MLM-based Typographical Error Correction of Unstructured Medical Texts for Named Entity Recognition

Eun Byul Lee
Yonsei University
Go Eun Heo
Yonsei University
Chang Min Choi
University of Ulsan College of Medicine
Min Song (min.song@yonsei.ac.kr)
Yonsei University

Research Article

Keywords: Bioinformatics, Named Entity Recognition, Language Model, Artificial Neural Network

Posted Date: February 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1349382/v1

Abstract

Background

Unstructured texts in the medical fields such as Electronic Health Records contain an enormous amount of valuable information for research, but it is difficult to extract and structure important information because of frequent typos. For text analysis, improving the quality of data with errors is an essential task, but just a few prior studies have been conducted so far. Therefore, we propose a new methodology for extracting important information from unstructured medical texts by overcoming the typographical problem in the surgical pathology records related to lung cancer in the medical field.

Methods

In order to solve the problem of typographical errors occurring in real-world medical data, we propose a typo correction model considering context based on the Masked Language Model. In addition, a word dictionary to be used for typo correction model was constructed based on PubMed abstracts. After refining the data through typo correction, fine tuning was performed on the pre-trained BERT, and deep learning based Named Entity Recognition was performed. By solving the quality problem of medical data, we strive to improve the accuracy of information extraction in unstructured text data.

Results

We compared the performance of the proposed typo correction model based on contextual information with the existing SymSpell, and confirmed the proposed one outperformed the existing model in the typographical correction task. The F1-score of the model improved about 5% and 9% than without the contextual information in the NCBI-disease and Surgical pathology record datasets, respectively. In addition, for the evaluation of the NER task, the F1-score of NER after typo correction increased by 2% in the NCBI-disease dataset. There was a significant performance difference about 25% between before and after typo correction in the Surgical pathology record dataset. It was confirmed that typos had a great influence on information extraction in the unstructured text.

Conclusion

We verified typos in text negatively affect the performance of natural language processing tasks and the proposed method outperforms the existing one. This study has implications proposed robust model can be applied to the real-world environment by focusing on the typo problem causes difficulties in analyzing unstructured medical text.

Background

With the improvement of hardware and software performance and the technological development of data collection and analysis methods, data have been digitized, and an interest in 'big data' is increasing due to the various uses of data. In particular, unstructured data such as text, image, video, and audio account

for an overwhelming proportion. As a large amount of data is accumulated, many studies related to text mining that extract and utilize meaningful information from a large amount of text data have been steadily conducted [1–8].

The digitization of data has led to major changes in many fields. In particular, the Electronic Health Records (EHRs), which collect and manage patient health information in the form of electronic documents, unlike the paper charts in the past is integrated with the hospital system and make it possible to utilize data in the medical fields [9]. EHRs include information such as diagnostic findings, past medical history, image reports, and surgical results sheets. EHRs contain a large amount of data and can be used as real-world evidence through appropriate analysis methods [10]. In addition to EHRs, PubMed and MEDLINE, which are premier bibliographic database of biomedical literature, provide more than 27 million of bibliographic information of biomedical journals such as medicine, nursing, and health medicine, and as such, text data in the medical fields is increasing exponentially.

In the case of unstructured text data, it is difficult to completely prevent errors in the typing process, although it contains a lot of meaningful information that can be studied. In particular, according to a study by Hersh et al. [11] and Zhou et al. [12], typos occur more frequently in texts written quickly such as EHR. Such typos are a common problem in clinical text data such as EHR, constituting about 5 ~ 17% of the total data. Typos appearing in these clinical data affect the performance of natural language processing (NLP) tasks in the medical field, such as part-of-speech tagging (POS), information extraction, and information retrieval. In particular, it was found that even a small amount of typos had a negative effect on the information retrieval task [13].

In the medical field, accuracy of information is an important issue that is directly related to patient safety as well as efficient communication. As a related example, an error in the breast imaging report resulted in the exchange of incorrect information and affected the patient's treatment [14]. Also, drugs with similar names cause confusion and lead to erroneous results in drug prescription [15]. While inaccurate information is a major problem that can lead to medical accidents, few studies have been conducted to improve the data quality of unstructured text data in the medical field [16]. Therefore, this study focused on the problems that hinder the accuracy of information and the performance of text-based analysis in the medical field. The present study helps prevent medical accidents related to patient safety directly and increasing the utilization of unstructured data by solving the problem of low-quality data due to frequently represented typos.

Named Entity Recognition (NER) is a major natural language processing task in the medical field and refers to the extraction of meaningful information from descriptively written unstructured text. A typo affects the extraction of important information from clinical data, and handling typos is a necessary task before extracting information through NER [17]. In this study, we attempted to solve the problem of inaccurate data appearing in the real-world medical environment and to identify how the improvement of the quality of such data affects the performance of the natural language processing task for unstructured

text. In particular, we evaluated the model and verified whether it is applicable to real data by extracting key information that is the basis for diagnosis from data in the real medical environment.

The rest of the paper is organized as follows. We describe the proposed method in the Methods section. We evaluate performance of typo correction with context and NER through data quality improvement in the Experimental result section. Finally, we conclude the research and suggest future works in Conclusions.

Related Work

Typo Correction in the medical field

Types of typos occurring in text are mainly divided into two types: 'Non-word typo error' and 'Context-sensitive typo error'. Since non-word typo error can be corrected in a simpler way than correcting context-sensitive typo error, this study focuses on context-sensitive typo error.

Context-sensitive typo error is complex task because it needs to be corrected considering the relationship between the word to be corrected and the surrounding word. Context-sensitive typo error can be classified into four types: homophone error, typographical error, grammatical error, and spacing error [18]. An example of each error is shown in Table 1. In order to solve context-sensitive typo error, studies on a frequency-based method, statistical-based method, and deep learning-based method have been conducted. The frequency-based method is a method of typo correction with word that is most likely to appear in a sentence based on how often a word appears within a sentence. statistical-based method can correct typos based on context and furthermore, through deep learning-based typo correction studies, typos have been corrected by understanding the context more deeply.

Table 1
type of errors and examples in context (Lee et al., 2020)

Error type	Cause of error	Example
Homophone error	words that sound the same but are spelled differently	peace/piece
Typographical error	striking an incorrect key on a keyboard	from/form
Grammatical error	the user did not know exatly what the difference between grammers	among/between
Spacing error	wrong blank between words	maybe/may be

In the medical fields, studies on typo correction have been continuously conducted for document accuracy. Senger et al. [19] conducted study to correct typos that occur when searching for drug through electronic drug information systems in which drug information is digitized. By applying Aspell based on the Metaphone and Double Metaphone algorithms, candidate words were created, and typos were corrected by ordering them based on the edit distance. The system that did not apply typo correction

missed the search results about 17.5%, but the system optimized with search auto-correction applying a typo correction algorithm reduced noise when searching for drugs, thereby reducing the search time delay.

The National Library of Medicine (NLM) has a system to receive health questions from users. Kilicoglu et al. [20] applied typo correction algorithm based on a phonetic algorithm and an edit distance algorithm to consumer health questions collected from NLM. A group of candidate words for typo correction was generated based on dictionary search, and word similarity, pronunciation similarity, and similarity calculating the number of matching words at the beginning and the end of a word were used.

Workman et al. [21] applied a typo correction algorithm to the surgical pathology reports, and emergency department progress and visit notes. To remove unnecessary words in the document, SPECIALIST Lexicon which is a large-scale biomedical vocabulary corpus was used and Word2Vec, word embedding model, was used to embed the sentences. In addition, typos were corrected by comparing the group of candidate words and typos by using the Levenshtein edit distance.

Language Model and NER

Language model can be divided into statistical-based methods and methods using artificial neural networks. The statistical-based language model is finding the word with the highest probability of appearing at a specific position based on previous words in the sentence through conditional probability. Since the probability is obtained from the corpus of a specific domain, it has a dependency on a specific domain, and there is a problem of long- term dependency as the length of the sentence increase. In addition, there is a problem of scarcity because a word that does not exist in the corpus has a probability of 0 even though it is an appropriate word in the given sentence. Therefore, in this case, the performance of the language model differs depending on the size of the corpus and how many different words the corpus contains.

Recently, in the field of natural language processing, transformer-based language models [22] such as BERT (Bi-directional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) using artificial neural networks are showing the best performance. Since many texts are pre-trained in these models, they are domain-independent and can compensate for the scarcity problem that appears in existing language models. In addition, they are contextual language model that learns contextual information within a sentence.

The BERT process is divided into two tasks: pre-training and fine-tuning [23]. In the pre-training process, language modeling is conducted, and in the fine-tuning process, additional natural language is learned for pre-trained model. Through this, BERT showed significantly better performance than the existing 11 natural language processing tasks.

Unlike existing word embedding models such as Word2vec [24], Fasttext [25], and Glove [26], BERT embeds words by considering contextual information. In the case of the existing embedding model, each word has a fixed vector value. In this case, information about homonym which is words that are the same

word but have different meanings depending on the context cannot be considered. BERT expresses a sentence as the sum of three embeddings: token embedding, segment embedding, and position embedding to reflect the contextual characteristics of a word. Token embedding uses the Word Piece embedding method to treat words that appear frequently in the document as a single word unit and divide words that rarely appear into sub-words. In an embedding model such as Word2Vec with a fixed vector, an out-of-vocabulary (OOV) problem occurs because a specific word does not exist in the word set. In the case of BERT, the OOV problem can be effectively dealt with by dividing rarely used words in the document into sub-words. Segment embedding separates sentences, and position embedding embeds positional information of each word.

Masked Language Model (MLM) is similar to the learning method of Word2Vec's CBOW model by covering some words in a sentence and predicting the hidden words based on context information. $P\left(s_i \middle| s_1, s_2, \ldots, s_{i-1}, [MASK], s_{i+1}, \ldots s_t\right)$ is obtained when the i-th of a sentence S consisting of number of words t is masked. In general, previous language models learn unidirectional context and predict entire words, whereas MLM learns context information bidirectionally from left to right and right to left and predicts only the masked part of words. The [MASK] token for masking words is used during pre-training to learn contextual information in sentences while matching the correct answer to the masked part. These [MASK] tokens are not used during fine tuning, which creates a gap between the pre-training process and the fine-tuning process. To reduce this gap, when randomly masking 15% of the total words, 80% of them are covered with [MASK] tokens, 10% are randomly replaced with other words, and rest 10% of the words are kept as they are.

Named entity recognition means a word or phrase with a specific meaning in the corpus, and the type of entity name is different according to domains. For example, country, organization, and person are defined as entity names and the entity names are extracted from within the corpus. NER is used for preprocessing of major natural language processing task such as chatbots and information retrieval. Unlike the general classification problem that outputs one value, it is a sequence labeling problem that receives an input sequence and outputs a sequence of the same length as the input sequence. The BIO format is used to recognize the object name, which is an abbreviation of Begin, Inside, and Outside. An example is shown in Fig. 1.

NER has been studied for text-based information extraction in various domains. In the medical fields, many studies have been conducted to recognize individual names such as diseases, drugs, and DNA and to extract information. An artificial neural network-based LSTM-CRF model has shown good performance, and recently, a pre-trained BERT-based model has shown the highest performance in NER.

Researches in clinical NER have been applied from machine learning approaches to deep learning models. An active learning (AL) algorithm to minimize the annotation process was proposed and the performance of F1-measure was 0.8 [27]. Recently, studies on neural word embeddings in the unlabeled clinical NER corpus [28], studies applying deep learning models [29], and transformer-based models [30] to extract clinical concept and evaluate the performance of clinical NER system have been conducted.

Methods

In this section, we describe the dataset, error generation, typo correction, MLM based candidate word selection, and deep learning based NER extraction. We examine whether low-quality data with many typos affect the performance of information extraction. To this end, two processes were performed to extract key information from low-quality data. First, we correct typos in two datasets, NCBI-disease data set and surgical pathology records. Second, we identify NER after data was refined through typo correction. Figure 2 illustrates the structure of the overall model. Detailed descriptions of each method are provided in the following sub sections.

Dataset

To evaluate performance of typo correction and NER after typo correction, we used the NCBI-disease dataset, which is mainly used as a benchmarking dataset for NER in the medical domain. In addition, EHRs related lung cancer was provided and used by Asan Medical Center.

A total of 40,443 diagnosis results were composed of five columns, including the date and time of the prescription, the test code, and the text of the test result. We extract the type of test, test institution, test location, result, and size of cancer from the text of the test result. Types of tests include PCNA, needle biopsy, bronchial washing, and pleural fluid etc. even if the name of the test written in the text of the test result is the same, there are some differences depending on the author. Figure 3 shows the example of the text of the test results, which is the input data of the model. The first part is the name of organ, location, operation name, histology diagnosis, tumor size, and invasion of lymph node in order. We excluded invasion of lymph node from the range of information extraction for our work. Table 2 shows the size of training and test data in two datasets each. In surgical pathology records, one line of the text content of the test result in excel file was defined as one sentence and used as the input of the model.

Because the content of the test result is written differently according to the authors without standardized rules, there are many exceptions that cannot be extracted by the rule-based extraction method. In addition, there are parts that cannot be extracted in a rule-based manner due to typos that occur while the author is typing. Therefore, in this study we adapted a deep learning-based algorithm to deal with the part that could not be extracted due to the exception cases and typos that cannot be handled by the rule-based information extraction.

Table 2
Data Specification

Datasets	Train/Test data	Sentences	Tokens
NCBI-disease	Train	6347	159670
	Test	940	24497
Surgical pathology record	Train	39443	2050125
	Test	1000	49668

Error Generation

Since there is no benchmarking dataset for evaluating typo correction related studies in the medical field, most studies use a method of randomly generating and evaluating typos for evaluation. Therefore, in this study, as in the previous study, the types of typos were defined as four types as shown in Table 3, and for types were randomly generated. Special characters, numbers, and acronyms are made to avoid typos.

Table 3
Type of Errors and Examples

Error type	Example
Insertion	differenece/difference
Delete	randm/random
Replace	appand/append
Transpose	money/moeny

Typo Correction

In this study, the candidate group of words to be corrected is selected using the Edit distance algorithm and typo is corrected by scoring each word candidate group in consideration of the frequency of words in the dictionary and the context within the sentence. The SymSpell algorithm was used to generate a candidate group of words to be modified [31]. In general, in order to create a word candidate group based on the Edit distance, four types of text correction processes are performed: delete, transpose, replace, and insert. However, the SymSpell algorithm reduced the amount of computation by using only the delete approach. In order to optimize the SymSpell algorithm for the medical domain, we trained PubMed abstract (about 25.4GB of literature updated in December 2019). In addition, word dictionary was created that summarized words and their frequency using the PubMed abstract. If the frequency of words appeared in the entire collections was less than 20, the words were excluded from the word dictionary. As a result, a total of 2,370,526 words were included in the dictionary. Based on the generated word dictionary, a group of word candidates was generated using the SymSpell algorithm.

MLM based Candidate Word Selection

In order to find an appropriate word to correct a typo among the generated word candidates, we used a score that combines a frequency-based score and context-based score. The formula is as follows.

$$(FinalScore) = \lambda(FrequencyScore) + (1 - \lambda)(ContextSensitiveScore)$$

Using the MLM method, scores were obtained in consideration of the context within the sentence. In particular, a BERT-base pre-trained MLM model was used. By adding a dense layer to the pre-trained model, we calculated the probability of a specific word in the masked part of the input sentence and used it as a score for the context. Table 4 shows the structure of the model added to the pre-trained BERT model.

Table 4
structure of the candidate word selection model

Layer	Output Shape
Input	(None, None, 768)
Dense	(768, 768)
Layer Normalization	(768,)
Output	(None, None, 30522)

Figure 4 shows the process of finding an appropriate cored word from the generated candidate word group. ‘Eting’ in Fig. 3 (a) should be corrected to ‘eating’. (b) shows the generation of candidate word to correct typos through the optimized SymSpell algorithm. In (c) and (d), the part where the typo appears is masked and replaced with a candidate word, and the probability of the word entering a specific position in the sentence is calculated through the MLM model.

Deep Learning based NER extraction

In order to extract key information from medical data, we fine-tuned by adding a dense layer to the pre-trained BERT. The hyper parameters used for fine tuning are shown in Table 5. The output value y_i of the model means the probability that the input value x_i belongs to n tags of each dataset, and it has a vector of the form (1, n). The softmax function is used to learn the probability that the model belongs to j tags. When a_k is the probability value of the k th tag among n tags, the probability that the k th tag is correct is y_k. The formula of softmax function is as follows:

$$y_k = \exp(a_k / \sum_{i=1}^{n} (\exp(a_i)$$

As a result of the model, each word is tagged as one of [B-Disease, I-Disease, O] for the NCBI-disease dataset, and [B-ORGAN, I-ORGAN, B-LOCATION, I-LOCATION, B-OPNAME, I-OPNAME, B-HISTOLOGIC DIAGNOSIS, I-HISTOLOGIC DIAGNOSIS, B-TUMOR_SIZE] for the surgical pathology record. Through this, it

is possible to extract the key information from the surgical pathology record, such as organ, location, operation name, histology diagnosis, and tumor size.

Table 5
hyper parameters

Hyperparameter	Value
learning rate	3e-5
epochs	3
max sequence length	178
batch size	16
optimizer	Adam
activation function	Softmax

Experimental Results

Validation of Context based Typo Correction

The NCBI-disease and surgical pathology record data were evaluated for the SymSpell algorithm optimized for medical data and proposed typo correction model considering contextual information. For data evaluation, about 16% and 7% of typos were randomly generated for each data. Compared with the performance of the SymSpell algorithm, we tried to verify the difference in the performance of typo correction when contextual information was included. Since a word with a typo among all words appears with a very low probability compared to a word without a typo, the number of values corresponding to each class is unbalanced. Therefore, F1-score was used for quantitative performance evaluation.

To evaluate the typo correction performance, typos were randomly generated for each dataset. Table 6 shows the types of typos created in the NCBI-disease and the surgical pathology record dataset and the number of words in which each typo appeared. As shown in Table 6, the performance of the typo correction model was performed in the generated sentences containing typos.

Table 6
Error type and number in two datasets

Datasets	replace	delete	transpose	insert	no error
NCBI-disease	1014	1073	911	946	20553
Surgical pathology record	965	948	877	827	46051

Table 7 shows the typo correction performance for each dataset. performance of the typo correction model was performed in the generated sentences containing typos. In the two data sets, the F1-score of

the model considering contextual information were 0.72 and 0.73, which improved performance about 5% and 9% compared to the case where contextual information was not included.

Table 7
performance of typo correction in NCBI-disease and surgical pathology record datasets

Datasets	algorithms	Typo	precision	recall	f1-score	support
NCBI-disease	SymSpell	Typo X	0.94	0.92	0.93	20553
		Typo O	0.62	0.7	0.67	3944
	Proposed model	Typo X	0.96	0.92	0.94	20553
		Typo O	0.65	0.81	**0.72**	3944
Surgical pathology record	SymSpell	Typo X	0.98	0.96	0.97	46051
		Typo O	0.59	0.7	0.64	3671
	Proposed model	Typo X	0.98	0.97	0.98	46051
		Typo O	0.70	0.76	**0.73**	3671

Figure 5 shows an example of correcting typos and typos in the surgical pathology record data. In the case of Fig. 5 (a), 'lmyph' is originally 'lymph' and it was confirmed that it was corrected through the model. In Fig. 5 (b), 'lug' and 'wedgoe' mean 'lung' and 'wedge' respectively, and they are the body organs and test names. If there was a typo in the part containing important information, it was confirmed that the word was corrected through the model.

Validation of NER through Data Quality Enhancement

We show the effect of typos in the document on the performance of named entity recognition by randomly generating 5 ~ 15% of typos in the NCBI-disease dataset. In the case of well-refined data without typos, the F1-score is 0.89, whereas the F1-score drops to 0.85 when 5% of typos are included, 0.82 when 10% are included, and 0.77 when 15% are included as shown in Fig. 5.

When 16% of typos were included in the NCBI-disease dataset, the average F1-score of name entity recognition was 0.77, and after the typo correction model was applied, the performance has been improved about 2% as 0.79. The detailed performance evaluation results in each case are shown in Table 8.

Table 8
performance of NER including 16% errors and after typo correction in NCBI-disease dataset

	precision		recall		f1-score		support
	Typos	Typo correction	Typos	Typo correction	Typos	Typo correction	
B-Disease	0.96	0.95	0.73	0.74	0.83	0.84	960
I-Disease	0.97	0.97	0.66	0.71	0.79	0.82	1087
	Typos			Typo correction			
accuracy	0.70			0.72			2047
f1-score	**0.77**			**0.79**			

In the case of the surgical pathology record dataset, the performance of named entity recognition was evaluated when 7% of typos were included. As shown in Table 9, the average F1-score was 0.6 when typos were included, and the performance of named entity recognition increased significantly as 0.85 after correcting typos. Through this, it was confirmed that the typo had a significant effect on extracting information from the unstructured text.

Table 9
performance of NER including 7% errors and after typo correction in surgical pathology record dataset

	precision		recall		f1-score		support
	Typos	Typo correction	Typos	Typo correction	Typos	Typo correction	
B-ORGAN	0.97	1.00	0.31	0.82	0.47	0.90	1022
I-ORGAN	1.00	1.00	0.37	0.73	0.54	0.84	196
B-LOCATION	1.00	1.00	0.25	0.79	0.41	0.88	625
I-LOCATION	1.00	1.00	0.26	0.81	0.41	0.89	1338
B-OPNAME	0.90	0.98	0.54	0.97	0.68	0.97	822
I-OPNAME	1.00	1.00	0.46	0.96	0.63	0.98	803
B-HISTOLOGIC DIAGNOSIS	1.00	1.00	0.83	0.99	0.91	0.99	70
I-HISTOLOGIC DIAGNOSIS	1.00	1.00	0.90	1.00	0.95	1.00	114
B-TUMOR_SIZE	1.00	1.00	0.99	0.99	1.00	1.00	222
	Typos			Typo correction			
accuracy	0.40			0.87			2047
f1-score	**0.60**			**0.85**			

Discussion

In this study, we focused on the problems that deteriorate the accuracy of entity extraction and the performance of text-based analysis in the medical field. A typo problem that frequently appears in unstructured texts in the real medical field has a negative effect on the performance of text-based information extraction. Therefore, we tried to improve the performance of extracting important information from unstructured medical texts by resolving the typo problem in the unrefined text. Through experiments, we verified that typos occurring in text have a negative effect on the performance of natural language processing tasks.

The improvement of NER performance before and after typo correction showed a difference between NCBI-disease and surgical pathological record dataset. As described in the experimental results section, in the surgical pathological record dataset, the F1-score of NER detection increased significantly by 25% through the typo correction model, whereas in NCBI-disease dataset increased only by 2%. Since NER is sequence labeling, it affects the recognition of NER according to the preceding and following words.

Therefore, in the case of a long entity name, it is less affected by typos compared to a short entity name in recognizing the entity name.

Figure 6 shows the result of comparing the entity proportion by length appearing in the two datasets. In the case of surgical pathology record dataset, the ratio of a single word entity occupied half of the total entities, and more than 90% of entities had a maximum length 3. On the other hand, in the NCBI-disease dataset, entities with a length of 4 or more occupied around 20%. Due to the nature of the entity length between the two datasets, the NCBI-disease dataset having many relatively long entities showed little difference in performance before and after typo correction. In addition, these characteristics differed depending on the entity name existing in the surgical pathology record dataset. For example, in Table 9, there are five entity names in surgical pathology record dataset such as ORGAN, LOCATION, OPNAME, HISTOLOGIC DIAGNOSIS, and TUMOR_SIZE. Among them, HISTOLOGIC DIAGNOSIS is the longest entity with an average length 6, and F1-score increased by 8% for Begin and 5% for Inside after correcting typo. This entity was less affected by typos than the other four shorter entities.

We confirmed that the typographical correction model proposed in this study helps extract accurate information especially from unstructured data. We also reviewed that some exceptional cases may occur infrequently in our real-world dataset. For example, in the case of the size about the cancer, from the contents written as ‘2.1 X 1.3 X 1 cm’, the largest value ‘2.1’ should be extracted. In the most cases, it was written in the form of ‘Number X Number X Number’ as in the example, but sometimes it was written in another expression such as ‘1.2 cm IN GREATEST DIMENSION’. Although this case occupies a small proportion of the total data, in order to apply the rule-based model, a person must make a rule by considering the number of all cases. This task is labor intensive and when a new exception appears, it is impossible to extract it with the existing rule, so the rule needs to be modified again. Therefore, we were able to supplement the limitation of rule-based information extraction by applying the BERT-based named entity recognition model to unstructured text in real-world medical fields. In this way, it was confirmed that the proposed typo correction model can contribute to accurate information extraction from unstructured data.

Conclusion

Due to the digitization of data, a huge amount of patient data is accumulated every day, such as Electronic Health Records in the real-world medical fields. Along with the changes in technology, there is an active movement to build and analyze accumulated data platforms in the medical fields. In particular, a large proportion of Electronic Health Records data is written descriptively without standard pattern. Extracting information, the basis for judging a patient's lung cancer staging from the unstructured text, is necessary in a real-world medical environment. In the case of unstructured data in the real-world medical fields, since the use of delimiters or abbreviations in text differs depending on the record writer, it was necessary to apply a robust methodology to resolve these exceptions. In addition, it is difficult to extract and structure important information because the writer frequently makes typos in the typing process.

The purpose of this study is to propose a robust model that extracts necessary information even in undefined exceptions and typos situations and to be applied to real-world medical fields. To this end, two tasks were carried out: correcting typos in the data where the typo appeared and extracting information by utilizing named entity recognition in the corrected text. The major contribution of the present study is that we proposed a model that can be applied to real world environment by focusing on problems that cause difficulties in analysis in real-world medical fields. In addition, our experiments have shown that typos occurring in text data have a negative effect on the performance of natural language processing tasks.

Studies of pre-trained BERT were conducted based on data from various domains. Regarding the medical domain, there are representative ones such as BioBERT and ClinicalBERT. In the case of the BERT-base model used in this study, it was trained with Wikipedia and BookCorpus, not domain dependent data, so additional verification is needed to determine whether it is suitable for the medical domain. Therefore, as a future study, we will select an appropriate pre-trained model by comparing the performance of BioBERT and ClinicalBERT trained as text in the medical domain when correcting typos in real-world medical data. In addition, it is expected that the performance of typo correction will be improved by constructing a word dictionary based on a corpus related to clinical data such as the MIMIC-III Clinical Database rather than the scientific literature PubMed abstract.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

This work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)). This work was also partly supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5B1104865). This study was also supported by a grant (Elimination of Cancer Project Fund) from Asan Cancer Institute of Asan Medical Center, Seoul, Korea.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EBL conceived and designed the study. EBL and GEH conducted the experiment and drafted the manuscript. MS designed and coordinated the study. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data availability

The data that support the findings of this study are available from Asan Medical Center, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Asan Medical Center.

References

1. Scherf M, Epple A, Werner T. **The next generation of literature analysis: integration of genomic analysis into text mining**. Briefings in bioinformatics 2005;**6**(3): 287–297.
2. Delen D, Crossland MD. **Seeding the survey and analysis of research literature with text mining**. Expert Systems with Applications 2008;**34**(3): 1707–1720.
3. Zhong N, Li Y, Wu ST. **Effective pattern discovery for text mining**. IEEE transactions on knowledge and data engineering 2010;**24**(1): 30–44.
4. Chen H, Chiang RH, Storey VC. **Business intelligence and analytics: From big data to big impact**. MIS quarterly 2012;1165–1188.
5. Das TK, Kumar PM. **Big data analytics: A framework for unstructured data analysis**. International Journal of Engineering Science & Technology 2013;**5**(1): 153.
6. Gandomi A, Haider M. **Beyond the hype: Big data concepts, methods, and analytics**. International journal of information management 2015;**35**(2): 137–144.
7. Moro S, Cortez P, Rita P. **Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation**. Expert Systems with Applications 2015;**42**(3): 1314–1324.
8. Bello-Orgaz G, Jung JJ, Camacho D. **Social big data: Recent achievements and new challenges**. Information Fusion 2016;*28*: 45–59.

9. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, … Schrag D. **Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports**. JAMA oncology 2019;**5**(10): 1421–1429.
10. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, … Shuren J. **Real-world evidence—what is it and what can it tell us**. N Engl J Med 2016;**375**(23): 2293–2297.
11. Hersh WR, Campbell EM, Malveau SE. **Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records**: **a lexical analysis**. In: *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association 1997. 580.
12. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, Rocha RA. **How many medication orders are entered through free-text in EHRs?-a study on hypoglycemic agents**. In: *AMIA annual symposium proceedings*. American Medical Informatics Association 2012, 1079.
13. Ruch P. **Using contextual spelling correction to improve retrieval effectiveness in degraded text collections**. In *COLING 2002: The 19th International Conference on Computational Linguistics*. 2002.
14. Basma S, Lord B, Jacks LM, Rizk M, Scaranelo AM. **Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription**. American Journal of Roentgenology 2011;**197**(4): 923–927.
15. Lambert BL. **Predicting look-alike and sound-alike medication errors**. American Journal of Health-System Pharmacy 1997;**54**(10):1161–1171.
16. Lai KH, Topaz M, Goss FR, Zhou L. **Automated misspelling detection and correction in clinical free-text records**. Journal of biomedical informatics 2015;55: 188–195.
17. Britz D, Goldie A, Luong MT, Le Q. **Massive exploration of neural machine translation architectures**. *arXiv preprint arXiv:1703.03906*, 2017.
18. Lee, JH, Kim M, Kwon HC. **Deep Learning-Based Context-Sensitive Spelling Typing Error Correction**. IEEE Access 2020;8: 152565–152578.
19. Senger C, Kaltschmidt J, Schmitt SP, Pruszydlo MG, Haefeli WE. **Misspellings in drug information system queries: characteristics of drug name spelling errors and strategies for their prevention**. International journal of medical informatics 2010;**79**(12): 832–839.
20. Kilicoglu H, Fiszman M, Roberts K, Demner-Fushman D. **An ensemble method for spelling correction in consumer health questions**. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association 2015, 727.
21. Workman TE., Shao Y, Divita G, Zeng-Treitler Q. **An efficient prototype method to identify and correct misspellings in clinical text**. BMC research notes 2019;**12**(1): 1–5.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, … Polosukhin I. **Attention is all you need.** In: *Advances in neural information processing systems*. 2017, 5998–6008.
23. Devlin J, Chang MW, Lee K, Toutanova K. **Bert**: **Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*, 2018.

24. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. **Distributed representations of words and phrases and their compositionality**. In: *Advances in neural information processing systems*. 2013, 3111–3119.
25. Bojanowski P, Grave E, Joulin A, Mikolov T. **Enriching word vectors with subword information**. Transactions of the Association for Computational Linguistics 2017;**5**: 135–146.
26. Pennington J, Socher R, Manning CD. **Glove**: **Global vectors for word representation**. In: *Proceedings of the* 2014 *conference on empirical methods in natural language processing (EMNLP)*. 2014;1532–1543.
27. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. **A study of active learning methods for named entity recognition in clinical text**. Journal of biomedical informatics, 2015;**58**: 11–18.
28. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. **A study of neural word embeddings for named entity recognition in clinical text**. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association. 2015. 1326.
29. Wu Y, Jiang M, Xu J, Zhi D, Xu H. **Clinical named entity recognition using deep learning models**. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association 2017. 1812.
30. Yang X, Bian J, Hogan WR, Wu Y. **Clinical concept extraction using transformers**. Journal of the American Medical Informatics Association, 2020;**27**(12): 1935–1942.
31. SymSpell. https://github.com/wolfgarbe/SymSpell. Accessed 20 August 2020.

Figures

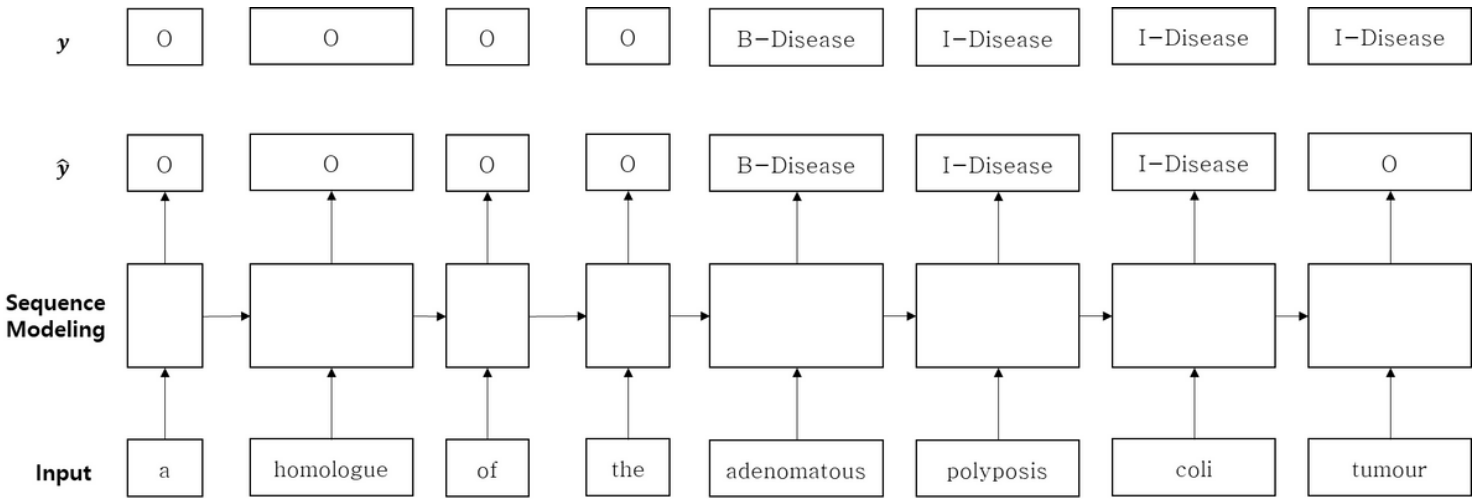

Figure 1

Example of Named Entity Recognition

Figure 2

Research overflow

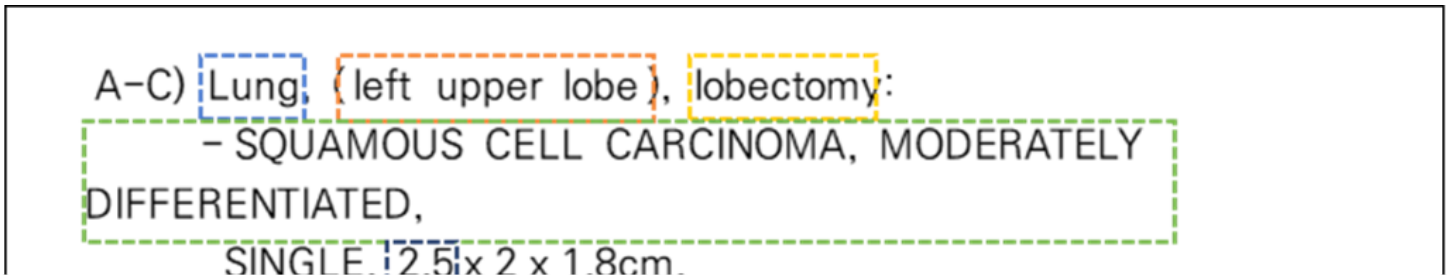

Figure 3

Example of the contents of the test result in surgical pathology record

Figure 4

Example of the candidate word selection process

Figure 5

Example of typo correction through model

Figure 6

NER performance by error rate in NCBI-disease dataset

Figure 7

Comparison of entity ratio by entity length in NCBI-disease and surgical pathology record dataset