Detecting Redundancy in Electronic Medical Records Using Clinical BERT

Faith W. Mutinda, Sumaila Nigo, Daisaku Shibata
Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki

Nara Institute of Science and Technology

{mutinda.faith_wavinya.mz2, sumaila.nigo.sl8, shibata.daisaku.rr8,
s-yada, wakamiya, aramaki}@is.naist.jp

Abstract

Semantic Textual Similarity (STS) computes the degree of semantic equivalence between text snippets. STS is used in various natural language processing tasks, including detecting redundant information in texts. Although STS tasks have been widely studied in the general English domain, there exists very few resources for STS tasks in the clinical domain. The methods used for general domain STS tasks might not work well in the clinical domain because of variability of natural language expressions, and clinical domain expressions are different from general domain expressions. We present a Clinical BERT-based model for redundancy detection in clinical texts. Our experiments show that domain-specific BERT improves performance of the model.

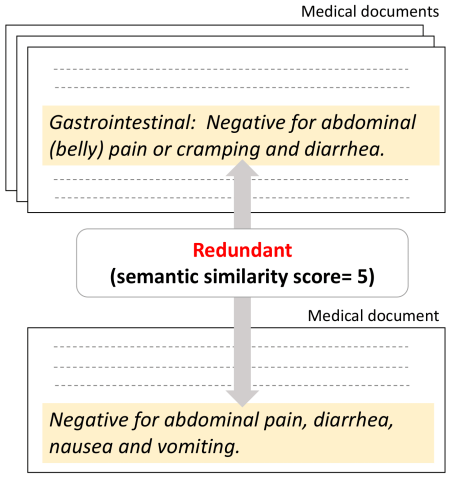

Figure 1: Redundant information detection. Given a set of medical documents, compute the semantic similarity score and find highly similar sentences.

1 Introduction

Electronic Health Records (EHRs) have been widely adopted to record patient's medical progress. EHRs have improved clinical documentation and decision support, because they provide a coordinated, quick, and efficient access to patient records. However, this comes with challenges such as copy-and-paste, use of templates, and smart phrases [23]. An analysis of 23,630 notes written by 460 clinicians found that 18% were manually entered; 46% were copied; and 36% were imported [22]. Redundancy reduces the quality of the EHR data and makes it difficult to extract relevant information for decision making [24]. Therefore, there is need to minimize redundancy so as to improve the quality of collected EHR data and make clinical decision making easier and efficient.

One method for detecting redundant information is to compute the degree of semantic equivalence between clinical texts to remove texts which are highly equivalent [23]. STS is a common task in general English domain and natural language processing (NLP) tasks, including text summarization, question answering, machine translation, information retrieval, dialog systems, plagiarism detection, and query ranking [3]. Although redundant information detection is similar to plagiarism detection, plagiarism detection finds whether texts are similar, whereas semantic similarity finds the degree of the similarity.

SemEval (Semantic Evaluation) shared tasks have been held since 2012 to encourage the development of automated methods for the STS task [1–4, 6]. However, these tasks focus on the general English domain and there exists very few resources for STS tasks in the clinical domain, ClinicalSTS [23]. Measuring semantic textual similarity is a challenging task due to variability of natural language expressions and limited annotated data.

This study focuses on the problem of detecting redundancy in clinical texts. Figure 1 shows redundant information detection using semantic textual simi-

larity. We solve a Clinical STS task, i.e., given a pair of two sentences, the objective is to compute their degree of semantic similarity on a scale [0, 5]. Zero means that the two sentences are completely dissimilar, i.e., their meanings do not overlap and 5 means that the sentence pairs are completely similar, semantically. We adopt a BERT-based model. BERT [8] set state-of-the-art performance in various NLP tasks [8, 15, 17]. In the experiments, two BERT models are used, i.e., BERT and Clinical BERT [5]. BERT is trained on general domain texts, while the Clinical BERT is trained on clinical notes. The results show that domain-specific BERT, i.e., Clinical BERT, improved the performance.

2 Related work

Due to its application across diverse tasks, many approaches to compute semantic similarity have been proposed. The existing approaches include; corpus-based and knowledge-based models [12], machine learning-based models [7, 19, 21, 26], and neural networks-based models [9–11, 14, 16, 18, 20].

Mihalcea et al. [12] proposed a method which uses corpus-based and six knowledge-based measures for semantic textual similarity. The corpus-based method measures the degree of similarity between texts by using information exclusively extracted from a large corpus while the knowledge-based method measures the semantic similarity based on information extracted from semantic networks.

Chen et al. [7] achieved the best performance in the ClinicalSTS shared task [23]. They proposed a method which employs traditional machine learning and deep learning. Similarly, [21] combined traditional NLP methods with deep learning. Zhao et al. [26] used latent semantic analysis to learn vector-space representations, together with handcrafted features. However, traditional NLP approaches, such as designing handcrafted features, suffer from sparsity due to lack of large annotated data and language ambiguity [10].

Kiros et al. [11] proposed skip-thoughts model, which extends word2vec skip-gram model from word level to sentence level. They train an encoder-decoder architecture to predict surrounding sentences. Prijatel et al. [16] proposed a model that uses various Long Short-Term Memory (LSTM) models with pre-trained word vectors and sentence embeddings. Mueller et al. [14] proposed Siamese LSTM network for labelled data consisting of sentence pairs with variable length. Their approach relies on pre-trained word-embeddings [13] and synonym augmentation.

Tai et al. [20] 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.

BERT [8] provides pretrained models which can be fine-tuned to produce state-of-the-art results in various NLP tasks [8, 15, 17]. BERT 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.

3 Material and Methods

3.1 Dataset

We use two datasets: n2c2/OHNLP dataset and STS-B dataset. The datasets consist of sentence pairs annotated on a scale [0,5], 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. The sizes of the datasets are as shown in Table 1.

3.1.1 n2c2/OHNLP dataset

This dataset was provided in the 2019 n2c2/OHNLP Clinical Semantic Textual Similarity shared task[1] and consists of 1642 sentence pairs. The sentences are derived from clinical notes obtained from Mayo Clinic's clinical data warehouse [23]. Each sentence pair was annotated by medical experts. The agreement between the annotators had a weighted Cohen's Kappa of 0.67.

3.1.2 STS-B dataset

This dataset comprises English data used in the SemEval STS tasks [6]. It consists of general domain English sentences derived from user forums, image captions, and news headlines. Similar to the n2c2/OHNLP dataset, the STS-B dataset is scored by human annotators. This dataset is used for fine-tuning of our model.

[1] https://n2c2.dbmi.hms.harvard.edu/track1

Table 1: Datasets

Dataset	Train	Development	Test
n2c2/OHNLP	1314	328	410
STS-B	5749	1500	1379

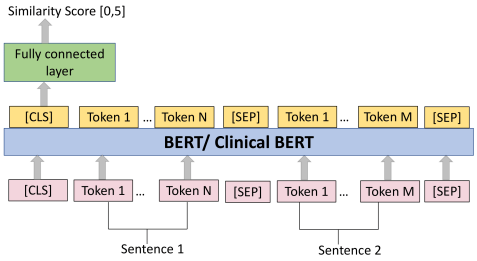

Figure 2: Overview of our model

3.2 Method

We adopt BERT [8], one of the state-of-the-art models in NLP tasks. BERT is a popular approach for transfer learning and has been proven to be effective in achieving good accuracy for small datasets [8, 15, 17]. In transfer learning, model weights are learned from a large dataset, and fine-tuned for the target task. In addition to increased precision and accuracy, transfer learning also reduces the computing time and memory usage.

Figure 2 shows the overview of our model. Under sentence pair tasks, the input for BERT consists of the tokens of the two sentences, separated by a special token, [SEP]. The input sequence also has the [SEP] token at the end. The first token of the input sequence is the BERT special classification token, [CLS]. We use BERT to encode the sentence pair, and pass the final hidden state of the [CLS] token to a fully connected linear layer to obtain the similarity score. Since the n2c2/OHNLP dataset is small (1642 sentence pairs), we increase the training instances by adding the STS-B dataset. During fine-tuning, the training set contains the n2c2/OHNLP training set and the STS-B dataset. In the experiments, we use two BERT pre-trained models: BERT [8] and Clinical BERT [5]. BERT is trained on general domain texts, whereas the Clinical BERT is trained on clinical notes.

4 Results and Discussion

We evaluate the model performance based on the Pearson correlation score between the predicted scores and gold scores on the n2c2/OHNLP test dataset. We also provided scores for a baseline model. The baseline model uses traditional NLP features and word embeddings. The NLP features were extracted manually and the semantic textual similarity score computed using an ensemble of five regression models, i.e., Random Forest, AdaBoost, Gradient Boosting, XGBoost, and CatBoost. The NLP features include different types of string similarity measures such as N-gram overlaps [19], word embeddings (Google[2] and PubMed[3]), machine translation metrics [25], token-based string similarity, and sequence-based string similarity [7].

Table 2: Test set Pearson correlation

Model	Test Correl.
Baseline	0.7804
BERT	0.6923
Clinical BERT	**0.8320**

Table 2 shows the results of the models. The BERT model achieved a Pearson correlation score of 0.6923, whereas the Clinical BERT model achieved a Pearson correlation score of 0.8320 on the n2c2/OHNLP test dataset. Note that the baseline model outperformed the BERT model. As expected, the Clinical BERT model achieved the best performance since it is trained on clinical texts. These results show that using domain-specific model improves the performance.

Table 3 shows some of the challenges experienced in the clinical STS task. The system output for the *aspirin-carvedilol* and *melatonin-eliquis* sentence pairs is similar. These sentence pairs have many similar words, and the only major difference is the drug names. The model also assigned a high score to the *ibuprofen-ibuprofen* sentence pair. In this sentence pair, the major differences are the drug dosage, strength, and frequency. It is difficult for the model to automatically understand these differences, and even humans without medical knowledge cannot correctly score the sentences. Therefore, there is need to use extra explicit medical knowledge to appropriately model and capture such differences.

5 Conclusion

This paper introduced a Clinical BERT-based model for detecting redundancy in electronic medical records. Although STS tasks have been widely studied in the general domain, there exists few resources in the clinical context. Our experiments showed that using domain-specific BERT, i.e., Clinical BERT, improved the performance. One limitation of this study

[2]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[3]https://github.com/ncbi-nlp/BioSentVec

Table 3: Challenges of clinical STS task

Examples	Gold score	System output
Sentence 1: Aspirin [BAYER] 81 mg tablet enteric coated 1 tablet by mouth one time daily. Sentence 2: Carvedilol [COREG] 25 mg tablet 1 tablet by mouth two times a day.	4	3.47
Sentence 1: Melatonin 3 mg tablet 1-2 tablets by mouth every bedtime as needed. Sentence 2: Eliquis 5 mg tablet 1 tablet by mouth two times a day.	2.5	3.49
Sentence 1: ibuprofen [MOTRIN] 600 mg tablet 1 tablet by mouth every 6 hours as needed. Sentence 2: ibuprofen [ADVIL] 200 mg tablet 2-3 tablets by mouth every 4 hours as needed.	0.5	2.63

is that our models compute the semantic similarity on a sentence level. We plan to extend this study to build a practical application for clinicians to perform semantic similarity on whole documents.

References

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proc. of SemEval-2015*, pages 252–263, 2015.

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proc. of SemEval-2014*, pages 81–91, 2014.

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proc. of SemEval-2016*, pages 497–511, 2016.

[4] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proc. of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, 2013.

[5] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[6] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proc. of SemEval-2017*, pages 1–14, 2017.

[7] Qingyu Chen, Jingcheng Du, Sun Kim, W John Wilbur, and Zhiyong Lu. Combining rich features and deep learning for finding similar sentences in electronic medical records. *BioCreative/OHNLP Challenge*, pages 5–8, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL'19*, volume 1, pages 4171–4186, 2019.

[9] Hua He, Kevin Gimpel, and Jimmy Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proc. of EMNLP'15*, pages 1576–1586, 2015.

[10] Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proc. of NAACL'16*, pages 937–948, 2016.

[11] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[12] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. of AAAI'06*, volume 1, pages 775–780, 2006.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS'13*, volume 2, pages 3111–3119, 2013.

[14] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *Proc. of AAAI'16*, page 2786–2792, 2016.

[15] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[16] Derek Prijatelj, Jonathan Ventura, and Jugal Kalita. Neural networks for semantic textual similarity. In *Proc. of ICON'17*, 2017.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Preprint*, 2018.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[19] Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. Takelab: Systems for measuring semantic text similarity. In *Proc. of *SEM'12*, pages 441–448, 2012.

[20] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proc. of ACL'15 and JCNLP'15 (Volume 1: Long Papers)*, pages 1556–1566, 2015.

[21] Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 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. In *Proc. of SemEval-2017*, pages 191–197, 2017.

[22] Michael D Wang, Raman Khanna, and Nader Najafi. Characterizing the source of text in electronic health record progress notes. *JAMA internal medicine*, 177(8):1212–1213, 2017.

[23] Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. Overview of the biocreative/ohnlp challenge 2018 task 2: Clinical semantic textual similarity. *BioCreative/OHNLP Challenge*, 2018, 2018.

[24] Rui Zhang, Serguei Pakhomov, Bridget T McInnes, and Genevieve B Melton. Evaluating measures of redundancy in clinical texts. In *AMIA annual symposium proceedings*, volume 2011, page 1612, 2011.

[25] Jiang Zhao, Man Lan, and Jun Feng Tian. ECNU: using traditional similarity measurements and word embedding for semantic textual similarity estimation. In *Proc. of SemEval-2015*, pages 117–122, 2015.

[26] Jiang Zhao, Tiantian Zhu, and Man Lan. ECNU: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. In *Proc. of SemEval-2014*, pages 271–277, 2014.