), the shadows slightly decreased.<br>S2: 前回(December 25, 2013)と概ね変化を認めません。<br>There is almost no change from the previous visit (December 25, 2013). | 3 | 1 |
| d | S1: 腹水なし<br>No ascites.<br>S2: 腹水あり。<br>There is ascites. | 1 | 3 |
| e | S1: 胸水貯留を認めません。<br>Pleural effusion is not recognized.<br>S2: 両側胸水、心嚢液貯留を認めます。<br>Bilateral pleural effusion and pericardial effusion are present. | 1 | 3 |

Abbreviation: EMR, electronic medical record.

case reports annotated with a semantic similarity score from 0 (low semantic similarity) to 5 (high semantic similarity).

We used a BERT-based approach to capture semantic similarity between clinical domain texts. In our experiments we achieved a high Pearson correlation score between the gold scores and human scores (0.904 in the CR dataset; 0.875 in the EMR dataset). In this study we also compared the performance of the general and clinical Japanese BERT models. Although both models achieved a good performance, the general Japanese BERT achieved the highest performance compared with the clinical Japanese BERT in our clinical domain datasets. Though this could be surprising because domain specific pretraining is known to perform better in general, the results suggest that semantic textual similarity relies more on fundamental linguistic features. This finding particularly encourages clinical applications based on semantic textual similarity, since widely available general domain BERT models would work well.

## Funding

This work was supported by a Japan Science and Technology Agency PRISM Grant (Grant No. JPMJCR18Y1).

## Conflict of Interest

None declared.

## References


1 Agirre E, Banea C, Cer D, et al. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. Paper presented at: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics; 2016

2 Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. Semeval-2017 task 1: semantic textual similarity—multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055;2017

3 Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. AAAI 2006; 6:775–780

4 Šarić F, Glavaš G, Karan M, Šnajder J, Dalbelo Bašić B TakeLab: Systems for measuring semantic text similarity. The First Joint Conference on Lexical and Computational Semantics—Volume 1. Paper presented at: Proceedings of the Main Conference and the Shared Task, and Volume 2. Paper presented at: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). Association for Computational Linguistics; 2012:441–448

5 Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Combining rich features and deep learning for finding similar sentences in electronic medical records. Paper resented at: Proceedings of the BioCreative/OHNLP Challenge 2018:5–8

6 Tian J, Zhou Z, Lan M, Wu Y. ECNU at Semeval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. Paper presented at: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics; 2017:191–197

7 Zhao J, Zhu TT, Lan M. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. Paper presented at: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin-Association for Computational Linguistics; 2014:271–277

8 Kiros R, Zhu Y, Salakhutdinov RR, et al. Skip-thought vectors. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, eds. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015:3294–3302

9 Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Paper presented at: Thirtieth AAAI Conference on Artificial Intelligence ; 2016

10 He H, Lin J. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. Paper presented at: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2016:937–948

11 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084;2019

12 He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks. Paper presented at: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2015:1576–1586

13 Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075;2015

14 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805;2018

15 Wang Y, Afzal N, Liu S, et al. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. Paper presented at: Proceedings of the BioCreative/OHNLP Challenge 2018

16 Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. Advances in Neural Information Processing Systems 26. Curran Associates, Inc.; 2013:3111–3119

17 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Accessed December 2, 2020 at: https://s3-us-west-2amazonaws.-com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

18 Agirre E, Diab M, Cer D, Gonzalez-Agirre A. Semeval-2012 task 6: A pilot on semantic textual similarity. Paper presented at: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1. Paper presented at: Proceedings of the Main Conference and the Shared Task, and Volume 2. Paper presented at: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics; 2012:385–393

19 Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. SEM 2013 shared task: Semantic textual similarity. Second Joint Conference on Lexical and Computational Semantics, Volume 1. Paper presented at: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Vol 1. 2013:32–43

20 Agirre E, Banea C, Cardie C, et al. Semeval-2014 task 10: Multilingual semantic textual similarity. Paper presented at: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics; 2014:81–91

21 Agirre E, Banea C, Cardie C, et al. Semeval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. Paper presented at: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics; 2015:252–263

22 Wang Y, Afzal N, Fu S, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 2020; 54:57–72

23 Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008;34(04):555–596

24 Yada S, Joh A, Tanaka R, Cheng F, Aramaki E, Kurohashi S. Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: starting from critical lung diseases. Paper presented at: Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association; 2020:4565–4572

25 Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365;2018

26 Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323;2019

27 Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed with huge size of Japanese clinical narrative. medRxiv 2020

28 Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. Paper presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2019: 4487–4496