(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0124536
(43) 공개일자 2023년08월25일

(51) 국제특허분류(Int. Cl.)
G06F 18/241 (2023.01) *G06F 16/31* (2019.01)
G06F 16/35 (2019.01) *G06F 18/214* (2023.01)
G06F 40/205 (2020.01) *G06F 40/232* (2020.01)
G06F 40/284 (2020.01) *G06N 20/20* (2019.01)
G16H 50/20 (2018.01)
(52) CPC특허분류
G06F 18/241 (2023.01)
G06F 16/31 (2019.01)
(21) 출원번호 10-2023-0107923(분할)
(22) 출원일자 2023년08월17일
심사청구일자 2023년08월17일
(62) 원출원 특허 10-2021-0003764
원출원일자 2021년01월12일
심사청구일자 2021년01월12일

(71) 출원인
주식회사 에임메드
서울특별시 강남구 도산대로 221, 3층 (신사동, 동남빌딩)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
임진환
서울시 강남구 도산대로 221 301호
김하영
서울시 서대문구 명지길 30 신원지벤스타 107동 602호
신동엽
서울시 서대문구 봉원사길 24, D103호
(74) 대리인
특허법인비엘티

전체 청구항 수 : 총 10 항

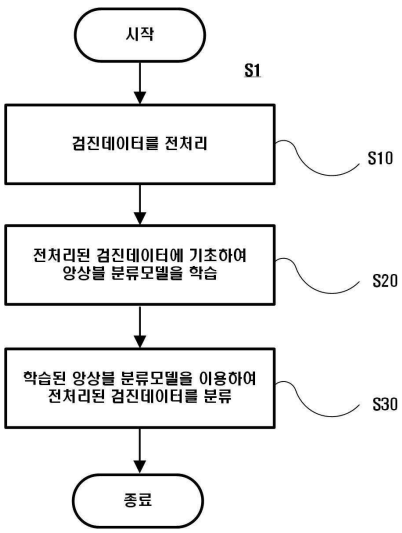
(54) 발명의 명칭 **데이터양 증식 기반의 검진데이터 수치화 장치, 방법및 프로그램**

(57) 요 약

본 발명은 복수의 소견 기반의 검진데이터 분류 방법을 제공하고, 상기 방법은, 검진데이터를 수신하는 단계; 상기 검진데이터를 전처리하는 단계; 전처리된 검진데이터에 기초하여 앙상블 분류모델을 생성하는 단계; 및 학습된 상기 앙상블 분류모델을 이용하여 상기 검진데이터의 분류결과를 획득하는 단계를 포함한다.

또한, 상기 생성하는 단계는, 상기 전처리된 검진데이터에 대한 고정 임베딩(Static Embedding) 및 문맥화 임베딩(Contextualized Embedding)을 수행하고, 상기 고정 임베딩 및 문맥화 임베딩이 수행된 결과 데이터를 기반으로 상기 앙상블 분류모델을 생성한다.

대 표 도 - 도4



(52) CPC특허분류
G06F 16/35 (2019.01)
G06F 18/214 (2023.01)
G06F 40/205 (2020.01)
G06F 40/232 (2020.01)
G06F 40/284 (2020.01)
G06N 20/20 (2021.08)
G16H 50/20 (2018.01)

명 세 서

청구범위

청구항 1

장치에 의해 수행되는, 데이터양 증식 기반의 검진데이터 수치화 방법에 있어서,

검진데이터를 수신하는 단계;

상기 검진데이터를 전처리하는 단계;

전처리된 검진데이터에 기초하여 앙상블 분류모델을 생성하는 단계; 및

학습된 상기 앙상블 분류모델을 이용하여 상기 검진데이터의 분류결과를 획득하는 단계를 포함하고,

상기 검진데이터는, 각각이 상기 복수의 소견 각각에 대응되는 복수의 그룹으로 분류되고,

상기 전처리하는 단계는,

상기 검진데이터에서 기 설정된 기준 특수문자를 제외한 나머지 특수문자를 삭제하는 단계;

기 설정된 기준 문법에 기초하여 상기 특수문자가 삭제된 검진데이터의 띄어 쓰기 규칙을 균일화하는 단계;

상기 균일화된 검진데이터를 수치화하는 단계; 및

상기 수치화된 데이터를 패딩(padding)하는 단계를 포함하고,

상기 수치화하는 단계는,

상기 복수의 그룹 중 데이터의 양이 가장 작은 최소그룹의 데이터의 양을 증식시키는 단계를 더 포함하며,

상기 증식시키는 단계는,

상기 최소그룹에 포함된 데이터를 문장 단위로 분할하여 복수의 문장을 생성하고,

상기 복수의 문장끼리 서로 연결하여 상기 최소그룹의 데이터의 양을 증식시키는 것을 특징으로 하는, 방법.

청구항 2

제1항에 있어서,

상기 생성하는 단계는,

상기 전처리된 검진데이터에 대한 고정 임베딩(Static Embedding) 및 문맥화 임베딩(Contextualized Embedding)을 수행하고,

상기 고정 임베딩 및 문맥화 임베딩이 수행된 결과 데이터를 기반으로 상기 앙상블 분류모델을 생성하는 것인, 방법.

청구항 3

제2항에 있어서,

상기 생성하는 단계는,

고정 임베딩된 데이터에 기초한 학습을 통해 적어도 하나의 제1 분류모델을 생성하는 단계;

문맥화 임베딩된 데이터에 기초한 학습을 통해 적어도 하나의 제2 분류모델을 생성하는 단계; 및

상기 적어도 하나의 제1 분류모델 및 상기 적어도 하나의 제2 분류모델에 기초하여 상기 앙상블 분류모델을 생성하는 단계를 포함하는, 방법.

청구항 4

제3항에 있어서,

상기 적어도 하나의 제2 분류모델을 생성하는 단계는,

기 획득된 코퍼스셋을 이용한 학습을 통해 초기 가중치를 획득하는 단계; 및

문맥화 임베딩된 데이터에 기초한 학습을 통해 상기 초기 가중치를 미세조정(fine-tuning)하여 상기 적어도 하나의 제2 분류모델을 생성하는 단계를 포함하는, 방법.

청구항 5

제4항에 있어서,

상기 검진데이터는, 토크나이징(tokenizing)을 통해 산출되는 토큰(token)의 최대 및 평균 중 적어도 하나의 개수가 기 설정된 기준 개수보다 큰 텍스트데이터를 포함하고,

상기 기준 개수는 상기 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수에 대응되는, 방법.

청구항 6

제5항에 있어서,

상기 검진데이터는, 복수의 진단명 각각에 대해 산출되며,

상기 전처리하는 단계는,

상기 텍스트데이터에서 상기 복수의 진단명 각각에 대응되는 키워드를 검색하는 단계;

상기 검색된 키워드를 포함하고, 토큰의 개수가 상기 기준 개수 이하인 문장을 추출하는 단계; 및

상기 추출된 문장에 기초하여 전처리를 수행하는 단계를 포함하는, 방법.

청구항 7

제6항에 있어서,

상기 검진데이터는 개별소견 텍스트, 수치 검사 결과 및 종합소견 텍스트를 포함하고,

상기 검진데이터는, 복수의 진단명 각각에 대해 산출되고,

상기 수치 검사 결과는 상기 복수의 진단명 각각과 대응되어 기 설정된 정형화된 데이터이고,

상기 종합소견 텍스트는, 토크나이징(tokenizing)을 통해 산출되는 토큰(token)의 최대 및 평균 중 적어도 하나의 개수가 기준 개수보다 크며,

상기 기준 개수는, 상기 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수에 대응되는, 방법.

청구항 8

제7항에 있어서,

상기 수치화하는 단계는,

상기 균일화된 검진데이터를 토크나이징(tokenizing)하는 단계;

상기 토크나이징을 통해 생성된 토큰(token)의 개수를 카운팅하는 단계; 및

상기 토큰의 개수에 기초하여 상기 균일화된 검진데이터를 수치화하는 단계를 포함하는, 방법.

청구항 9

통신부;

데이터베이스부; 및

상기 데이터베이스부에 저장된 검진데이터 또는 상기 통신부를 통해 외부서버에서 수신된 검진데이터를 기 설정된 복수의 소견으로 분류하는 제어부를 포함하고,

상기 검진데이터는, 각각이 상기 복수의 소견 각각에 대응되는 복수의 그룹으로 분류되고,

상기 제어부는,

상기 검진데이터를 전처리하고,

상기 전처리된 검진데이터에 대한 고정 임베딩(Static Embedding) 및 문맥화 임베딩(Contextualized Embedding)을 수행하고,

상기 고정 임베딩 및 문맥화 임베딩이 수행된 결과 데이터를 기반으로 앙상블 분류모델을 생성하며,

학습된 상기 앙상블 분류모델을 이용하여 상기 검진데이터의 분류결과를 획득하고,

상기 검진데이터에서 기 설정된 기준 특수문자를 제외한 나머지 특수문자를 삭제하고,

기 설정된 기준 문법에 기초하여 상기 특수문자가 삭제된 검진데이터의 띄어 쓰기 규칙을 균일화하고,

상기 균일화된 검진데이터를 수치화하고,

상기 수치화된 데이터를 패딩(padding)하고,

상기 복수의 그룹 중 데이터의 양이 가장 작은 최소그룹의 데이터의 양을 증식시키고,

상기 최소그룹에 포함된 데이터를 문장 단위로 분할하여 복수의 문장을 생성하고,

상기 복수의 문장끼리 서로 연결하여 상기 최소그룹의 데이터의 양을 증식시키는 것을 특징으로 하는, 데이터양 증식 기반의 검진데이터 수치화 장치.

청구항 10

하드웨어인 컴퓨터와 결합되어, 제1항 내지 제8항 중 어느 한 항의 방법을 실행시키기 위하여 매체에 저장된, 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 데이터양 증식 기반의 검진데이터 수치화 장치, 방법 및 프로그램에 관한 것이다.

배 경 기 술

[0002] 머신러닝 기술이 발전됨에 따라 다양한 머신러닝 기법이 개발되었으며, 이러한 머신러닝 기법은 다양한 분야에 적용되고 있다.

[0003] 자연어 처리 분야에 있어서도 다양한 머신러닝 기법이 개발되었으며, 문자로 이루어진 데이터에 대하여 보다 정확한 분류를 수행하기 위해 다양한 방법들이 시도되고 있다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 대한민국 등록특허공보 제10-1713487호, 2017. 02. 28 등록

발명의 내용

해결하려는 과제

[0005] 본 발명은, 단어 및 문맥에 대한 특징이 모두 반영된 앙상블 자연어처리 모델을 통해 검진데이터를 분류하는 검진데이터 분류장치 및 분류방법을 제공하는 것을 일 목적으로 한다.

[0006] 또한, 본 발명은, 사전훈련 언어모델을 이용하여 분류 성능이 향상된 앙상블 자연어처리 모델을 제공하는 검진데이터 분류장치 및 분류방법을 제공하는 것을 일 목적으로 한다.

[0007] 또한, 본 발명은, 진단명과 대응되는 키워드를 포함하고 토큰의 개수가 사전훈련에 사용된 코퍼스의 토큰의 최대 개수 이하인 문장을 검진데이터로부터 추출하고, 추출된 문장을 이용해 학습된 앙상블 자연어처리 모델을 제공하는 검진데이터 분류장치 및 분류방법을 제공하는 것을 일 목적으로 한다.

[0008] 또한, 본 발명은, 검진데이터에 적합화된 전처리 과정을 통해 학습된 앙상블 자연어처리 모델을 제공하는 검진데이터 분류장치 및 분류방법을 제공하는 것을 일 목적으로 한다.

[0009] 본 발명이 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상술한 과제를 해결하기 위하여, 본 발명의 일 실시예에 따른 검진데이터 분류방법은, 장치에 의해 수행되는, 복수의 소견 기반의 검진데이터 분류 방법으로서, 검진데이터를 수신하는 단계; 상기 검진데이터를 전처리하는 단계; 전처리된 검진데이터에 기초하여 앙상블 분류모델을 생성하는 단계; 및 학습된 상기 앙상블 분류모델을 이용하여 상기 검진데이터의 분류결과를 획득하는 단계를 포함한다.

[0011] 또한, 상기 생성하는 단계는, 상기 전처리된 검진데이터에 대한 고정 임베딩(Static Embedding) 및 문맥화 임베딩(Contextualized Embedding)을 수행하고, 상기 고정 임베딩 및 문맥화 임베딩이 수행된 결과 데이터를 기반으로 상기 앙상블 분류모델을 생성하는 것이다.

[0012] 또한, 상기 생성하는 단계는, 고정 임베딩된 데이터에 기초한 학습을 통해 적어도 하나의 제1 분류모델을 생성하는 단계; 문맥화 임베딩된 데이터에 기초한 학습을 통해 적어도 하나의 제2 분류모델을 생성하는 단계; 및 상기 적어도 하나의 제1 분류모델 및 상기 적어도 하나의 제2 분류모델에 기초하여 상기 앙상블 분류모델을 생성하는 단계를 포함한다.

[0013] 또한, 상기 적어도 하나의 제2 분류모델을 생성하는 단계는, 기 획득된 코퍼스셋을 이용한 학습을 통해 초기 가중치를 획득하는 단계; 및 문맥화 임베딩된 데이터에 기초한 학습을 통해 상기 초기 가중치를 미세조정(fine-tuning)하여 상기 적어도 하나의 제2 분류모델을 생성하는 단계를 포함한다.

[0014] 또한, 상기 적어도 하나의 제1 분류모델의 학습에는 CNN(Convolution Neural Network) 및 DCNN(Deep Convolution Neural Network) 중 적어도 하나가 사용되고, 상기 적어도 하나의 제2 분류모델의 학습에는 LSTM(Long Short-Term Memory models), KoBERT(Korean Bidirectional Encoder Representations from Transformers), KoELECTRA(Korean Efficiently Learning an Encoder that Classifies Token Replacements Accurately) BERT(Bidirectional Encoder Representations from Transformers) 및 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 중 적어도 하나가 사용될 수 있다.

[0015] 또한, 상기 수신된 검진데이터는, 토크나이징(tokenizing)을 통해 산출되는 토큰(token)의 최대 및 평균 중 적어도 하나의 개수가 기 설정된 기준 개수보다 큰 텍스트데이터를 포함하고, 상기 기준 개수는 상기 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수에 대응된다.

[0016] 또한, 상기 검진데이터는, 복수의 진단명 각각에 대해 산출되며, 상기 전처리하는 단계는, 상기 텍스트데이터에서 상기 복수의 진단명 각각에 대응되는 키워드를 검색하는 단계; 상기 검색된 키워드를 포함하고, 토큰의 개수가 상기 기준 개수 이하인 문장을 추출하는 단계; 및 상기 추출된 문장에 기초하여 전처리를 수행하는 단계를 포함한다.

[0017] 또한, 상기 검진데이터는 개별소견 텍스트, 수치 검사 결과 및 종합소견 텍스트를 포함하고, 상기 검진데이터는, 복수의 진단명 각각에 대해 산출된다.

[0018] 또한, 상기 수치 검사 결과는 상기 복수의 진단명 각각과 대응되어 기 설정된 정형화된 데이터이고, 상기 종합소견 텍스트는, 토크나이징(tokenizing)을 통해 산출되는 토큰(token)의 최대 및 평균 중 적어도 하나의 개수가 기준 개수보다 크며, 상기 기준 개수는, 상기 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수에 대응될 수 있다.

[0019] 또한, 상기 전처리하는 단계는, 상기 검진데이터에서 기 설정된 기준 특수문자를 제외한 나머지 특수문자를 삭제하는 단계; 기 설정된 기준 문법에 기초하여 상기 특수문자가 삭제된 검진데이터의 띄어 쓰기 규칙을 균일화하는 단계; 상기 균일화된 검진데이터를 수치화하는 단계; 및 상기 수치화된 데이터를 패딩(padding)하는 단계를 포함한다.

[0020] 또한, 상기 수치화하는 단계는, 상기 균일화된 검진데이터를 토크나이징(tokenizing)하는 단계; 토크나이징을 통해 생성된 토큰(token)의 개수를 카운팅하는 단계; 및 상기 토큰의 개수에 기초하여 상기 균일화된 검진데이터를 수치화하는 단계를 포함한다.

[0021] 또한, 상기 검진데이터는, 각각이 상기 복수의 소견 각각에 대응되는 복수의 그룹으로 분류되고, 상기 수치화하는 단계는, 상기 복수의 그룹 중 데이터의 양이 가장 작은 최소그룹의 데이터의 양을 증식시키는 단계를 더 포함한다.

[0022] 또한, 상기 증식시키는 단계는, 상기 최소그룹에 포함된 데이터를 문장 단위로 분할하여 복수의 문장을 생성하는 단계; 및 상기 복수의 문장끼리 서로 연결하여 상기 최소그룹의 데이터의 양을 증식시키는 단계를 포함한다.

[0023] 또, 본 발명의 실시 예에 따른 검진데이터 분류장치는, 통신부; 데이터베이스부; 및 상기 데이터베이스부에 저장된 검진데이터 또는 상기 통신부를 통해 외부서버에서 수신된 검진데이터를 기 설정된 복수의 소견으로 분류하는 제어부를 포함한다.

[0024] 또한, 상기 제어부는, 상기 검진데이터를 전처리하고, 상기 전처리된 검진데이터에 대한 고정 임베딩(Static Embedding) 및 문맥화 임베딩(Contextualized Embedding)을 수행하고, 상기 고정 임베딩 및 문맥화 임베딩이 수행된 결과 데이터를 기반으로 상기 앙상블 분류모델을 생성하며, 학습된 상기 앙상블 분류모델을 이용하여 상기 검진데이터의 분류결과를 획득한다.

[0025] 이 외에도, 본 발명을 구현하기 위한 다른 방법, 다른 시스템 및 상기 방법을 실행하기 위한 컴퓨터 프로그램을 기록하는 컴퓨터 판독 가능한 기록 매체가 더 제공될 수 있다.

발명의 효과

[0026] 상기와 같은 본 발명에 따르면 다음과 같은 효과가 도출될 수 있다.

[0027] 먼저, 본 발명에 따르면, 단어 및 문맥에 대한 특징이 모두 반영된 앙상블 자연어처리 모델을 통해 검진데이터가 분류되므로, 검진데이터에 대한 분류 성능이 향상될 수 있다.

[0028] 또한, 본 발명에 따르면, 사전훈련 언어모델을 이용한 앙상블 자연어처리 모델을 통해 검진데이터가 분류되므로, 검진데이터에 대한 분류 성능이 향상될 수 있다.

[0029] 또한, 본 발명에 따르면, 단어에 대한 특징이 반영된 자연어처리 모델이 문맥에 대한 특징이 반영된 자연어처리 모델과 함께 사용되므로, 검진데이터의 토큰의 개수가 문맥의 특징에 대한 사전훈련에 사용된 코퍼스의 토큰의 최대 개수 이상인 경우에도 정확한 분류가 이루어질 수 있다.

[0030] 또한, 본 발명에 따르면, 진단명과 대응되는 키워드를 포함하고 토큰의 개수가 사전훈련에 사용된 코퍼스의 토큰의 최대 개수 이하인 문장을 검진데이터로부터 추출하고, 추출된 문장을 이용해 학습된 앙상블 자연어처리 모델을 통해 검진데이터의 분류가 이루어진다. 이를 통해, 검진데이터의 토큰의 개수가 사전훈련에 사용된 코퍼스의 토큰의 최대 개수 이상인 경우에도 정확한 분류가 이루어질 수 있다.

[0031] 또한, 본 발명에 따르면, 검진데이터에 적합화된 전처리 과정을 통해 학습된 앙상블 자연어처리 모델을 통해 검진데이터가 분류되므로, 검진데이터에 대한 분류 성능이 향상되 수 있다.

[0032] 본 발명의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0033] 도 1은 본 발명의 실시예에 따른 검진데이터 분류시스템의 구성을 도시하는 예시도이다.

도 2는 본 발명의 실시예에 따른 검진데이터 분류장치의 구성을 도시하는 블록도이다.

도 3은 도 2의 제어부의 구성을 도시하는 블록도이다.

도 4는 본 발명의 실시예에 따른 검진데이터 분류방법의 과정을 도시하는 흐름도이다.

도 5는 본 발명의 실시예에 따른 검진데이터 분류방법의 과정을 예시적으로 도시하는 개념도이다.

도 6은 도 4의 S10단계의 일 실시 예의 과정을 도시하는 흐름도이다.

도 7은 도 6의 S13단계의 과정을 도시하는 흐름도이다.

도 8은 도 7의 S133단계에 따라 수치화된 데이터를 예시적으로 도시하는 개념도이다.

도 9는 도 4의 S20단계의 과정을 도시하는 흐름도이다.

도 10은 도 9의 제1 분류모델을 예시적으로 도시하는 개념도이다.

도 11은 도 9의 제2 분류모델을 예시적으로 도시하는 개념도이다.

도 12는 도 9의 S23단계의 과정을 도시하는 흐름도이다.

도 13은 도 4의 앙상블모델을 예시적으로 도시하는 개념도이다.

도 14는 도 4의 S10단계의 다른 실시 예의 과정을 도시하는 흐름도이다.

도 15는 도 14의 S102단계에 따라 추출된 문장을 예시적으로 도시하는 개념도이다.

발명을 실시하기 위한 구체적인 내용

[0034] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야의 통상의 기술자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0035] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0036] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0037] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0038] 설명에 앞서 본 명세서에서 사용하는 용어의 의미를 간략히 설명한다. 그렇지만 용어의 설명은 본 명세서의 이해를 돕기 위한 것이므로, 명시적으로 본 발명을 한정하는 사항으로 기재하지 않은 경우에 본 발명의 기술적 사상을 한정하는 의미로 사용하는 것이 아님을 주의해야 한다.

[0039] 다만, 몇몇 실시예에서 검진데이터 분류장치(10)는 도 2 및 도 3에 도시된 구성요소보다 더 적은 수의 구성요소나 더 많은 구성요소를 포함할 수도 있다.

[0040] 본 명세서에서 설명되는 검진데이터 입력단말(2)에는 예를 들면 휴대폰, 스마트 폰(smart phone), 노트북 컴퓨터(laptop computer), 디지털방송용 단말기, PDA(personal digital assistants), PMP(portable multimedia player), 네비게이션, 슬레이트 PC(slate PC), 태블릿 PC(tablet PC), 울트라북(ultrabook), 웨어러블 디바이스(wearable device, 예를 들어, 와치형 단말기 (smartwatch), 글래스형 단말기 (smart glass), HMD(head mounted display)) 등이 포함될 수 있다.

[0041] 그러나, 본 명세서에 기재된 실시 예에 따른 구성은 단말기에만 적용 가능한 경우를 제외하면, 디지털 TV, 데스크탑 컴퓨터, 디지털 사이니지 등과 같은 고정 단말기에도 적용될 수도 있음을 본 기술분야의 당업자라면 쉽게 알 수 있을 것이다.

[0042] 도 1을 참조하면, 본 발명의 실시 예에 따른 검진데이터 분류 시스템은 서버(1) 및 다수의 검진데이터 입력단말

(2)이 네트워크(3)를 통해서 통신 가능하도록 서로 연결될 수 있다.

[0043] 서버(1)는 다수의 검진데이터 입력단말(2)로부터 수신한 검진데이터에 대한 전처리를 수행하고, 전처리된 검진데이터에 기초하여 앙상블 분류모델을 생성하며, 생성된 앙상블 분류모델을 이용해 검진데이터의 분류결과를 획득한다.

[0044] 도 2를 참조하면, 서버(1)는 검진데이터 분류장치(10)를 포함할 수 있다. 다만, 이에 한정되는 것은 아니며 검진데이터 분류장치(10)는 서버(1)와 별도로 구비될 수 있다.

[0045] 검진데이터 분류장치(10)는 통신부(11), 데이터베이스부(12) 및 제어부(13)를 포함한다. 통신부(11), 데이터베이스부(12) 및 제어부(13)는 서로를 연결하는 시스템 버스에 의해 연결될 수 있다.

[0046] 검진데이터 분류장치(10)와 검진데이터 분류장치(10) 사이, 검진데이터 분류장치(10)와 검진데이터 입력단말(2) 사이 또는 검진데이터 분류장치(10)와 외부 서버 사이의 정보교환은 통신부(11)를 통해 수행될 수 있다.

[0047] 통신부는 유선통신모듈, 무선통신모듈 및 근거리통신모듈 중 적어도 하나를 통해 구현될 수 있다. 무선 인터넷 모듈은 무선 인터넷 접속을 위한 모듈을 말하는 것으로 각 장치에 내장되거나 외장될 수 있다. 무선 인터넷 기술로는 WLAN(Wireless LAN)(Wi-Fi), Wibro(Wireless broadband), Wimax(World Interoperability for Microwave Access), HSDPA(High Speed Downlink Packet Access), LTE(long term evolution), LTE-A(Long Term Evolution-Advanced) 등이 이용될 수 있다.

[0048] 데이터베이스부(12)는 통신부(11)를 통해 수신된 각종 정보, 제어부(13)의 기능수행을 위한 각종 정보, 제어부(13)에서 연산된 각종 정보를 저장한다.

[0049] 데이터베이스부(12)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(random access memory; RAM), SRAM(static random access memory), 롬(read-only memory; ROM), EEPROM(electrically erasable programmable read-only memory), PROM(programmable read-only memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 또한, 데이터베이스부(12)는 웹스토리지 형태로 구현될 수 있다.

[0050] 제어부(13)는 전처리모듈(14) 및 분류모듈(15)을 포함한다.

[0051] 전처리모듈(14)은 수신한 검진데이터를 앙상블 분류모델의 입력데이터로 변환하는 전처리과정을 수행한다.

[0052] 전처리과정은 정형화, 수치화, 패딩 및 데이터증식과정을 포함한다.

[0053] 이를 위하여, 전처리모듈(14)은 정형화 유닛(141), 수치화 유닛(142), 패딩 유닛(143) 및 데이터증식 유닛(144)을 포함한다. 전처리과정에 대해서는 본 발명에 따른 검진데이터 분류방법과 함께 뒤에서 상세히 설명한다.

[0054] 분류모듈(15)은 전처리된 데이터를 임베딩(Embedding)하는 임베딩 유닛(151)을 포함한다.

[0055] 또한, 분류모듈(15)은 임베딩된 데이터를 이용한 학습을 수행하거나 학습된 모델을 이용하여 임베딩된 데이터를 분류하는 분류 유닛(152)을 포함한다.

[0056] 분류 유닛(152)은 앙상블 기법을 이용하여 앙상블 분류모델을 학습시키거나 학습된 앙상블 분류모델을 이용하여 임베딩된 데이터에 대한 분류를 수행한다.

[0057] 일 실시 예에서, 앙상블 기법(Ensemble Learning)에는 보팅(Voting), 배깅(Bagging) 및 부스팅(Boosting) 방법이 사용될 수 있다. 다만, 이에 한정되는 것은 아니다.

[0058] 즉, 분류 유닛(152)은 다양한 알고리즘을 이용한 앙상블 기법을 통해 검진데이터를 분류한다.

[0059] 앙상블 기법에는 CNN(Convolution Neural Network) 및 DCNN(Deep Convolution Neural Network), LSTM(Long Short-Term Memory models), KoBERT(Korean Bidirectional Encoder Representations from Transformers), KoELECTRA(Korean Efficiently Learning an Encoder that Classifies Token Replacements Accurately) BERT(Bidirectional Encoder Representations from Transformers) 및 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 중 적어도 하나의 알고리즘이 사용될 수 있다.

[0060] 일 실시 예에서, 분류 유닛(152)은 CNN(Convolution Neural Network) 및 DCNN(Deep Convolution Neural Network) 중 적어도 하나의 알고리즘을 이용하여 적어도 하나의 제1 분류모델을 생성할 수 있다.

[0061] 일 실시 예에서, 분류 유닛(152)은 LSTM(Long Short-Term Memory models), KoBERT(Korean Bidirectional Encoder Representations from Transformers), KoELECTRA(Korean Efficiently Learning an Encoder that Classifies Token Replacements Accurately) BERT(Bidirectional Encoder Representations from Transformers) 및 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 중 적어도 하나의 알고리즘을 이용하여 적어도 하나의 제2 분류모델을 생성할 수 있다.

[0062] 일 실시 예에서, 분류 유닛(152)은 제1 분류모델과 제2 분류모델을 이용하여 앙상블 분류모델을 생성할 수 있다.

[0063] 제어부(13)는, 하드웨어적으로, ASICs(applicationspecific integrated circuits), DSPs(digital signal processors), DSPDs(digital signal processing devices), PLDs(programmable logic devices), FPGAs(field programmable gate arrays), 프로세서(processors), 제어기(controllers), 마이크로컨트롤러(micro-controllers), 마이크로 프로세서(microprocessors), 기타 기능 수행을 위한 전기적인 유닛 중 적어도 하나를 이용하여 구현될 수 있다.

[0064] 또한, 소프트웨어적으로, 본 명세서에서 설명되는 절차 및 기능과 같은 실시 예들은 별도의 소프트웨어 모듈들로 구현될 수 있다. 상기 소프트웨어 모듈들 각각은 본 명세서에서 설명되는 하나 이상의 기능 및 작동을 수행할 수 있다. 소프트웨어 코드는 적절한 프로그램 언어로 쓰여진 소프트웨어 애플리케이션으로 소프트웨어 코드가 구현될 수 있다. 상기 소프트웨어 코드는 메모리에 저장되고, 제어부(13)에 의해 실행될 수 있다.

[0066] 아래에서는 도 4 내지 도 15를 참조하여 본 발명의 실시 예에 따른 검진데이터 분류방법(S1)에 대해서 상세히 설명한다.

[0067] 도 4를 참조하면, 검진데이터 분류방법(S1)은 검진데이터를 전처리하는 단계(S10), 전처리된 검진데이터에 기초하여 앙상블 분류모델을 학습하는 단계(S20) 및 학습된 앙상블 분류모델을 이용하여 전처리된 검진데이터를 분류하는 단계(S30)를 포함한다.

[0068] 도 5를 참조하면, 검진데이터 분류장치(10)가 갑상선에 대한 검진데이터에 대한 분류모델을 생성하고 생성된 분류모델을 이용해 기 설정된 복수의 소견에 따른 분류결과를 도출하는 과정이 개념적으로 도시된다.

[0069] 도시된 실시 예에서, 검진데이터는 건강검진데이터로부터 도출될 수 있다. 건강검진데이터는 건강검진에서 확인될 수 있는 각종 진단명에 대한 개별소견 텍스트, 각종 이상과 관련된 수치 검사 결과 및 종합소견 텍스트로 구성될 수 있다.

[0070] 검진데이터는 특정 진단명에 대한 개별소견 텍스트, 특정 진단명과 관련된 수치 검사 결과 및 종합소견 텍스트로 구성될 수 있다. 일 실시 예에서, 검진데이터는 개별소견 텍스트, 수치 검사 결과 및 종합소견 텍스트를 연결한 자연어의 나열로 구성될 수 있다.

[0071] 도시된 실시 예에서, 검진데이터는 갑상선에 대한 것이며, 갑상선 개별소견 텍스트, 호르몬 수치 검사 결과 및 종합소견 텍스트로 구성된다. 도시되지 않은 실시 예에서, 검진데이터는 갑상선 외에 다른 진단명에 대한 검진데이터로 구성될 수 있다.

[0072] 검진데이터가 검진데이터 분류장치(10)로 수신되면, 검진데이터 분류장치(10)의 전처리모듈(14)이 검진데이터를 전처리한다(S10).

[0073] 전처리가 완료되면, 검진데이터 분류장치(10)의 분류모듈(15)은 전처리된 검진데이터에 기초하여 앙상블 분류모델을 학습한다(S20).

[0074] 학습이 완료되면, 검진데이터 분류장치(10)의 분류모듈(15)은 학습된 앙상블 분류모델을 이용하여 전처리된 검사데이터에 대한 분류결과를 획득한다(S30).

[0075] 분류결과는 기 설정된 복수의 소견으로 분류된 결과이며, 일 실시 예에서, 복수의 소견은 정상(Class 0) 및 유소견(Class 1)일 수 있다. 또한, 일 실시 예에서, 복수의 소견은 정상(Class 0), 유소견(Class 1) 및 중증유소견(Class 2)일 수 있다.

[0076] 아래에서는, 도 6 내지 도 8을 참조하여, 검진데이터를 전처리하는 단계(S10)에 대해 구체적으로 설명한다.

[0078] (1) 검진데이터를 전처리하는 단계(S10)의 설명

[0079] 도 6을 참조하면, 검진데이터를 전처리하는 단계(S10)는 기 설정된 특수문자를 제외한 나머지 특수문자를 삭제하는 단계(S11), 띄어쓰기 규칙을 균일화하는 단계(S12), 데이터를 수치화하는 단계(S13) 및 데이터를 패딩(padding)하는 단계(S14)를 포함한다.

[0080] 먼저, 검진데이터 분류장치(10)가 수신된 검진데이터에서 기 설정된 특수문자를 제외한 나머지 특수문자를 삭제한다(S11). 특수문자의 삭제는 전처리모듈(14)의 정형화 유닛(141)에 의해 수행될 수 있다.

[0081] 검진데이터에는 측정단위와 같은 특수문자가 많이 포함되어 있어 검진데이터의 시퀀스(sequence)가 과도하게 증가될 수 있다. 따라서, 검진데이터의 시퀀스를 감소시키기 위해 기 설정된 특수문자를 제외한 나머지 특수문자를 삭제한다.

[0082] 일 실시 예에서, 기 설정된 특수문자는 " , . & ' / ~ ^{2} -"와 같은 8개의 특수문자일 수 있다.

[0083] 또한, 검진데이터에 포함된 수치 검사 결과는 진단명과 매칭되어 정형화된 형태의 수치데이터이며, 정형화 유닛(141)은 정형화된 형태의 수치데이터를 텍스트 데이터로 변환시킬 수 있다.

[0084] 예를 들어, 검진데이터가 갑상선에 연관된 경우, 수치 결과 데이터는 호르몬 수치에 대한 데이터일 수 있다. 정형화 유닛(141)은 호르몬 수치에 대한 데이터를 "호르몬 검사 수치 정상입니다."와 같은 텍스트 형태로 변환시킬 수 있다.

[0085] 특수문자 삭제가 완료되면, 검진데이터 분류장치(10)가 기 설정된 문법규칙에 기초하여 검진데이터에 포함된 텍스트의 띄어쓰기 규칙을 획일화시킨다(S12). 띄어쓰기 규칙의 획일화는 정형화 유닛(141)에 의해 수행될 수 있다.

[0086] 작성하는 사람에 따라 검진데이터에 포함된 텍스트에는 띄어쓰기 규칙에 오류가 있거나 서로 다른 형태의 띄어쓰기 규칙이 사용될 수 있다. 예를 들어, "할수 있다", "할수있다" 및 "할 수 있다"가 함께 포함되어 있을 수 있다. 해당 텍스트들은 모두 같은 내용의 자연어 이므로, 이들을 기 설정된 띄어쓰기 규칙에 기초하여 통일시킬 수 있다. 예를 들어, "할수 있다", "할수있다"를 모두 "할 수 있다"로 통일시킬 수 있다.

[0087] 일 실시 예에서, 띄어쓰기 외에도 다양한 문법규칙이 획일화될 수 있다.

[0088] 띄어쓰기 규칙의 획일화에는 다양한 토크나이저(tokenizer)가 사용될 수 있다. 일 실시 예에서, Mecab-Ko, WordPiece from KoBERT, WordPiece from KoELECRA, Mecab-Ko & WordPiece 중 적어도 하나의 토크나이저가 사용될 수 있다. 토크나이저에 의해 검진데이터에 포함된 텍스트가 복수의 토큰(token)으로 분할된다.

[0089] 띄어쓰기 규칙에 대한 획일화가 완료되면, 검진데이터 분류장치(10)는 획일화된 검진데이터를 수치화한다(S13).

[0090] 도 7을 참조하면, 검진데이터를 수치화하는 단계(S13)의 구체적인 과정이 도시된다.

[0091] 먼저, 수치화 유닛(142)에 의해 검진데이터에 대한 토크나이징이 수행된다(S131). 도시되지 않은 실시 예에서, 수치화 유닛(142)은 별도로 토크나이징을 수행하지 않고 정형화 유닛(141)에서 토크나이징된 결과를 사용할 수 있다.

[0092] 토크나이징에 의해 검진데이터가 복수의 토큰으로 분리되면, 수치화 유닛(142)이 토큰의 개수를 카운팅한다(S132).

[0093] 카운팅이 완료되면, 수치화 유닛(142)이 토큰의 개수에 기초하여 데이터를 수치화한다(S133).

[0094] 도 8을 참조하면, 상기 S133단계에 의해 수치화된 결과가 예시적으로 도시된다.

[0095] "호르몬" 토큰은 검진데이터에 포함된 토큰 중 13023번째로 배치된 토큰이고, "검사" 토큰은 검진데이터에 포함된 토큰 중 6911번째로 배치된 토큰이다. 즉, 토큰의 배열 순서에 따라 수치화가 이루어진다.

[0096] 다시 도 6을 참조하면, 수치화가 완료됨에 따라 패딩 유닛(143)이 수치화가 완료된 검진데이터에 대한 패딩(padding)을 수행한다.

[0097] 패딩 유닛(143)은 복수의 수치화된 검진데이터의 시퀀스(sequence)를 비교하고, 복수의 수치화된 검진데이터의 최대 시퀀스를 확인한다.

[0098] 또한, 최대 시퀀스가 확인되면, 확인된 최대 시퀀스에 기초하여 복수의 수치화된 검진데이터의 시퀀스를 증가시

킨다. 즉, 복수의 수치화된 검진데이터 모두의 길이를 통일시킨다.

[0099] 상술한 과정을 통해 검진데이터에 대한 전처리가 완료된다.

[0100] 전처리가 완료되면, 전처리모듈(14)의 데이터증식 유닛(144)은 추가적으로 데이터 증식을 수행할 수 있다.

[0101] 앙상블 분류모델의 학습과정에서 사용되는 검진데이터는, 각각이 기 설정된 복수의 소견 각각에 대응되는 복수의 그룹으로 분류될 수 있다. 즉, 학습과정에서 사용되는 검진데이터는 복수의 그룹으로 라벨링될 수 있다.

[0102] 예를 들어, 학습과정에 사용되는 검진데이터는 정상(Class 0), 유소견(Class 1) 및 중증유소견(Class 2)으로 분류될 수 있다.

[0103] 다만, 분류된 특정 그룹의 데이터양이 너무 작은 경우 해당 그룹의 특징이 학습모델에 제대로 반영되지 않는 문제점이 발생될 수 있다. 따라서, 데이터증식 유닛(144)은 데이터 양이 과도하게 적은 그룹에 대한 데이터 증식을 수행할 수 있다.

[0104] 데이터증식 유닛(144)은 데이터량이 가장 적은 최소그룹에 포함된 데이터를 문장 단위로 분할하여 복수의 문장을 생성하고, 생성된 복수의 문장끼리 서로 연결하여 최소그룹의 데이터의 양을 증식시킬 수 있다.

[0105] 예를 들어, 제1 문장, 제2 문장 및 제3 문장으로 분류된 경우, 제1 및 제2 문장의 결합, 제2 및 제3 문장의 결합, 제1, 제2 및 제3 문장 결합을 통해 데이터의 양을 증식시킬 수 있다.

[0106] 특수문자 생략, 띄어쓰기 규칙 균일화, 데이터 수치화 및 데이터 패딩 등 검진데이터에 적합화된 전처리과정을 통해 앙상블 분류모델이 생성되므로, 생성된 앙상블 분류모델의 분류 정확도가 향상될 수 있다.

[0108] (2) 전처리된 검진데이터에 기초하여 앙상블 분류모델을 학습하는 단계(S20)의 설명

[0109] 도 9를 참조하면, 전처리된 검진데이터에 기초하여 앙상블 분류모델을 학습하는 단계(S20)의 구체적인 과정이 도시된다.

[0110] 먼저, 검진데이터 분류장치(10)가 전처리된 검진데이터에 대하여 고정 임베딩(Static Embedding) 및 문맥화 임베딩(Contextualized Embedding)을 수행한다. 임베딩을 수행하는 과정은 제어부(13)의 임베딩 유닛(151)에 의해 수행될 수 있다.

[0111] 도 10을 참조하면, 고정 임베딩이 예시적으로 도시된다.

[0112] 도시된 실시 예에서, 검사데이터에 포함된 복수의 토큰 중 일부 토큰에 대하여 기 설정된 윈도우 사이즈에 기초한 고정 임베딩이 수행된다.

[0113] 도시된 실시 예에서, 상측에 도시된 모델에서는 “혈압”, ”이” 및 ”평균” 토큰에 대하여 고정 임베딩이 수행되고, “약간”, “높”, “습니다” 및 “.” 토큰에 대하여 고정 임베딩이 수행된다.

[0114] 도시된 실시 예에서, 하측에 도시된 모델에서는 “혈압”,”이” 및 ”평균” 토큰에 대하여 고정 임베딩이 수행되고, “수치”, “검사” 및 “결과” 토큰에 대하여 고정 임베딩이 수행된다.

[0115] 일 실시 예에서, 고정 임베딩에는 일정한 크기의 윈도우가 사용될 수 있으며, 다양한 크기위 윈도우가 복합적으로 사용될 수 있다.

[0116] 도 11을 참조하면, 문맥화 임베딩이 예시적으로 도시된다.

[0117] 도시된 실시 예에서, 문맥과 관련된 특징을 고려하기 위하여 기 설정된 개수 이상의 토큰에 대하여 문맥화 임베딩을 수행한다.

[0118] 일 실시 예에서, 문맥화 임베딩은 토큰 임베딩, 새그먼트 임베딩 및 포지션 임베딩을 포함할 수 있다.

[0119] 토큰 임베딩(Token Embedding)은 각 문자 단위로 임배딩을 하고, 자주 등장하면서 가장 긴 길이의 sub-word를 하나의 단위로 만든다. 자주 등장하지 않는 단어를 OOV(Out Of Vacabulary)처리하여 모델링 성능 저하를 방지할 수 있다.

[0120] 또한, 세그먼트 임베딩(Segment Embedding)은 토큰 시킨 단어들을 다시 하나의 문장으로 구성한다. 두 개의 문장 사이에는 구분자 [SEP]를 활용하고 그 두 문장을 하나의 Segment로 지정하여 입력한다.

[0121] 포지션 임베딩(Position Embedding)은 토큰의 순차적으로 인코딩한다.

[0122] 다시 도 9를 참조하면, 검진데이터 분류장치(10)는 고정 임베딩이 수행된 결과 데이터에 기초한 학습을 통해 제1 분류모델을 생성한다(S22).

[0123] 일 실시 예에서, 분류 유닛(152)은 CNN(Convolution Neural Network) 및 DCNN(Deep Convolution Neural Network) 중 적어도 하나의 알고리즘을 이용하여 적어도 하나의 제1 분류모델을 생성할 수 있다.

[0124] 도 10을 참조하면, CNN 및 DCNN 알고리즘에 의해 제1 분류모델이 생성되는 과정이 개념적으로 도시된다.

[0125] 다시 도 9를 참조하면, 검진데이터 분류장치(10)는 문맥화 임베딩이 수행된 결과 데이터에 기초한 학습을 통해 제2 분류모델을 생성한다(S23).

[0126] 일 실시 예에서, 분류 유닛(152)은 LSTM(Long Short-Term Memory models), KoBERT(Korean Bidirectional Encoder Representations from Transformers), KoELECTRA(Korean Efficiently Learning an Encoder that Classifies Token Replacements Accurately) BERT(Bidirectional Encoder Representations from Transformers) 및 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 중 적어도 하나의 알고리즘을 이용하여 적어도 하나의 제2 분류모델을 생성할 수 있다.

[0127] 도 11을 참조하면, BERT(Bidirectional Encoder Representations from Transformers) 알고리즘에 의해 제2 분류모델이 생성되는 과정이 개념적으로 도시된다.

[0128] 문맥의 특징이 반영된 제2 분류모델이 제1 분류모델과 함께 사용되므로, 앙상블 분리모델의 분류성능이 향상될 수 있다.

[0129] 도 12를 참조하면, 제2 분류모델을 생성하는 단계(S23)의 일 실시 예의 구체적인 과정이 도시된다.

[0130] 먼저, 검진데이터 분류장치(10)는 기 획득된 코퍼스셋을 이용한 학습을 통해 초기 가중치를 획득한다(S231).

[0131] 초기 가중치 획득이 완료되면, 검진데이터 분류장치(10)는 문백화 임베딩이 수행된 결과 데이터에 기초한 학습을 통해 초기 가중치를 미세조정(fine-tuning)하여 제2 분류모델을 생성한다(S232).

[0132] 일 실시 예에서, KoBERT(Korean Bidirectional Encoder Representations from Transformers), KoELECTRA(Korean Efficiently Learning an Encoder that Classifies Token Replacements Accurately) BERT(Bidirectional Encoder Representations from Transformers) 및 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately) 중 적어도 하나의 알고리즘이 사용될 수 있다.

[0133] 기 획득된 코퍼스셋을 이용하여 사전학습을 수행되므로, 상대적으로 적은 양의 검진데이터를 이용하여 높은 정확도의 분류성능을 갖는 분류모델을 생성될 수 있다.

[0134] 다만, 기 획득된 코퍼스셋에 의해서 학습된 경우, 미세조정 또는 분류과정에서 제2 분류모델에는 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수보다 짧은 길이의 데이터가 입력되야하는 문제점이 존재했다.

[0135] 따라서, 검진데이터에 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수보다 큰 길이의 데이터가 포함되는 경우, 이를 분할하여 사용해야하며, 분할된 데이터에 원하는 키워드가 포함되지 않을 수 있다. 특히, 검사데이터의 종합소견 텍스트의 경우 그 길이가 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수보다 긴 경우가 빈번하였다.

[0136] 본 발명에 따르면, 코퍼스의 토큰의 최대 개수에 영향을 받지 않는 제1 분류 모델이 함께 사용되므로, 검사데이터의 길이에 무관하게 정확한 분류가 이루어질 수 있다.

[0137] 일 실시 예에서, 검진데이터는, 토크나이징(tokenizing)을 통해 산출되는 토큰(token)의 최대 및 평균 중 적어도 하나의 개수가 기 설정된 기준 개수보다 큰 텍스트데이터를 포함하고, 기준 개수는 상기 기 획득된 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수에 대응될 수 있다. 예를 들어, 종합소견 텍스트의 토큰의 최대 및 평균 중 적어도 하나의 개수가 기준 개수보다 클 수 있다.

[0138] 다시 도 9를 참조하면, 검진데이터 분류장치(10)가 생성된 적어도 하나의 제1 분류모델 및 적어도 하나의 제2 분류모델에 기초하여 앙상블 분류모델을 생성한다(S24).

[0139] 일 실시 예에서, 앙상블 기법(Ensemble Learning)에는 보팅(Voting), 배깅(Bagging) 및 부스팅(Boosting) 방법이 사용될 수 있다. 다만, 이에 한정되는 것은 아니다.

[0141] (3) 학습된 앙상블 분류모델을 이용하여 전처리된 검진데이터를 분류하는 단계(S30)의 설명

[0142] 다시 도 4를 참조하면, 앙상블 분류모델이 생성되면 검진데이터 분류장치(10)는 전처리된 검진데이터를 분류한다(S30).

[0143] 도 13을 참조하면, DCNN 알고리즘에 의해 생성된 제1 분류모델과 사전학습을 이용한 KoELECTRA 알고리즘에 의해 생성된 제2 분류모델을 함께 사용하는 앙상블 분류모델에 대한 테스트 결과가 도시된다.

[0144] 도시된 실시 예에서, 복수의 제1 분류모델, 복수의 제2 분류모델 및 앙상블 분류모델에 대한 정밀도(Macro P) 및 재현율(Macro R)에 기초해 산출되는 F1 스코어(Macro F1)가 도시된다. DCNN 및 KoELECTRA 알고리즘을 통한 앙상블 분류모델의 정밀도, 재현율 및 F1 스코어가 제1 분류모델 및 제2 분류모델만을 사용하는 것에 비해 향상됨이 도시된다.

[0145] 일 실시 예에서, 앙상블 분류모델에는 CNN, DCNN 및 LSTM 등의 알고리즘을 이용한 제1 및 제2 분류모델과 사전훈련을 이용한 방식으로 학습된 제2 분류모델(KoELECTRA 등)이 함께 사용될 수 있다.

[0146] 이를 통해, 사전훈련을 이용한 방식으로 학습된 제2 분류모델(KoELECTRA 등)을 사용함으로써 분류도의 정확성이 향상됨과 동시에, 검사데이터의 길이에 제약받지 않고 정확한 분류를 수행할 수 있다.

[0148] (4) 다른 실시 예에 따른 검진데이터를 전처리하는 단계(S10)의 설명

[0149] 도 14를 참조하면, 다른 실시 예에 따른 검진데이터를 전처리하는 단계(S10)의 과정이 도시된다.

[0150] 먼저, 검진데이터 분류장치(10)는 코퍼스셋에 포함된 코퍼스의 토큰의 최대 개수인 기준 개수에 기초하여 검진데이터를 재가공하고, 재가공된 검진데이터에 대하여 전처리를 수행할 수 있다.

[0151] 특정 진단명에 대응되는 검진데이터에 대한 학습 및 분류를 수행하는 경우, 검진데이터 분류장치(10)는 종합소견 텍스트로부터 소정 요건을 만족시키는 문장을 추출하고, 개별소견 텍스트, 수치 검사 결과 및 추출된 문장에 대하여 전처리를 수행할 수 있다.

[0152] 먼저, 검진데이터 분류장치(10)는 종합소견 텍스트에서 진단명에 대응되는 키워드를 검색한다(S101).

[0153] 예를 들어, 개별소견 텍스트가 갑상선에 관련된 경우, 갑상선과 관련하여 기 설정된 키워드에 대해 검색할 수 있다.

[0154] 키워드가 검색되면, 검진데이터 분류장치(10)는 검색된 키워드를 포함하고 토큰의 개수가 기준 개수 이하인 문장을 추출한다(S102).

[0155] 도 15를 참조하면, 추출된 문장이 예시적으로 도시된다.

[0156] 다시 도 14를 참조하면, 검진데이터 분류장치(10)는 추출된 문장에 기초하여 전처리를 수행한다(S103).

[0157] 일 실시 예에서, 검진데이터 분류장치(10)는 개별소견 텍스트, 수치 검사 결과 및 추출된 문장에 대하여 전처리를 수행한다.

[0158] 이를 통해, 검진데이터의 토큰의 개수가 사전훈련에 사용된 코퍼스의 토큰의 최대 개수 이상인 경우에도 정확한 학습 및 분류가 이루어질 수 있다.

[0160] 이상으로 설명한 본 발명의 실시예에 따른 검진데이터 분류방법은 도 1 내지 도 3를 통해 설명한 검진데이터 분류 시스템 및 장치와 발명의 카테고리만 다를 뿐, 동일한 내용이므로 중복되는 설명, 예시는 생략하도록 한다.

[0161] 이상에서 전술한 본 발명의 실시예에 따른 방법은, 하드웨어인 서버와 결합되어 실행되기 위해 프로그램(또는 어플리케이션)으로 구현되어 매체에 저장될 수 있다.

[0162] 상기 전술한 프로그램은, 상기 컴퓨터가 프로그램을 읽어 들여 프로그램으로 구현된 상기 방법들을 실행시키기 위하여, 상기 컴퓨터의 프로세서(CPU)가 상기 컴퓨터의 장치 인터페이스를 통해 읽힐 수 있는 C, C++, JAVA, 기계어 등의 컴퓨터 언어로 코드화된 코드(Code)를 포함할 수 있다. 이러한 코드는 상기 방법들을 실행하는 필요

한 기능들을 정의한 함수 등과 관련된 기능적인 코드(Functional Code)를 포함할 수 있고, 상기 기능들을 상기 컴퓨터의 프로세서가 소정의 절차대로 실행시키는데 필요한 실행 절차 관련 제어 코드를 포함할 수 있다. 또한, 이러한 코드는 상기 기능들을 상기 컴퓨터의 프로세서가 실행시키는데 필요한 추가 정보나 미디어가 상기 컴퓨터의 내부 또는 외부 데이터베이스부의 어느 위치(주소 번지)에서 참조되어야 하는지에 대한 데이터베이스부 참조관련 코드를 더 포함할 수 있다. 또한, 상기 컴퓨터의 프로세서가 상기 기능들을 실행시키기 위하여 원격(Remote)에 있는 어떠한 다른 컴퓨터나 서버 등과 통신이 필요한 경우, 코드는 상기 컴퓨터의 통신 모듈을 이용하여 원격에 있는 어떠한 다른 컴퓨터나 서버 등과 어떻게 통신해야 하는지, 통신 시 어떠한 정보나 미디어를 송수신해야 하는지 등에 대한 통신 관련 코드를 더 포함할 수 있다.

[0163] 상기 저장되는 매체는, 레지스터, 캐쉬, 데이터베이스부 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상기 저장되는 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등이 있지만, 이에 제한되지 않는다. 즉, 상기 프로그램은 상기 컴퓨터가 접속할 수 있는 다양한 서버 상의 다양한 기록매체 또는 사용자의 상기 컴퓨터상의 다양한 기록매체에 저장될 수 있다. 또한, 상기 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장될 수 있다.

[0164] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 데이터베이스부(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

[0165] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

부호의 설명

[0166] 1: 서버

2: 검진데이터 입력단말

3: 네트워크

10: 검진데이터 분류장치

11: 통신부

12: 데이터베이스부

13: 제어부

14: 전처리모듈

141: 정형화 유닛

142: 수치화 유닛

143: 패딩 유닛

144: 데이터증식 유닛

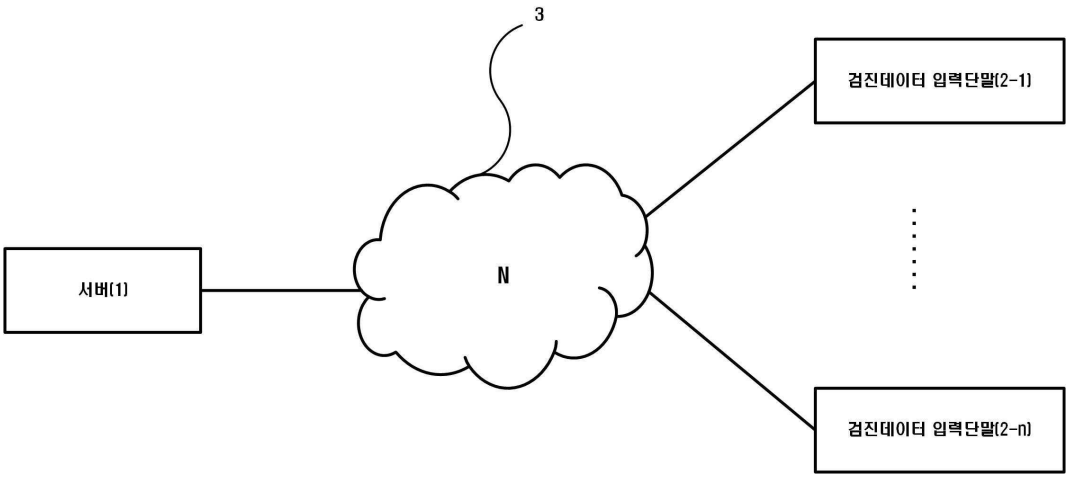
15: 분류모듈

151: 임베딩 유닛

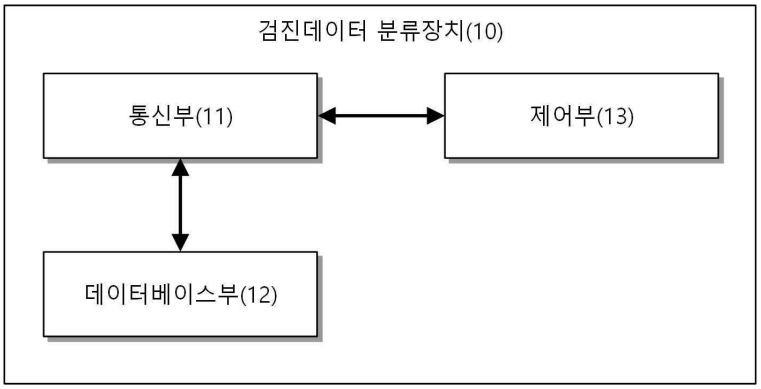
152: 분류 유닛

도면

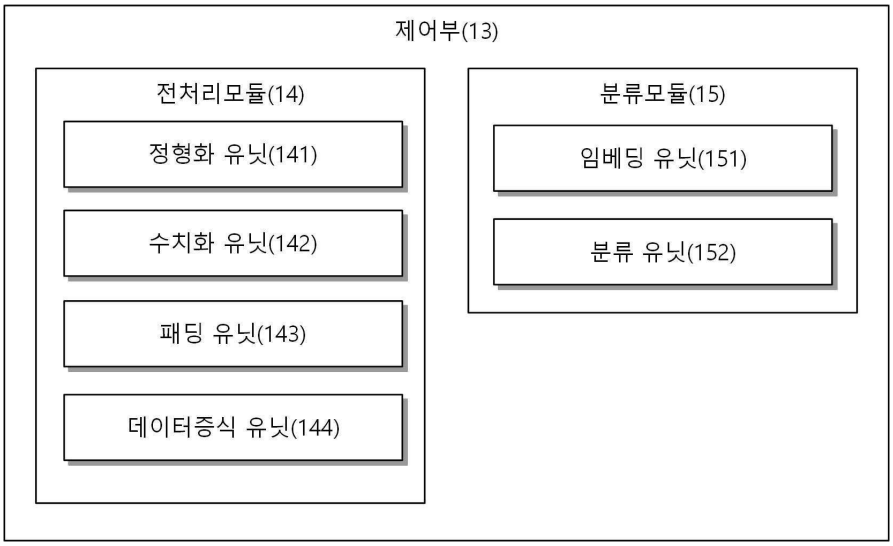
도면1



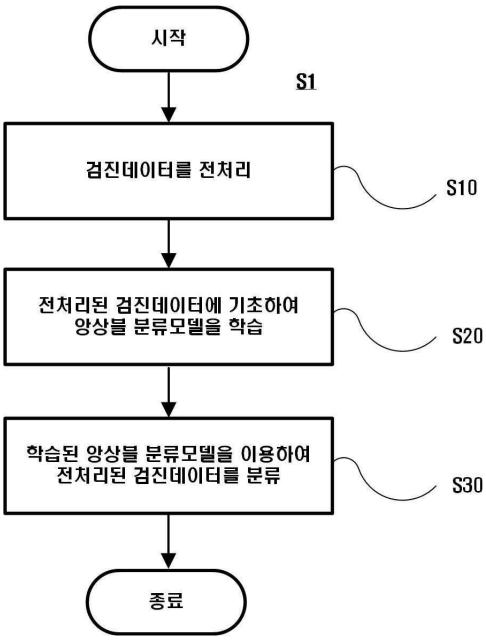
도면2



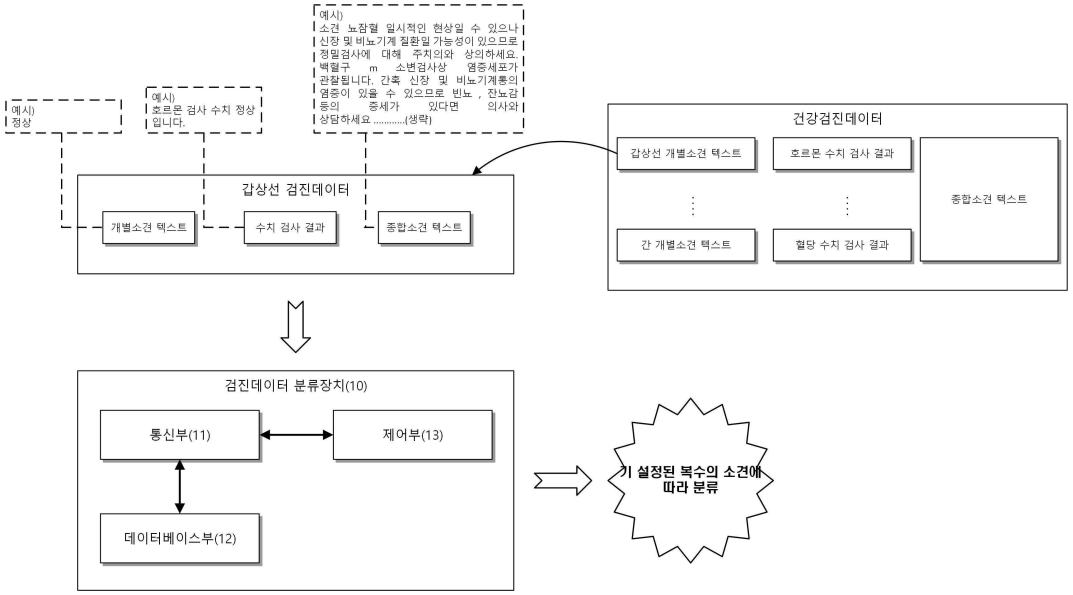
도면3



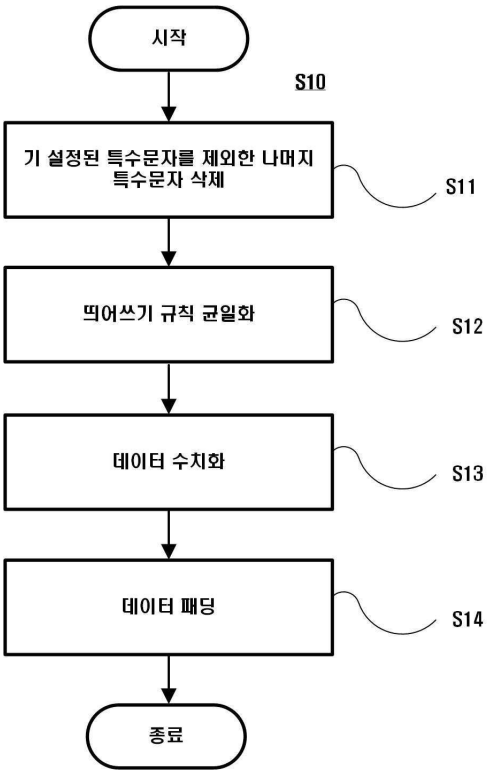
도면4


시작
S1
검진데이터를 전처리
S10
전처리된 검진데이터에 기초하여 앙상블 분류모델을 학습
S20
학습된 앙상블 분류모델을 이용하여 전처리된 검진데이터를 분류
S30
종료

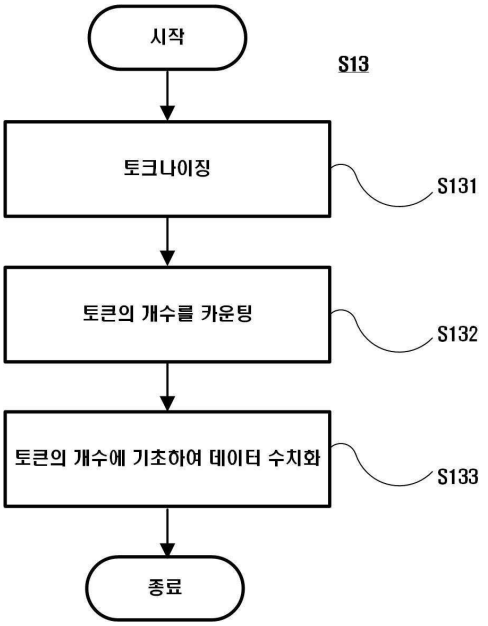
도면5


예시)
정상
예시)
호르몬 검사 수치 정상입니다.
예시)
소견 노잠혈 일시적인 현상일 수 있으나 신장 및 비뇨기계 질환일 가능성이 있으므로 정밀검사에 대해 주치의와 상의하세요. 백혈구 m 소변검사상 염증세포가 관찰됩니다. 간혹 신장 및 비뇨기계통의 염증이 있을 수 있으므로 빈뇨 , 잔뇨감 등의 증세가 있다면 의사와 상담하세요 .....(생략)
건강검진데이터
갑상선 개별소견 텍스트
호르몬 수치 검사 결과
종합소견 텍스트
간 개별소견 텍스트
혈당 수치 검사 결과
갑상선 검진데이터
개별소견 텍스트
수치 검사 결과
종합소견 텍스트
검진데이터 분류장치(10)
통신부(11)
제어부(13)
데이터베이스부(12)
기 설정된 복수의 소견에 따라 분류

도면6



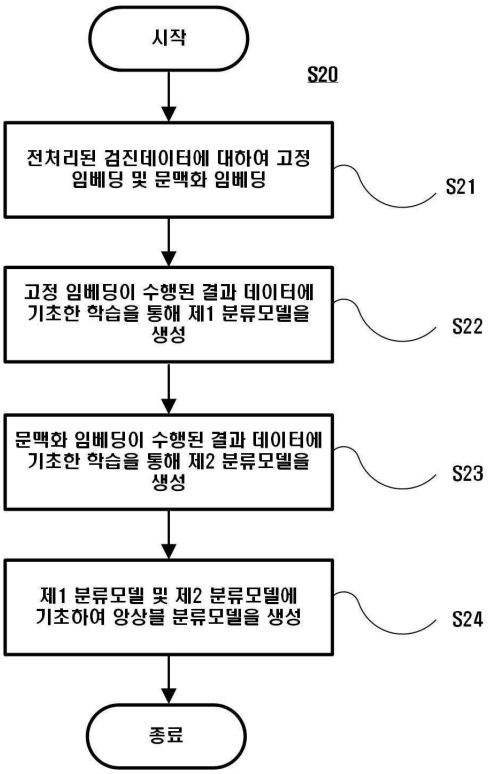
도면7



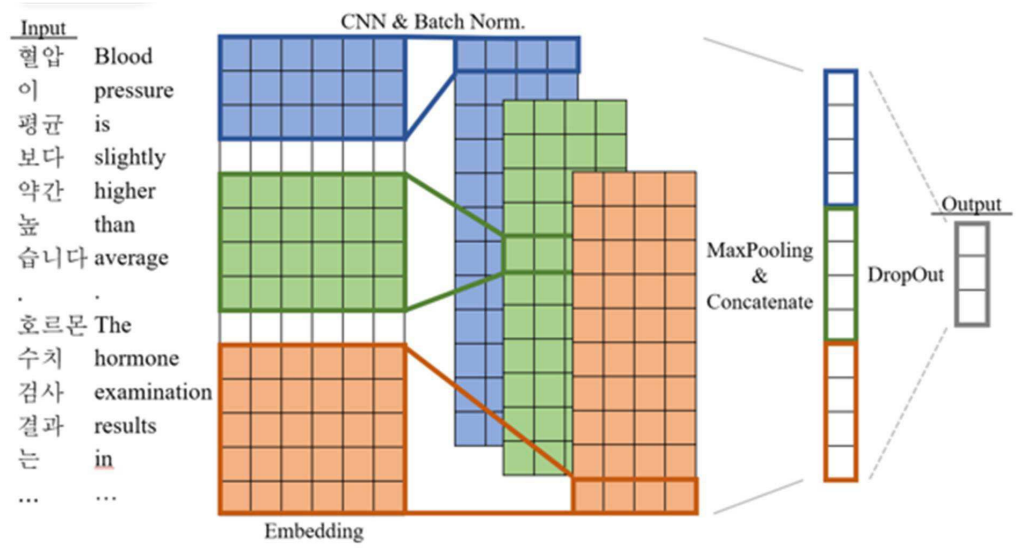
도면8

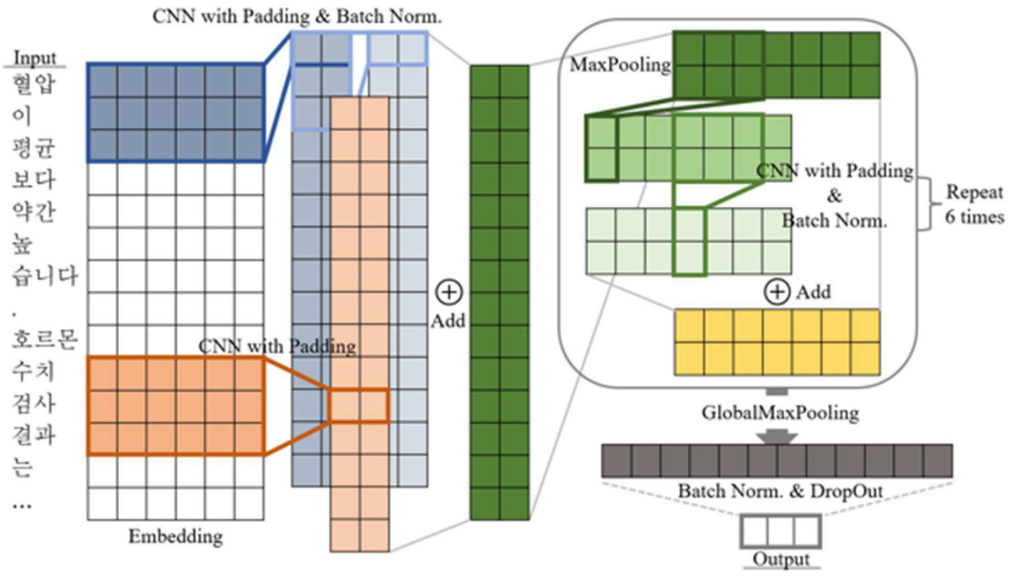
토큰	' 호르몬'	' 검사'	' 수치'	' 정상'	' 입니다'
수치화	13023	6911	8113	6648	6286

도면9

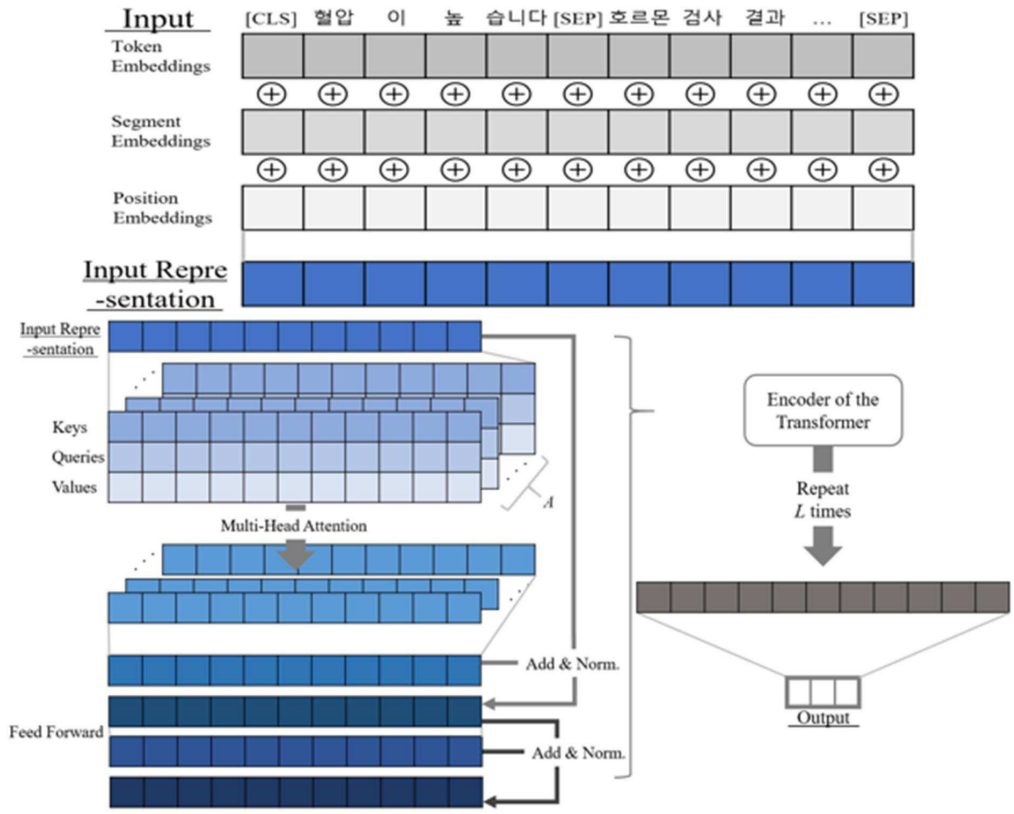

시작
S20
전처리된 검진데이터에 대하여 고정 임베딩 및 문맥화 임베딩
S21
고정 임베딩이 수행된 결과 데이터에 기초한 학습을 통해 제1 분류모델을 생성
S22
문맥화 임베딩이 수행된 결과 데이터에 기초한 학습을 통해 제2 분류모델을 생성
S23
제1 분류모델 및 제2 분류모델에 기초하여 앙상블 분류모델을 생성
S24
종료

도면10

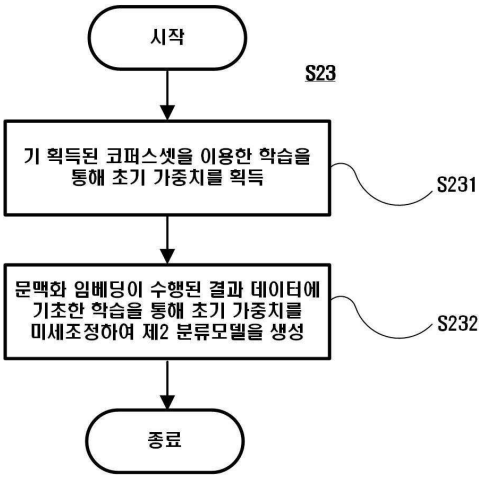

Input
혈압 Blood
이 pressure
평균 is
보다 slightly
약간 higher
높 than
습니다 average
. .
호르몬 The
수치 hormone
검사 examination
결과 results
는 in
... ...
CNN & Batch Norm.
Embedding
MaxPooling & Concatenate
DropOut
Output


CNN with Padding & Batch Norm.
Input
혈압
이
평균
보다
약간
높
습니다
.
호르몬
수치
검사
결과
는
...
CNN with Padding
Embedding
Add
MaxPooling
CNN with Padding & Batch Norm.
Repeat 6 times
Add
GlobalMaxPooling
Batch Norm. & DropOut
Output

도면11


Input
[CLS] 혈압 이 높 습니다 [SEP] 호르몬 검사 결과 ... [SEP]
Token Embeddings
Segment Embeddings
Position Embeddings
Input Repre -sentation
Input Repre -sentation
Keys
Queries
Values
A
Multi-Head Attention
Add & Norm.
Feed Forward
Add & Norm.
Encoder of the Transformer
Repeat L times
Output

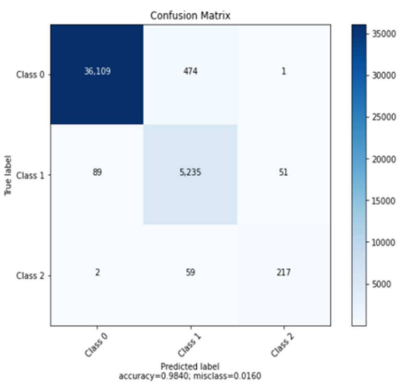
도면12


시작
S23
기 획득된 코퍼스셋을 이용한 학습을 통해 초기 가중치를 획득
S231
문맥화 임베딩이 수행된 결과 데이터에 기초한 학습을 통해 초기 가중치를 미세조정하여 제2 분류모델을 생성
S232
종료

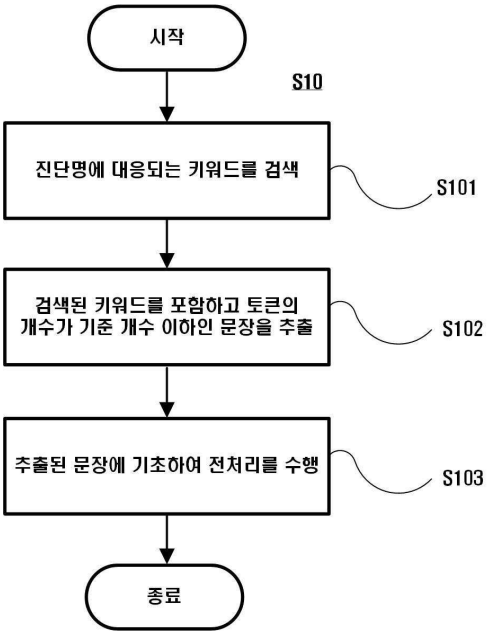
도면13

실험 모형	토크나이저	Macro P	Macro R	Macro F1
CNN-rand with batch normalization	MeCab-ko	0.8462	0.8706	0.8568
CNN-static with batch normalization	MeCab-ko	0.8316	0.8810	0.8543
CNN-non-static with batch normalization	MeCab-ko	0.8561	0.8927	0.8733
5 medium regularized LSTMs with trainable Word2Vec	MeCab-ko	0.8738	0.8826	0.8752
DPCNN with batch normalization and randomly initialized embedding	MeCab-ko	0.8751	0.9023	0.8880
DPCNN with batch normalization and non-trainable Word2Vec	MeCab-ko	0.8717	0.9125	0.8910
DPCNN with batch normalization and trainable Word2Vec	MeCab-ko	0.8885	0.9000	0.8937
Fine-tuning pretrained KoBERT	WordPiece	0.8635	0.8876	0.8719
Fine-tuning pretrained KoELECTRA_v1	WordPiece	0.8735	0.8839	0.8758
Fine-tuning pretrained KoELECTRA_v2	MeCab-ko & WordPiece	**0.9118**	0.8865	0.8987
Ensemble of DPCNN and KoELECTRA_v2	MeCab-ko & WordPiece	0.9039	**0.9138**	**0.9084**

앙상블 모델



도면14



도면15

검사데이터-종합소견 텍스트

복부비만입니다. 허리둘레(복위)가 증가하면 만성질환에 이환될 위험이 증가합니다.
* 안저검사상 대부분 큰 이상 없는 고도근시에 의한 안저변화가 일차적으로 의심됩니다. 눈에 증상이 있으면 안과 진료 받으시기 바랍니다.
* 흉부 X-선 검사에서 좌하흉부, 흉막 유착 소견입니다. 정기검진 요합니다.
* 복부초음파 검사에서 지방간 경증 추정입니다. 간에 지방이 과다 축적된 상태로 비만, 음주, 고지혈증, 당뇨병, 약물 등이 원인으로 알려져있습니다. 향후 동맥경화, 간경화 등의 위험요인이 되므로 본인에게 해당되는 사항이 있으면 교정하시고 정기적으로 검사하십시오.
* 복부초음파 검사에서 좌측 신장에 석회화를 동반한 1.0cm 크기의 낭종이 의심됩니다. 소변, 혈액검사와 함께 정기적인 추적검사로 모양 및 크기 변화를 관찰하십시오.
* 전립선초음파 검사에서 전립선 내 국소석회화가 의심됩니다. 전립선 석회화는 비교적 흔한 소견으로서 임상적으로 문제가 되는 사항은 아니므로 정기적인 추적검사로 확인하십시오.
* **갑상선**초음파 검사에서 우측갑상선 결절 의심(0.6cm) 소견입니다. 추적검사 요합니다.
* 위내시경 검사상 점막의 주름이 위축된 경우로 속쓰림, 소화장애 등의 위장증상이 없으시면 치료 받으실 필요는 없고 정기적인 검사를 받으시기 바랍니다.
* 체성분 분석 결과 근육량 및 체지방률 모두 증가한 비만입니다. 비만에 대한 적절한 식사요법 및 무.유산소 운동으로 건강한 체형 유지 바랍니다.(체지방률 기준)
* 정신건강 척도 및 직무스트레스 평가는 별지의 결과지를 참조 하십시요.
* LDL-콜레스테롤이 높습니다. 이는 심혈관 질환의 발생위험을 높이는 주요원인입니다. 고혈압, 낮은 HDL콜레스테롤(40mg/dl미만),흡연, 나이(남 45세,여55세 이상), 심혈관질환의 가족력중 2개 이상 해당하거나 당뇨병시 약물요법이 필요한지 전문의와 상담하십시오.
* 혈당이 정상보다 다소 높지만 당뇨병 진단기준에 비해서는 낮아 공복시 혈당장애로 분류됩니다. 정상체중 유지 및 지속적인 운동하시고 3개월 후 재검하여 추적관찰 하십시오.
* HBsAb가 종전 검사상 양성에서 금번에는 약양성으로 나왔습니다. 역가가 낮아서 검사상 약양성일 뿐 질환에 대한 면역력은 존재합니다. 더 이상의 추가 접종은 하지 않으셔도 됩니다.
* 그 외 검사결과는 정상 또는 다소 벗어난 수치가 있어도 특별한 임상적 의의가 없는 것으로 사료됩니다.

추출된 문장 [토큰 개수< 기준 개수]

* 전립선초음파 검사에서 전립선 내 국소석회화가 의심됩니다. 전립선 석회화는 비교적 흔한 소견으로서 임상적으로 문제가 되는 사항은 아니므로 정기적인 추적검사로 확인하십시오.
* **갑상선**초음파 검사에서 우측갑상선 결절 의심(0.6cm) 소견입니다. 추적검사 요합니다.
* 위내시경 검사상 점막의 주름이 위축된 경우로 속쓰림, 소화장애 등의 위장증상이 없으시면 치료 받으실 필요는 없고 정기적인 검사를 받으시기 바랍니다.
* 체성분 분석 결과 근육량 및 체지방률 모두 증가한 비만입니다. 비만에 대한 적절한 식사요법 및 무.유산소 운동으로 건강한 체형 유지 바랍니다.(체지방률 기준)