BMC Medical Informatics and Decision Making

RESEARCH Open Access

Korean clinical entity recognition from diagnosis text using BERT

Young-Min Kim[1,2*] and Tae-Hoon Lee[2]

From The 13th International Workshop on Data and Text Mining in Biomedicine
Beijing, China. 07 November 2019

Abstract

Background: While clinical entity recognition mostly aims at electronic health records (EHRs), there are also the demands of dealing with the other type of text data. Automatic medical diagnosis is an example of new applications using a different data source. In this work, we are interested in extracting Korean clinical entities from a new medical dataset, which is completely different from EHRs. The dataset is collected from an online QA site for medical diagnosis. Bidirectional Encoder Representations from Transformers (BERT), which is one of the best language representation models, is used to extract the entities.

Results: A slightly modified version of BERT labeling strategy replaces the original labeling to enhance the separation of postpositions in Korean. A new clinical entity recognition dataset that we construct, as well as a standard NER dataset, have been used for the experiments. A pre-trained multilingual BERT model is used for the initialization of the entity recognition model. BERT significantly outperforms a character-level bidirectional LSTM-CRF, a benchmark model, in terms of all metrics. The micro-averaged precision, recall, and f1 of BERT are 0.83, 0.85 and 0.84, whereas that of bi-LSTM-CRF are 0.82, 0.79 and 0.81 respectively. The recall values of BERT are especially better than that of the other model. It can be interpreted that the trained BERT model could detect out of vocabulary (OOV) words better than bi-LSTM-CRF.

Conclusions: The recently developed BERT and its WordPiece tokenization are effective for the Korean clinical entity recognition. The experiments using a new dataset constructed for the purpose and a standard NER dataset show the superiority of BERT compared to a state-of-the-art method. To the best of our knowledge, this work is one of the first studies dealing with clinical entity extraction from non-EHR data.

Keywords: Clinical entity recognition, BERT, Korean, Diagnosis text

Background

Clinical entity recognition traditionally aims at electronic health records (EHRs) [1] generated by healthcare providers. EHRs contain clinical information about patients including diagnoses, laboratory tests, clinical notes, etc [2]. The target entities are mostly technical terms precisely written by medical specialists. The medical problem, treatment, and test are typical entity types of these texts [3]. The extracted entities are fundamental to build clinical informatics applications [4]. Identification of patient cohorts, extraction of adverse drug events, and finding the relationships of drug-disease treatment are some of the applications [1, 5, 6].

While the above applications concern the traditional biomedical informatics, there is also the demand for extracting clinical entities for new applications from another domain. It is automatic medical diagnosis via a

*Correspondence: yngmnkim@hanyang.ac.kr
[1] Graduate School of Technology & Innovation Management, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, South Korea
[2] Division of Interdisciplinary Industrial Studies, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, South Korea

dialogue system. There is a recent work dealing with medical diagnosis data extracted from user-generated QA set [7]. Its goal is to build an automatic medical diagnosis system, which operates via a conversational process. The clinical entities expressed in the data are essential to construct the system. However, entity recognition is not the main concern of the work although it is necessary for the actual use. The authors focus on building a system by reinforcement learning. Considering the increasing interests in healthcare applications, this kind of expanded use of clinical entity extraction will be a valuable research theme.

In this paper, we are interested in extracting clinical named entities from a new medical dataset, which is completely different from EHRs. The novelty of the dataset mainly comes from the originality of the data source and the annotated entity types. The dataset is collected from an online QA site for medical diagnosis, and the target language is Korean. Three different entity types, which are essential for the diagnosis, are defined and annotated. We use a state-of-the-art NLP technique, Bidirectional Encoder Representations from Transformers (BERT) [8] for the entity extraction. We show BERT using WordPiece tokenization is effective for the domain-specific named entity recognition in Korean.

Our clinical entity recognition is originally designed to be a part of a dialogue system for health advice. The final goal of the system is to provide simple health advice to users via conversation. This work will be the first step for the final goal. Figure 1 shows the overall process of the target system following the recent dialogue system architectures [9–11]. The interest of this paper is emphasized with red rectangles.

From the initial user utterance, the system first extracts the user intent and useful entities. In the example, the extracted entities are stomach and hurt as body part and symptom respectively. Then a dialogue manager tries to understand the current state of the conversation to derive the best answer. In this case, the system tries to get the start time of the illness to exactly identify the current condition of the user. Here, the entity recognition is an important part that influences the most the final system performance.

Named entity recognition (NER) is usually considered as a sequence labeling task. The recent developments of long-short term memory (LSTM) variants such as bidirectional LSTM-CRF [12] and bi-LSTM-CNNs-CRF [13] have achieved a success in both NER and biomedical NER [14, 15]. Another popular research direction is to apply the attention mechanism [16]. Various sequence labeling approaches using the mechanism have been developed in different tasks, such as slot filling [17], role labeling [18], and bio-medical information detection [19].

The Korean NLP community has adopted similar methodologies for sequence labeling and also obtained better performances than traditional CRF-based approaches. However, because of the linguistic property that the words are not always clearly separated, the tokenization and input encoding influence a lot to the final performance in general. The character-level n-gram encoding with additional linguistic information is one of the state-of-the-art approaches for Korean NER [20]. A recent work reports that *jamo* (Korean alphabet) level representation extracts well the word semantics in terms of word similarity [21]. Another work on NER proposed to use a hybrid representation of morpheme vectors [22]. These works focus on finding the best input representation in common.

BERT is a recently developed language representation model. It trains a deep bidirectional representation of a large unlabeled corpus using stacked Transformer encoders. Then, the representation is fine-tuned with an additional output layer for downstream NLP tasks. The main difference with Embeddings from Language Models (ELMo) [23] and OpenAI Generative Pre-training Transformer (GPT) [24], the precedent models, is that BERT is bidirectional when applying self-attention.

In this work, we use BERT to train a NER model for medical entity recognition. We investigate the effects of proper tokenization as well as labeling strategy for evaluation. We empirically show the simple WordPiece representation is effective for the domain-specific NER in Korean even with a small dataset. An additional experiment is also provided to verify the effectiveness of BERT for Korean NER on a standard dataset. A bi-LSTM-CRF

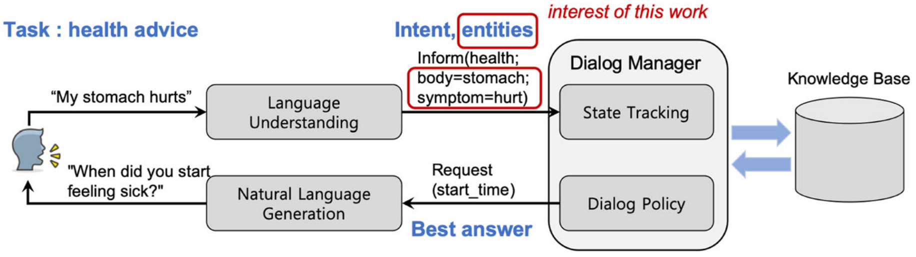


Fig. 1 Architecture of a dialogue system

model is selected as a benchmark to show the superiority of BERT for Korean medical NER.

Methods

We constructed a clinical NER dataset that contains medical experts' diagnoses to the questions of an online QA service. BERT is applied to the dataset to extract the clinical entities. We slightly modified the BERT labels to separate the postpositions, which are very common in Korean. Different labeling formats for the evaluation are proposed to find the most effective one to assess the entity recognition results.

Bidirectional encoder representations from transformers

Figure 2 shows the architecture of BERT. The pre-trained model weights are re-used for the training of an NLP task, NER, in our case. Unlike the precedent models, BERT pre-trains the model using a masked language model (LM). While a standard LM aims to predict the next token for each token in a sentence, a masked LM predicts a randomly masked token given a sentence. This simple but effective mechanism allows the bi-directionality in the model. BERT produces surprisingly good results on most NLP tasks. It outperforms the existing state-of-the-art methods on 11 tasks including NER. An interesting part for us is that it uses a simple WordPiece tokenization [25] for input that is more appropriate for Korean than the other English text encodings.

WordPiece tokenization separates words into different tokens as shown in the 'input' row of Fig. 3. Most meaningful words are kept and the other words are tokenized into pieces. To express the continuity of the tokens, two sharps (##) are attached in front of the token when it is a part of the precedent token. In the example, the word 'playing' is split into 'play' and '##ing'. It means the word was less occurring than the other words when training the WordPiece representation. The separated tokens are called sub-words. Two special tokens [CLS] and [SEP] are added to express the start of the data instance and the end of a sentence.

The segmentation embeddings are used to distinguish different sentences. The position embeddings represent the token positions in the instance. Each token is finally represented by the sum of the token embedding, the segmentation embedding, and the position embedding. With this tokenization, BERT keeps a very small number of sub-words. In English, the vocabulary size of a trained BERT model is only 30,522. The final input embedding dimension is 768.

BERT uses the encoders of the Transformer architecture. Using the Transformers has become very common in language modeling [26]. The Transformer has Multi-Head Attention, which concatenates h different attention layers with different initializations [16]. Multi-Head Attention is written as follows:

$$Multihead(Q,K,V) = Concat(head_1, ..., head_h)W^O$$
$$where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

The $head_i$ is i^{th} attention head. $W_i^Q \in \mathbb{R}^{d_{emb} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{emb} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{emb} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{emb}}$ are projection matrix parameters. Q, K and V are different input matrices. At the beginning, input matrix X are used for all three matrices. Then their projections XW_i^Q, XW_i^K and XW_i^V become Q_i, K_i, and V_i. These matrices are used to compute the following Scaled Dot-Product Attention.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}V\right)$$

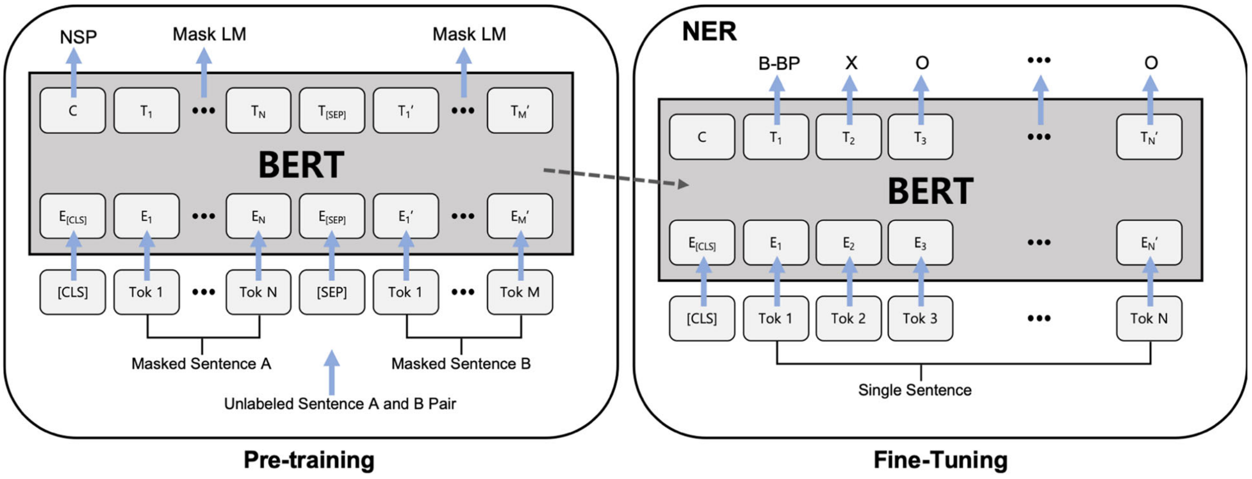


Fig. 2 BERT architecture and fine-tuning

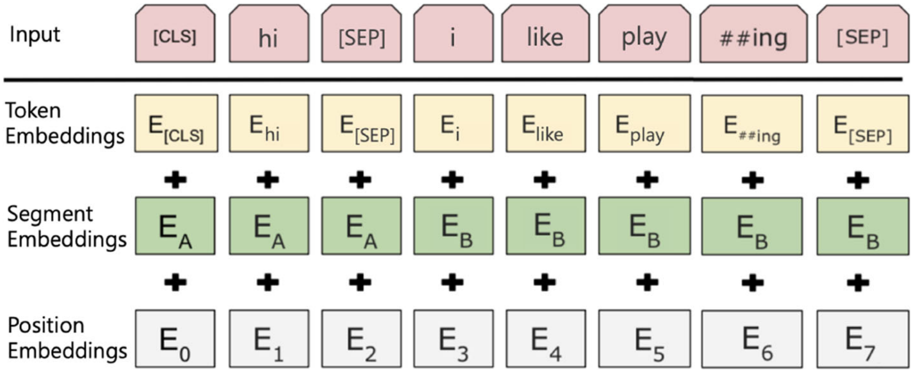


Fig. 3 BERT input representation. (modified image from [8])

We use a pre-trained BERT model's weights for the initialization of the clinical entity recognition task. The BERT architecture is reused and the input and output are adjusted to our task.

Data collection and training set

The source data is collected from the biggest Korean online QA service[1] that the platform is provided by a web portal. The questioners and answerers are all the portal users. The questions are categorized into different sections according to the nature of the question. In the medical section, there are 24 different departments. Each medical department has a special subcategory, which contains the QA pairs answered by medical specialists.

We select four departments most relevant for the automatic diagnosis. The department of neurology, neurosurgery, internal medicine, and otorhinolaryngology are those four. We collect QA data answered by medical specialists from Jan. 2009 to Aug. 2018. Two medical specialists who wrote most of the answers per department have been selected. Then QA sets answered by the selected specialists have been all collected.

Among the data, we randomly selected 200 QA pairs per department. To construct the training set, we first use the answers only because they include much more formulaic expressions than the questions. After testing the effectiveness of BERT for clinical entity recognition using answer sets, we also test question sets, which include much more colloquial expressions.

By filtering out some duplicated answers and unnecessary questions such as MRI reading or military service exemption issue, we have a total of 536 answers. As a long answer tends to include supplementary information at the end, we kept 5 first sentences the maximum excluding greeting messages. The statistics of the QA dataset is given in Table 1.

For the construction of the training set, three annotators discussed the annotation guidelines. Each instance, which corresponds to an answer, is annotated by one annotator and is reviewed by another annotator. After a detailed review of the source data, we select three different clinical entity types that are essential for diagnosis. These are 'disease', 'symptom' and 'body part'. The definition of the entity types is given in Table 2.

For the annotation, we first collected the entity dictionaries from "National Health Information Portal" and hospital web sites. The number of terms is 2,191 for disease names, 142 for symptoms, and 139 for body parts. Some conflict terms have been preprocessed before annotation. Terms not found in the dictionaries have been searched using the other health information portals such as "Infectious Disease Portal" of Korea Centers for Disease Control and Prevention (KCDC) and "Medical Encyclopedia" of Seoul National University Hospital.

The main annotation guidelines for the training set construction are as follows:

- Frequent informal disease names such as "허리(waist) 디스크(disc)", which means "lumbar herniated intervertebral disc", are considered as a disease.
- Symptoms are usually nouns but can be the combination of adjective and symptom when the expression is very common. For example, "저린(benumbed) 증상(symptom)".

Table 1 Statistics of the QA dataset for diagnosis

department	# anwsers	# sents
neurology	126	630
neurosurgery	156	780
internal medicine	131	655
otorhinolaryngology	123	615
total	536	2189

[1] https://kin.naver.com

Table 2 Entity definition for clinical NER

Entity type	definition
Disease (DZ)	disease name. used for final diagnosis
Symptom (SX)	symptom which can be detected by users
Body Part (BP)	body part where the symptom reveals

Table 3 Characteristics of the annotated diagnosis data

	DZ	SX	BP
# annotated unique entities	297	228	199
# annotated entities	915	1,267	1,010

- Body tissues such as muscle, ligament or bone are not the target but organs such as stomach, liver, or brain are because they can be the location of a symptom.

Table 3 shows the characteristics of the annotated data. The number of annotated unique terms and that of total annotated terms are given. The annotated dataset will be available on request.[2]

Tokenization and labeling for BERT model

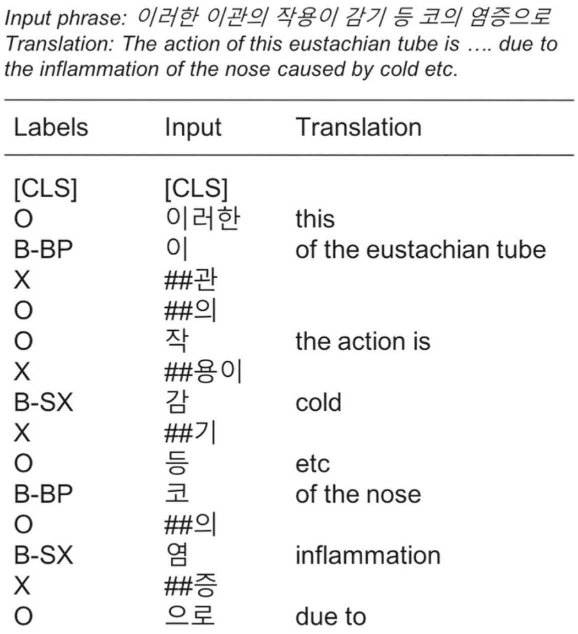
In BERT, WordPiece tokenization and three different embeddings are used to represent input tokens. After a punctuation splitting and whitespace tokenization, WordPiece tokenization separates words into different subwords as explained in the previous section. Figure 4 shows an example of a tokenized sentence in our dataset. The translation in the table is done in word-level separated by whitespaces regardless of the grammatical order. It is because the one-to-one translation from Korean token to English is not possible.

There are two symptoms and two body parts to be detected in the example. The first symptom is ‘감기(cold)’, which is tokenized as ‘감’ and ‘##기’. Its first token is annotated as ‘B-SX’ according to the IOB (Inside-Outside-Beginning) tagging and the remainder is annotated as ‘X’. The second symptom, ‘염증(inflammation)’ is also annotated in the same way.

In the case of the body parts, the annotation is a little different. The first body part, ‘이관(eustachian tube)’ is originally attached with a genitive postposition, ‘의(of)’. After WordPiece tokenization, each syllable(character) is separated, but the third token becomes ‘##의’ because it was a part of the word ‘이관의(of the eustachian tube)’. However, in the annotation, only the tokens ‘이’ and ‘##관##’ are annotated as body parts such as ‘B-BP’ and ‘X’ respectively, but not the token ‘##의’. Therefore, the tag of ‘##의’ should be ‘O’ (outside), instead of ‘X’. In the original version of BERT does not consider this kind of postposition processing for Korean. It is the main difference when we label the training set compared to the original BERT.

During the evaluation, the tokens attached with ‘X’ tags are ignored to compute word-level frequency in the case of general English NER. However, in Korean, this is not a common approach, especially when using character-level embeddings. Because of the postpositions, more than one label can be attached to a word. Therefore, character-level evaluation can be more exact to verify NER performance for Korean. For the comparison with the existing approaches introducing character embeddings, we propose to additionally use two modified versions of tag representations as shown in Fig. 5 when evaluating the results.

Figure 5 contains two tables representing different labeling strategies for WordPiece embeddings (left) and character embeddings (right) respectively. An example of the input phrase is given on top of the tables. The right side table shows the tokenized input features in character-level and their labels in two different formats. The first and second columns correspond to the typical input tokenization and labeling. Whitespace in the input is replaced by a special character, ‘∧’. The foremost token of an entity is annotated as a beginning tag, starting with ‘B’, whereas the others are annotated as ‘I’. When an entity consists of more than a word, the whitespace between words is also annotated as a part of the entity. Many LSTM-based models use this representation as a benchmark.

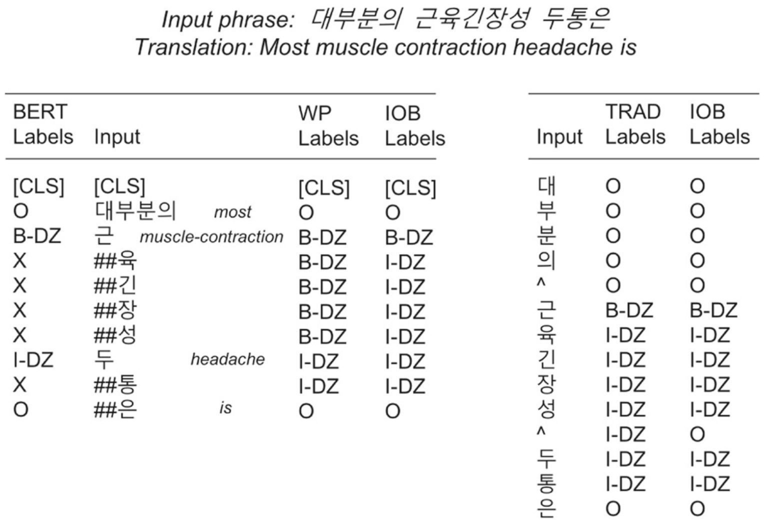


Input phrase: 이러한 이관의 작용이 감기 등 코의 염증으로
Translation: The action of this eustachian tube is …. due to the inflammation of the nose caused by cold etc.

Labels	Input	Translation
[CLS]	[CLS]	
O	이러한	this
B-BP	이	of the eustachian tube
X	##관	
O	##의	
O	작	the action is
X	##용이	
B-SX	감	cold
X	##기	
O	등	etc
B-BP	코	of the nose
O	##의	
B-SX	염	inflammation
X	##증	
O	으로	due to

Fig. 4 An example of tokenized sentence in diagnosis dataset. Each column corresponds to the output labels, input tokens, and the translation of input in English

[2] https://github.com/labihem/BERT-NER-KO



Input phrase: 대부분의 근육긴장성 두통은
Translation: Most muscle contraction headache is

BERT Labels	Input		WP Labels	IOB Labels
[CLS]	[CLS]		[CLS]	[CLS]
O	대부분의	*most*	O	O
B-DZ	근	*muscle-contraction*	B-DZ	B-DZ
X	##육		B-DZ	I-DZ
X	##긴		B-DZ	I-DZ
X	##장		B-DZ	I-DZ
X	##성		B-DZ	I-DZ
I-DZ	두	*headache*	I-DZ	I-DZ
X	##통		I-DZ	I-DZ
O	##은	*is*	O	O

Input	TRAD Labels	IOB Labels
대	O	O
부	O	O
분	O	O
의	O	O
^	O	O
근	B-DZ	B-DZ
육	I-DZ	I-DZ
긴	I-DZ	I-DZ
장	I-DZ	I-DZ
성	I-DZ	I-DZ
^	I-DZ	O
두	I-DZ	I-DZ
통	I-DZ	I-DZ
은	O	O

Fig. 5 An example of WordPiece input representation with three different labeling strategies (WP, IOB, TRAD) for the evaluation. BERT labels are used for the training, whereas WP and IOB labels are used only for the evaluation

There is a problem when we use the TRAD labels for the comparison. As BERT does not need to consider whitespaces, there is no label between two separated words in an entity for WordPiece representation. For an effective comparison when evaluating, we propose to use the IOB labels (third column), where the only difference with the TRAD labels is to ignore whitespace's labels by marking them as 'O'.

The left table in Fig. 5 shows the WordPiece input representation with three different labeling strategies. The first column corresponds to the default BERT labels that we use for the training. For the token-level evaluation, we define a new labeling strategy, called WP(WordPiece) labeling. The WP labeling strategy simply replaces 'X' labels by the precedent token's label when the token is a part of an entity. In this way, the predicted word boundaries can be correctly evaluated. The IOB labeling strategy is the same as that of the right table. In summary, the default BERT labels are used for the training, whereas the other strategies are used only for the evaluation. The IOB

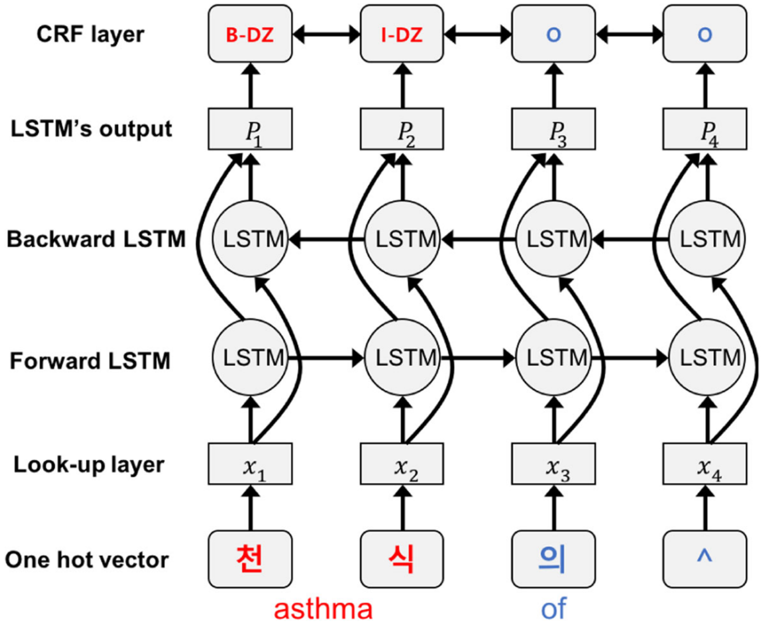


Fig. 6 Bi-LSTM-CRF architecture

Table 4 NER performance comparison results of BERT and bi-LSTM-CRF on the Exobrain dataset

	BERT (IOB)		bi-LSTM-CRF (IOB)	
	macro-avg	micro-avg	macro-avg	micro-avg
Precision	0.93	0.93	0.90	0.89
Recall	0.94	0.94	0.87	0.86
F1	0.94	0.93	0.89	0.88

Table 5 NER performance comparison results of BERT and bi-LSTM-CRF on the diagnosis dataset

	BERT (IOB)		bi-LSTM-CRF (IOB)	
	macro-avg	micro-avg	macro-avg	micro-avg
Precision	0.82	0.83	0.81	0.82
Recall	0.85	0.85	0.78	0.79
F1	0.83	0.84	0.79	0.81

labels are used especially for the comparison of BERT and bi-LSTM-CRF, a benchmark sequence labeling technique.

Results and discussion

The goal of the experiments is to verify the effectiveness of BERT for the clinical entity recognition with our novel dataset. However, for the more reliable experiments, we also conducted the experiments on another Korean NER dataset, Exobrain Korean named entity dataset.[3] It consists of 10,000 sentences with five different entity types: person, location, organization, date, and time. Although the corpus does not aim at medical NER, we expect that this extra experiment would help to justify the effectiveness of BERT for Korean NER. We choose bi-LSTM-CRF as a benchmark model because the model has achieved state-of-the-art performance in Korean NER [20, 22]. The architecture of bi-LSTM-CRF is shown in Fig. 6. It is a bi-directionally connected LSTM variant, which has a CRF layer at the end of the network for the entity recognition.

Experimental setting

The publicly available TensorFlow version of the official BERT implementation has been used for the experiments. A pre-trained multilingual cased model that is applicable for 104 languages, is selected to deal with Korean. It consists of 12-layer, 768-hidden, 12-heads, and 110M parameters. For the NER task, we need to modify the code to prepare the proper input and output. We modified a publicly available code[4], which is also based on the original BERT implementation. The experiments are carried out on an NVIDIA P100 16GB GPU.

Maximum sequence length, train batch size, and learning rate are set to default values, 128, 32, and 2e-5 respectively. The number of train epochs is set to 16. We used 5-fold cross validation for the experiments. Therefore, the obtained results are the averaged values on 5 experiments with different train-test splits.

Experimental results

We first show the NER performance comparison results of BERT and bi-LSTM-CRF on the Exobrain dataset in Table 4. We used the character-level embeddings for the bi-LSTM-CRF model as the example in Fig. 6. The IOB labels introduced in Results and discussion section have been used for the evaluation.

BERT much outperforms bi-LSTM-CRF in terms of all metrics. The micro-averaged precision, recall, and f1 of BERT are 0.93, 0.94 and 0.93, whereas those of bi-LSTM-CRF are 0.89, 0.86, and 0.88. BERT achieved similar results in terms of macro-averaged metrics also. Considering that BERT does not use the CRF layer in the end, the superiority is impressive.

Table 5 represents the NER performance comparison results of BERT and bi-LSTM-CRF on the diagnosis dataset. The IOB labels are also used for the evaluation. BERT again outperforms bi-LSTM-CRF for all metrics. The micro-averaged precision, recall, and f1 of BERT are 0.83, 0.85 and 0.84, whereas those of bi-LSTM-CRF are 0.82, 0.79, and 0.81 respectively. The recall values of BERT are especially better than that of the other model. It can be interpreted that the trained BERT model could detect out of vocabulary(OOV) words, which have not been observed in the training set, better than bi-LSTM-CRF.

Now, we show the results of the BERT model evaluated with the labels more suitable for WordPiece embeddings in Table 6. Two previously introduced formats, the original BERT labels, and WP labels are tested. With the BERT labels, we obtain 0.83, 0.84, and 0.84 in micro-averaged precision, recall, and f1. These results are similar to that of BERT evaluated with the IOB labels. However, the macro-averaged metrics produce much worse results than the micro versions at this time. It is because of the imbalance between classes such that few inside tags are found in the training set. Macro average simply takes the mean value of the different classes, whereas the micro version considers the class proportion. Therefore, the worse result of a

Table 6 NER results of BERT on the diagnosis dataset evaluated with BERT labels and WP label

	BERT labels		WP labels	
	macro-avg	micro-avg	macro-avg	micro-avg
Precision	0.72	0.83	0.70	0.82
Recall	0.72	0.84	0.73	0.84
F1	0.72	0.84	0.72	0.83

[3] http://aiopen.etri.re.kr/

[4] https://github.com/kyzhouhzau/BERT-NER

Table 7 Detailed evaluation result with BERT for the diagnosis dataset

	BERT labels			WP labels			IOB labels		
	precision	recall	f1 score	precision	recall	f1 score	precision	recall	f1 score
B-DZ	0.85	0.88	0.87	0.85	0.89	0.87	0.85	0.88	0.87
I-DZ	0.83	0.82	0.82	0.83	0.84	0.83	0.85	0.88	0.87
B-SX	0.86	0.85	0.86	0.86	0.84	0.84	0.86	0.85	0.86
I-SX	0.53	0.54	0.54	0.54	0.55	0.55	0.83	0.81	0.82
B-BP	0.83	0.86	0.84	0.78	0.83	0.81	0.83	0.86	0.84
I-BP	0.38	0.39	0.39	0.36	0.42	0.39	0.72	0.79	0.75

small class affects a lot the final performance when using the macro versions.

We obtain similar results using WP labels. The evaluation results with BERT labels are slightly better but statistically insignificant. It means that the trained BERT model successfully split a word at the exact word boundary when inferring.

To verify more in detail the result, we show the detailed performance for each entity-tag in Table 7. In the cases of the BERT labels and WP labels, while the disease tags show good results, the inside tags of the symptom and body part show much worse results than the others. The f1 score for I-DZ is 0.83 whereas that of I-SX and I-BP are 0.55 and 0.39 respectively. The main reason is the imbalance among tags. Many disease names consist of more than a word whereas symptoms and body parts usually consist of a word. This phenomenon cannot be captured using IOB labels because the inside tags are attached to from the second character (syllable) of an entity. Therefore, the IOB labeling strategy is not appropriate for an exact evaluation.

Transfer learning

In addition to the previous experiments, we also try applying the trained model to the question data. As the writing style of question set is significantly different from that of answer set, we suppose that two sets are from different domains. Therefore, this additional experiment is a kind of transfer learning, more specifically domain adaptation. The source domain is the answer set and the target domain is question set. For the evaluation of this experiment, we also annotated the question set.

Table 8 represents the detailed result of transfer learning. For the entity types disease and body part, we obtain a slightly worse performance than the standard learning. F1 scores of B-DZ, I-DZ, B-BP, and I-BP are 0.84, 0.85, 0.81 and 0.39 respectively. It means that disease and body part of user questions are automatically extractable using the BERT model trained with the diagnosis (answer) dataset.

On the other hand, the performance in terms of symptom is considerably worse than the others. The recall values are particularly bad such as 0.52 for B-SX and 0.07 for B-BP. We assume that this result comes from the difference between the two sets when representing symptoms. The question set has a lot of mimetic words such as 'pitapat' and 'pricking'. It has also many adjectives symptoms such as 'dizzy' and 'stabbing' whereas the answer set's symptoms are mainly nouns. This difference was likely to influence performance. Therefore we cannot apply this simple transfer learning when detecting symptoms from user questions.

Conclusions

In this paper, we show that the recently developed Bidirectional Encoder Representations from Transformers (BERT) model is effective for the Korean clinical entity recognition. A new dataset for the clinical entity recognition and a standard NER dataset have been tested to verify the effectiveness. The result suggests that the WordPiece tokenization is sufficient in the BERT framework to obtain a state-of-the-art result for the entity recognition in the Korean language. For future work, we are interested in enhancing the prediction performance by combing the core BERT framework with a modified end layer of the network. Then we will move into the next step of the diagnosis system, the extraction of question intent and the construction of the dialogue manager.

Table 8 Transfer learning result tested with the question data using BERT model trained with the diagnosis (answer) dataset

	WP labels		
	precision	recall	f1 score
B-DZ	0.87	0.81	0.84
I-DZ	0.84	0.85	0.85
B-SX	0.78	0.52	0.63
I-SX	0.19	0.07	0.10
B-BP	0.81	0.81	0.81
I-BP	0.69	0.26	0.37
macro-avg	0.70	0.55	0.62
micro-avg	0.81	0.67	0.73

Abbreviations

BERT: Bidirectional encoder representations from transformers; EHRs: Electronic health records; LSTM: Long-short term memory; CRF: Conditional random fields; bi-LSTM-CRF: Bidirectional LSTM-CRF; IOB: Inside-outside-beginning; OOV: Out of vocabulary; CNN: Convolutional neural network; NER: Named entity recognition; ELMo: Embeddings from language models; GPT: Generative pre-training transformer; NLP: Natural language processing; LM: Language model; DZ: Disease; SX: Symptom; BP: Body part

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of BMC Medical Informatics and Decision Making, Volume 20 Supplement 7 2020: Proceedings of the 13th International Workshop on Data and Text Mining in Biomedical Informatics (DTMBIO 2019): medical informatics and decision making. The full contents of the supplement are available at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-7.

Authors' contributions

YK developed the main idea and wrote the manuscript. YK and TL annotated the dataset and conducted the experiments. All authors have read and approved the manuscript.

Funding

This work was funded by two projects, Classification of The Artists using Deep Neural Networks (201600000002255) of Hanyang University, and Smart Multimodal Environment of AI Chatbot Robots for Digital Healthcare (P0000536) of the Ministry of Trade, Industry and Energy (MOTIE). Publication costs are funded by both projects.

Availability of data and materials

The datasets generated and analysed during the current study are available on request in the GitHub repository, https://github.com/labihem/BERT-NER-KO.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

Received: 4 August 2020 Accepted: 2 September 2020
Published: 30 September 2020

References

1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77:34–49.
2. Shickel B, Tighe P, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances on deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform. 2018;22(5):1589–604.
3. Uzuner O, South B, Shen S, DuVall S. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552–6.
4. Jiang M, Chen Y, Liu M, Rosenbloom S, Mani S, Denny J, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc. 2011;18(5):601–6.
5. Feng L, Chiam Y, Lo S. Text-mining techniques and tools for systematic literature reviews: A systematic literature review. In: 2017 24th Asia-Pacific Software Engineering Conference (APSEC). Nanjing: IEEE; 2017. p. 41–50.
6. Zhang R, Cairelli M, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch T, Pakhomov S, Melton G. Using semantic predications to uncover drug-drug interactions in clinical data. J Biomed Inform. 2014;(49):134–47.
7. Wei Z, Liu Q, Peng B, Tou H, Chen T, Huang X, Wong K-F, Dai X. Task-oriented dialogue system for automatic diagnosis. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne: ACL; 2018. p. 201–7.
8. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL; 2019. p. 4171–86.
9. Gao J, Galley M, Li L. Neural approaches to conversational AI. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics:Tutorial Abstracts. Melbourne: ACL; 2018. p. 2–7.
10. Chen H, Liu X, Yin D, Tang J. A survey on dialogue systems: Recent advances and new frontiers. SIGKDD Explor Newsl. 2017;19(2):25–35.
11. Li X, Chen Y-N, Li L, Gao J, Celikyilmaz A. End-to-end task-completion neural dialogue systems. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei: ACL; 2017. p. 733–43.
12. Huang Z, Xu W, Yu K. Bidirectional lstm-crf models for sequence tagging. CoRR. 2015;abs/1508.01991:1–10.
13. Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin: ACL; 2016. p. 1064–74.
14. Chalapathy R, Zare Borzeshi E, Piccardi M. Bidirectional LSTM-CRF for clinical concept extraction. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP). Osaka: COLING; 2016. p. 7–12.
15. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, Xu H. Entity recognition from clinical texts via recurrent neural network. BMC Med Informatics Decis Mak. 2017;67(17):53–61.
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser Lu, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems 30. Long Beach: Curran Associates, Inc.; 2017. p. 5998–6008.
17. Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling. In: INTERSPEECH; 2016.
18. Tan Z, Wang M, Xie J, Chen Y, Shi X. Deep semantic role labeling with self-attention. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence; 2018. p. 4929–36.
19. Rei M, Crichton G, Pyysalo S. Attending to characters in neural sequence labeling models. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: COLING; 2016. p. 309–18.
20. Kwon S, Ko Y, Seo J. A robust named-entity recognition system using syllable bigram embedding with eojeol prefix information. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM; 2017. p. 2139–42.
21. Park S, Byun J, Baek S, Cho Y, Oh A. Subword-level word vector representations for korean. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018. p. 2429–38.
22. Na S, Kim H, Min J, Kim K. Improving LSTM crfs using character-based compositions for korean named entity recognition. Comput Speech Lang. 2019;54:106–21.
23. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: ACL; 2018. p. 2227–37.
24. Alec Radford TS, Karthik Narasimhan, Sutskever I. Improving language understanding with unsupervised learning. Technical report, OpenAI. 2018. https://openai.com/blog/language-unsupervised/.
25. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR. 2016;abs/1609.08144:.
26. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag. 2017;13(3):55–75.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.