(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년09월20일
(11) 등록번호 10-2445098
(24) 등록일자 2022년09월15일

(51) 국제특허분류(Int. Cl.)
G06F 40/126 (2020.01) *G06F 40/237* (2020.01)
G06K 9/62 (2022.01) *G10L 15/26* (2006.01)
G16H 10/20 (2018.01) *G16H 20/00* (2018.01)
G16H 50/20 (2018.01)
(52) CPC특허분류
G06F 40/126 (2020.01)
G06F 40/237 (2022.01)
(21) 출원번호 10-2021-0178444
(22) 출원일자 2021년12월14일
심사청구일자 2021년12월14일
(56) 선행기술조사문헌
US20190130282 A1*
(뒷면에 계속)

(73) 특허권자
(주)아이케어닥터
서울특별시 서초구 강남대로 479, 지하1층 108호 (반포동)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김민승
서울시 서초구 논현로 151 도곡이스타빌 B동 203호
이호익
서울시 성동구 성수일로 8길 47 롯데캐슬 107동 301호
(뒷면에 계속)
(74) 대리인
특허법인비엘티

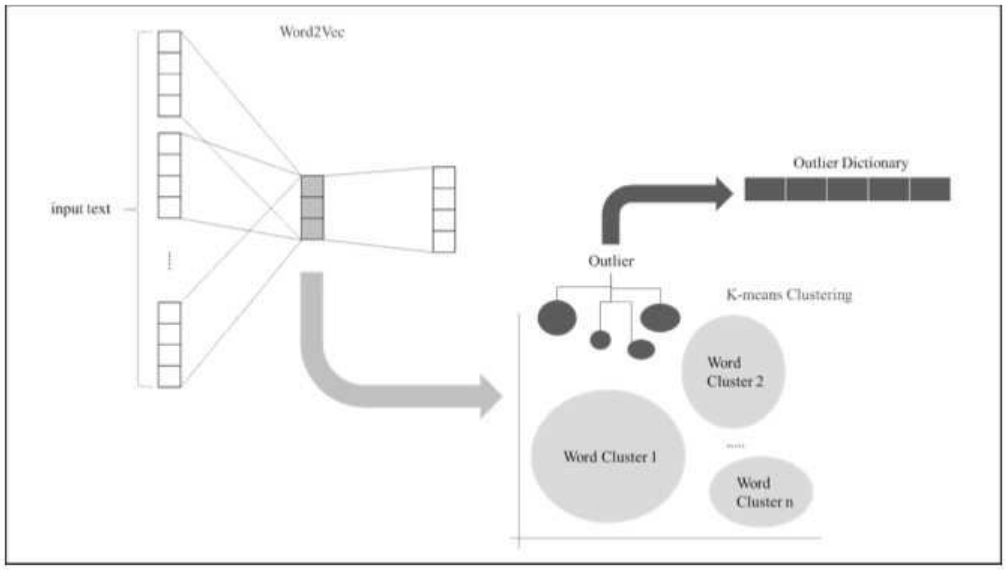
전체 청구항 수 : 총 8 항 심사관 : 경연정

(54) 발명의 명칭 **인공 지능 기반 의료 텍스트의 노이즈 데이터 필터링 방법, 장치 및 프로그램**

(57) 요 약

원격 진료 과정에서의 대화 내용으로부터 추출한 의료 텍스트의 노이즈 데이터를 필터링할 수 있는 인공 지능 기반 의료 텍스트의 노이즈 데이터 필터링 방법, 장치 및 프로그램에 관한 것으로, (a) 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하는 단계, (b) 상기 의료 텍스트를 하나의 문장마다 단어별로 임베딩하는 단계, (c) 상기 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하고, 상기 식별한 노이즈 단어 데이터를 노이즈 사전에 저장하는 단계, (d) 상기 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성하는 단계, (e) 상기 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성되는지를 확인하는 단계, 및 (f) 상기 새로운 의료 텍스트가 생성되면 상기 노이즈 필터를 통해 상기 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성하는 단계를 포함하는 것을 특징으로 한다.

대 표 도 - 도2



(52) CPC특허분류
G06K 9/6223 (2022.01)
G10L 15/26 (2013.01)
G16H 10/20 (2021.08)
G16H 20/00 (2021.08)
G16H 50/20 (2018.01)
(72) 발명자
최상민
서울특별시 은평구 연서로 23길 3-9, 502호
한요섭
서울특별시 서대문구 연세로 50

(56) 선행기술조사문헌
KR101806151 B1
KR1020210004057 A
US08090724 B1
M. Ester et al., A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, on Proceedings of KDD-96, pp.226-231, (1996)*
*는 심사관에 의하여 인용된 문헌

명 세 서

청구범위

청구항 1

장치에 의해 수행되는 방법에 있어서,

(a) 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하는 단계;

(b) 상기 의료 텍스트를 하나의 문장마다 단어별로 임베딩하는 단계;

(c) 상기 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하고, 상기 식별된 노이즈 단어 데이터를 노이즈 사전에 저장하는 단계;

(d) 상기 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성하는 단계;

(e) 상기 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성되는지를 확인하는 단계; 및

(f) 상기 새로운 의료 텍스트가 생성되면 상기 노이즈 필터를 통해 상기 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성하는 단계를 포함하고,

상기 (c) 단계는,

상기 임베딩된 단어들의 위치 정보를 기반으로 k-평균 클러스터링(k-means clustering) 알고리즘을 통해 단어들을 군집화하여 다수의 클러스터들을 생성하고, 상기 생성된 클러스터들에 속하지 않는 단어 데이터가 존재하면 해당 단어 데이터를 노이즈 단어 데이터로 간주하고,

상기 단어들을 군집화하여 다수의 클러스터들이 생성되면 상기 다수의 클러스터들을 진료 특징을 기반으로 재분류하고, 상기 재분류된 클러스터들로부터 노이즈 단어 데이터를 재식별하며, 상기 재식별된 노이즈 단어 데이터를 상기 노이즈 사전에 저장하는 것을 특징으로 하는 의료 텍스트의 노이즈 데이터 필터링 방법.

청구항 2

삭제

청구항 3

제1 항에 있어서,

상기 (c) 단계는,

상기 생성된 클러스터들 중 최소의 단어 수인 k개 미만의 단어로 구성된 클러스터가 존재하면 해당 클러스터에 포함되는 단어들을 노이즈 단어 데이터로 간주하는 것을 특징으로 하는 의료 텍스트의 노이즈 데이터 필터링 방법.

청구항 4

삭제

청구항 5

제1 항에 있어서,

상기 (d) 단계는,

상기 노이즈 사전에 저장된 일반 정보 기반 노이즈 단어 데이터로부터 일반 노이즈 필터를 생성하고, 상기 노이즈 사전에 저장된 상기 재식별된 노이즈 단어 데이터로부터 진료 특징 기반 노이즈 필터를 생성하는 것을 특징으로 하는 의료 텍스트의 노이즈 데이터 필터링 방법.

청구항 6

제5 항에 있어서,

상기 (d) 단계는,

상기 진료 특징 기반 노이즈 필터를 생성할 때, 처방 정보 기반 노이즈 필터와 환자 정보 기반 노이즈 필터를 포함하는 상기 진료 특징 기반 노이즈 필터를 생성하는 것을 특징으로 하는 의료 텍스트의 노이즈 데이터 필터링 방법.

청구항 7

제1 항에 있어서,

상기 (f) 단계는,

상기 새로운 의료 텍스트가 생성되면 상기 새로운 의료 텍스트의 각 문장에 상응하는 노이즈 필터를 기반으로 상기 새로운 의료 텍스트의 각 문장에 포함되는 노이즈 단어 데이터를 제거함으로써, 상기 새로운 의료 텍스트를 상기 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성하는 것을 특징으로 하는 의료 텍스트의 노이즈 데이터 필터링 방법.

청구항 8

제7 항에 있어서,

상기 (f) 단계는,

상기 새로운 의료 텍스트의 문장이 일반 정보 관련 문장이면 일반 노이즈 필터를 기반으로 상기 새로운 의료 텍스트의 일반 정보 관련 문장에 포함되는 노이즈 단어 데이터를 제거하고, 상기 새로운 의료 텍스트의 문장이 진료 특징 관련 문장이면 진료 특징 기반 노이즈 필터를 기반으로 상기 새로운 의료 텍스트의 진료 특징 문장에 포함되는 노이즈 단어 데이터를 제거하는 것을 특징으로 하는 의료 텍스트의 노이즈 데이터 필터링 방법.

청구항 9

하드웨어인 컴퓨터와 결합되어, 제1 항, 제3 항, 제5 항 내지 제8 항 중 어느 한 항의 의료 텍스트의 노이즈 데이터 필터링 방법을 수행시키기 위해 매체에 저장된, 의료 텍스트의 노이즈 데이터 필터링 장치의 의료 텍스트의 노이즈 데이터 필터링 방법을 제공하는 컴퓨터 프로그램.

청구항 10

의료 텍스트의 노이즈 데이터 필터링 방법을 제공하기 위한 컴퓨팅 장치로서,

하나 이상의 코어를 포함하는 프로세서; 및

메모리;

를 포함하고,

상기 프로세서는,

원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하고,

상기 의료 텍스트를 하나의 문장마다 단어별로 임베딩하며,

상기 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하여 상기 식별된 노이즈 단어 데이터를 노이즈 사전에 저장하고,

상기 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성하며,

상기 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성되는지를 확인하고, 및

상기 새로운 의료 텍스트가 생성되면 상기 노이즈 필터를 통해 상기 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성하고,

상기 프로세서는 상기 식별된 노이즈 단어 데이터를 노이즈 사전에 저장 시,

상기 임베딩된 단어들의 위치 정보를 기반으로 k-평균 클러스터링(k-means clustering) 알고리즘을 통해 단어들

을 군집화하여 다수의 클러스터들을 생성하고, 상기 생성된 클러스터들에 속하지 않는 단어 데이터가 존재하면 해당 단어 데이터를 노이즈 단어 데이터로 간주하고,

상기 단어들을 군집화하여 다수의 클러스터들이 생성되면 상기 다수의 클러스터들을 진료 특징을 기반으로 재분류하고, 상기 재분류된 클러스터들로부터 노이즈 단어 데이터를 재식별하며, 상기 재식별된 노이즈 단어 데이터를 상기 노이즈 사전에 저장하는 것을 특징으로 하는 컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 의료 텍스트의 노이즈 데이터 필터링 방법에 관한 것으로, 보다 구체적으로 원격 진료 과정에서의 대화 내용으로부터 추출한 의료 텍스트의 노이즈 데이터를 필터링할 수 있는 인공 지능 기반 의료 텍스트의 노이즈 데이터 필터링 방법, 장치 및 프로그램에 관한 것이다.

배 경 기 술

[0002] 최근 들어, 새로운 유형의 전염병이 유행하면서 비대면 의료 서비스에 대한 니즈가 높아지고 있다.

[0003] 이러한, 비대면 의료 서비스 니즈에 의해, 온라인을 통해 진단과 치료 및 자문 등의 의료 서비스를 제공하는 원격 진료가 증가하고 있는 추세이다.

[0004] 원격 진료는, 온라인을 통해 의사가 하는 다섯 가지 진찰 방법(문진, 시진, 촉진, 타진, 청진) 중 적어도 세 가지 이상을 사용하여 진찰을 하고, 소변검사, 혈액 검사, 심전도 검사 등 병원과 마찬가지의 검사를 실시하여 진단, 처방 및 치료를 시행하는 것을 의미한다.

[0005] 원격 진료가 실시되면 먼 곳에 떨어져 있는 환자에게도 전문적인 의료를 제공할 수 있어 의료 서비스의 지역 편중을 없애고, 의료 관련 자원을 최대한 효율적으로 운영할 수 있게 되어 궁극적으로는 의료비를 절감할 수 있는 효과를 기대할 수 있다.

[0006] 원격 진료는, 모바일 상에서 웹이나 앱 형태로 진행될 수 있는데, 음성 텍스트 변환 기술을 기반으로 의사와 환자간의 진료 상담 중에 대화 내용을 텍스트 형식으로 추출한다.

[0007] 하지만, 추출한 텍스트 데이터의 정확도가 음성 텍스트 변환 기술에 의존적이므로, 음성 텍스트 변환 기술 자체에 문제가 존재할 경우, 부정확한 텍스트가 추출되어 의사와 환자간의 진료 데이터에 오류가 발생할 수 있다.

[0008] 이러한 진료 데이터의 오류는, 인공 지능의 학습뿐만 아니라 환자의 진료에도 악영향을 미쳐 잘못된 처방을 내리는 문제가 발생할 수 있다.

[0009] 따라서, 향후, 원격 진료 과정 중 대화 내용으로부터 추출한 의료 텍스트의 노이즈 데이터를 필터링하여 정확성 및 신뢰성이 높은 의료 텍스트를 제공할 수 있는 의료 텍스트의 노이즈 데이터 필터링 기술의 개발이 요구되고 있다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 대한민국 등록특허 10-1909094호 (2018. 10. 11)

발명의 내용

해결하려는 과제

[0011] 상술한 바와 같은 문제점을 해결하기 위한 본 발명의 일 목적은, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환한 의료 텍스트를 단어별로 임베딩하고 군집화하여 노이즈 단어 데이터를 식별하고, 식별한 노이즈 단어 데이터를 기반으로 새로운 텍스트 데이터를 필터링함으로써, 정확성 및 신뢰성이 높은 의료 텍스트를 제공할 수 있는 의료 텍스트의 노이즈 데이터 필터링 방법, 장치 및 프로그램을 제공하는 것이다.

[0012] 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

[0013] 상술한 과제를 해결하기 위한 본 발명의 일 실시예에 따른 의료 텍스트의 노이즈 데이터 필터링 방법은, (a) 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하는 단계, (b) 상기 의료 텍스트를 하나의 문장마다 단어별로 임베딩하는 단계, (c) 상기 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하고, 상기 식별한 노이즈 단어 데이터를 노이즈 사전에 저장하는 단계, (d) 상기 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성하는 단계, (e) 상기 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성되는지를 확인하는 단계, 및 (f) 상기 새로운 의료 텍스트가 생성되면 상기 노이즈 필터를 통해 상기 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성하는 단계를 포함하는 것을 특징으로 한다.

[0014] 실시 예에 있어서, 상기 (a) 단계는, 상기 원격 진료 대화에 상응하는 음성 데이터를 STT(Speech-to-Text)를 통해 텍스트 데이터로 변환 처리하여 의료 텍스트를 생성하는 것을 특징으로 한다.

[0015] 실시 예에 있어서, 상기 STT(Speech-to-Text)의 의해 생성된 의료 텍스트는, 병명 및 복약 지도 중 적어도 어느 하나를 포함하는 처방 정보와, 나이, 성별, 거주 지역 중 적어도 어느 하나를 포함하는 환자 정보를 포함하는 것을 특징으로 한다.

[0016] 실시 예에 있어서, 상기 (b) 단계는, 상기 의료 텍스트의 문장 데이터를 미리 학습된 뉴럴 네트워크 모델에 입력하여 입력 문장 데이터에 대해 단어(word)별로 임베딩하는 것을 특징으로 한다.

[0017] 실시 예에 있어서, 상기 (c) 단계는, 상기 임베딩된 단어들의 위치 정보를 기반으로 k-평균 클러스터링(k-means clustering) 알고리즘을 통해 단어들을 군집화하여 다수의 클러스터들을 생성하고, 상기 생성된 클러스터들에 속하지 않는 단어 데이터가 존재하면 해당하는 단어 데이터를 노이즈 단어 데이터로 간주하는 것을 특징으로 한다.

[0018] 실시 예에 있어서, 상기 (c) 단계는, 상기 생성된 클러스터들 중 최소의 단어 수인 k개 미만의 단어로 구성된 클러스터가 존재하면 해당 클러스터에 포함되는 단어들을 노이즈 단어 데이터로 간주하는 것을 특징으로 한다.

[0019] 실시 예에 있어서, 상기 (c) 단계는, 상기 단어들을 군집화하여 다수의 클러스터들이 생성되면 상기 다수의 클러스터들을 진료 특징을 기반으로 재분류하고, 상기 재분류한 진료 특징 기반 클러스터들로부터 노이즈 단어 데이터를 식별하며, 상기 식별한 진료 특징 기반 노이즈 단어 데이터를 노이즈 사전에 저장하는 것을 특징으로 한다.

[0020] 실시 예에 있어서, 상기 (c) 단계는, 상기 진료 특징 중 처방 정보를 기반으로 클러스터들을 재분류하고, 상기 재분류한 클러스터들로부터 처방 정보 기반 노이즈 단어 데이터를 식별하여 상기 노이즈 사전에 저장하는 것을 특징으로 한다.

[0021] 실시 예에 있어서, 상기 (c) 단계는, 상기 진료 특징 중 환자 정보를 기반으로 클러스터들을 재분류하고, 상기 재분류한 클러스터들로부터 환자 정보 기반 노이즈 단어 데이터를 식별하여 상기 노이즈 사전에 저장하는 것을 특징으로 한다.

[0022] 실시 예에 있어서, 상기 (d) 단계는, 상기 노이즈 사전에 저장된 일반 정보 기반 노이즈 단어 데이터로부터 일반 노이즈 필터를 생성하고, 상기 노이즈 사전에 저장된 진료 특징 기반 노이즈 단어 데이터로부터 진료 특징 기반 노이즈 필터를 생성하는 것을 특징으로 한다.

[0023] 실시 예에 있어서, 상기 (d) 단계는, 상기 진료 특징 기반 노이즈 필터를 생성할 때, 처방 정보 기반 노이즈 필터와 환자 정보 노이즈 필터를 포함하는 진료 특징 기반 노이즈 필터를 생성하는 것을 특징으로 한다.

[0024] 실시 예에 있어서, 상기 (f) 단계는, 상기 새로운 의료 텍스트가 생성되면 상기 새로운 의료 텍스트의 각 문장에 상응하는 노이즈 필터를 기반으로 상기 의료 텍스트의 각 문장에 포함되는 노이즈 단어를 제거하고, 상기 노이즈 단어가 제거된 의료 텍스트를 재구성하는 것을 특징으로 한다.

[0025] 실시 예에 있어서, 상기 (f) 단계는, 상기 새로운 의료 텍스트의 문장이 일반 정보 관련 문장이면 일반 노이즈 필터를 기반으로 상기 의료 텍스트의 일반 정보 관련 문장에 포함되는 노이즈 단어를 제거하고, 상기 새로운 의

료 텍스트의 문장이 진료 특징 관련 문장이면 진료 특징 기반 노이즈 필터를 기반으로 상기 의료 텍스트의 진료 특징 문장에 포함되는 노이즈 단어를 제거하는 것을 특징으로 한다.

[0026] 실시 예에 있어서, 상기 (f) 단계는, 상기 새로운 의료 텍스트의 문장이 처방 정보 문장이면 처방 정보 기반 노이즈 필터를 기반으로 상기 의료 텍스트의 처방 정보 문장에 포함되는 노이즈 단어를 제거하고, 상기 새로운 의료 텍스트의 문장이 환자 정보 문장이면 환자 정보 기반 노이즈 필터를 기반으로 상기 의료 텍스트의 환자 정보 문장에 포함되는 노이즈 단어를 제거하는 것을 특징으로 한다.

[0027] 또한, 본 발명 일 실시예에 따른 컴퓨팅 장치는, 의료 텍스트의 노이즈 데이터 필터링 방법을 제공하기 위한 컴퓨팅 장치로서, 하나 이상의 코어를 포함하는 프로세서 및 메모리를 포함하고, 상기 프로세서는, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하고, 상기 의료 텍스트를 하나의 문장마다 단어별로 임베딩하며, 상기 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하여 상기 식별한 노이즈 단어 데이터를 노이즈 사전에 저장하고, 상기 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성하며, 상기 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성되는지를 확인하고, 및 상기 새로운 의료 텍스트가 생성되면 상기 노이즈 필터를 통해 상기 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성하는 것을 특징으로 한다.

[0028] 상술한 과제를 해결하기 위한 본 발명의 다른 실시 예에 따른 의료 텍스트의 노이즈 데이터 필터링 방법을 제공하는 컴퓨터 프로그램은, 하드웨어인 컴퓨터와 결합되어 상술한 방법 중 어느 하나의 방법을 수행하기 위해 매체에 저장된다.

[0029] 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

[0030] 상기와 같이 본 발명에 따르면, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환한 의료 텍스트를 단어별로 임베딩하고 군집화하여 노이즈 단어 데이터를 식별하고, 식별한 노이즈 단어 데이터를 기반으로 새로운 텍스트 데이터를 필터링함으로써, 정확성 및 신뢰성이 높은 의료 텍스트를 제공할 수 있다.

[0031] 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

[0032] 도 1은, 본 발명의 일 실시예에 따라, 의료 텍스트의 노이즈 데이터 필터링 방법을 제공하기 위한 동작을 수행하는 컴퓨팅 장치의 블록 구성도를 도시한 도면이다.

도 2 내지 도 7은, 본 발명의 일 실시예에 따라, 의료 텍스트의 노이즈 데이터 필터링 방법을 설명하기 위한 개념도이다.

도 8은, 본 발명의 일 실시예에 따라, 의료 텍스트의 노이즈 데이터 필터링 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0033] 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.

[0034] 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 구성

요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0035] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야의 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.

[0036] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0037] 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.

[0038] 본 명세서에서 신경망, 인공 신경망, 네트워크 함수는 종종 상호 교환 가능하게 사용될 수 있다.

[0039] 또한, 본 명세서에 걸쳐, 뉴럴 네트워크(neural network), 신경망 네트워크, 네트워크 함수는, 동일한 의미로 사용될 수 있다. 뉴럴 네트워크는, 일반적으로 “노드”라 지칭될 수 있는 상호 연결된 계산 단위들의 집합으로 구성될 수 있다. 이러한 “노드”들은, “뉴런(neuron)”들로 지칭될 수도 있다. 뉴럴 네트워크는, 적어도 둘 이상의 노드들을 포함하여 구성된다. 뉴럴 네트워크들을 구성하는 노드(또는 뉴런)들은 하나 이상의 “링크”에 의해 상호 연결될 수 있다.

[0040] 도 1은, 본 발명의 일 실시예에 따라, 의료 텍스트의 노이즈 데이터 필터링 방법을 제공하기 위한 동작을 수행하는 컴퓨팅 장치의 블록 구성도를 도시한 도면이다.

[0041] 도 1에 도시된 컴퓨팅 장치(100)의 구성은 간략화하여 나타낸 예시일 뿐이다. 본 발명의 일 실시예에서 컴퓨팅 장치(100)는 컴퓨팅 장치(100)의 컴퓨팅 환경을 수행하기 위한 다른 구성들이 포함될 수 있고, 개시된 구성들 중 일부만이 컴퓨팅 장치(100)를 구성할 수도 있다.

[0042] 컴퓨팅 장치(100)는, 프로세서(110), 메모리(130), 네트워크부(150)를 포함할 수 있다.

[0043] 본 발명에서, 프로세서(110)는, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하고, 의료 텍스트를 하나의 문장마다 단어별로 임베딩하며, 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하여 식별한 노이즈 단어 데이터를 노이즈 사전에 저장하고, 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성하며, 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성되는지를 확인하고, 새로운 의료 텍스트가 생성되면 노이즈 필터를 통해 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성할 수 있다.

[0044] 여기서, 프로세서(110)는, 원격 진료 대화에 상응하는 음성 데이터를 STT(Speech-to-Text)를 통해 텍스트 데이터로 변환 처리하여 의료 텍스트를 생성할 수 있다.

[0045] 일 예로, STT(Speech-to-Text)의 의해 생성된 의료 텍스트는, 병명 및 복약 지도 중 적어도 어느 하나를 포함하는 처방 정보와, 나이, 성별, 거주 지역 중 적어도 어느 하나를 포함하는 환자 정보를 포함할 수 있는데, 이는 일 실시예일 뿐, 이에 한정되지는 않는다.

[0046] 다음, 프로세서(110)는, 의료 텍스트의 문장 데이터를 미리 학습된 뉴럴 네트워크 모델에 입력하여 입력 문장 데이터에 대해 단어(word)별로 임베딩할 수 있다.

[0047] 일 예로, 뉴럴 네트워크 모델은, 워드투벡터(Word2Vec) 모델의 스킵-그램(skip-gram) 알고리즘을 포함할 수 있는데, 이는 일 실시예일 뿐, 이에 한정되지는 않는다.

[0048] 여기서, 프로세서(110)는, 의료 텍스트에서 하나의 문장 데이터가 n개의 단어로 구성되면 각 단어에 대해 문장 데이터 내의 위치를 기반으로 원-핫-벡터(one-hot-vector)를 진행할 수 있다.

[0049] 또한, 프로세서(110)는, 중심 단어에 상응하는 하나의 원-핫-벡터가 프로젝션 레이어(projection layer)를 거쳐 주변 단어에 상응하는 다수의 원-핫-벡터로 출력되고, 각 출력 데이터를 소프트맥스(softmax) 알고리즘을 활용하여 변환하며, 변환된 출력 데이터와 실제 데이터(real data) 사이의 오차를 크로스 엔트로피(cross-entropy) 알고리즘을 활용하여 산출할 수 있다.

[0050] 여기서, 프로세서(110)는, 그레디언트 디센트(gradient descent) 알고리즘을 활용하여 출력 데이터와 실제 데이터 사이의 오차를 최소화할 수 있다.

[0051] 이어, 프로세서(110)는, 임베딩된 단어들의 위치 정보를 기반으로 k-평균 클러스터링(k-means clustering) 알고리즘을 통해 단어들을 군집화하여 다수의 클러스터들을 생성하고, 생성된 클러스터들에 속하지 않는 단어 데이터가 존재하면 해당하는 단어 데이터를 노이즈 단어 데이터로 간주할 수 있다.

[0052] 여기서, 프로세서(110)는, 생성된 클러스터들 중 최소의 단어 수인 k개 미만의 단어로 구성된 클러스터가 존재하면 해당 클러스터에 포함되는 단어들을 노이즈 단어 데이터로 간주할 수 있다.

[0053] 또한, 프로세서(110)는, 단어들을 군집화하여 다수의 클러스터들이 생성되면 다수의 클러스터들을 진료 특징을 기반으로 재분류하고, 재분류한 진료 특징 기반 클러스터들로부터 노이즈 단어 데이터를 식별하며, 식별한 진료 특징 기반 노이즈 단어 데이터를 노이즈 사전에 저장할 수 있다.

[0054] 여기서, 프로세서(110)는, 진료 특징 중 처방 정보를 기반으로 클러스터들을 재분류하고, 재분류한 클러스터들로부터 처방 정보 기반 노이즈 단어 데이터를 식별하여 노이즈 사전에 저장할 수 있다.

[0055] 일 예로, 처방 정보는, 병명 및 복약 지도 중 적어도 어느 하나를 포함할 수 있는데, 이는 일 실시예일 뿐, 이에 한정되지는 않는다.

[0056] 경우에 따라, 프로세서(110)는, 진료 특징 중 환자 정보를 기반으로 클러스터들을 재분류하고, 재분류한 클러스터들로부터 환자 정보 기반 노이즈 단어 데이터를 식별하여 노이즈 사전에 저장할 수도 있다.

[0057] 일 예로, 환자 정보는, 나이, 성별, 거주 지역 중 적어도 어느 하나를 포함할 수 있는데, 이는 일 실시예일 뿐, 이에 한정되지는 않는다.

[0058] 다음, 프로세서(110)는, 노이즈 사전에 저장된 일반 정보 기반 노이즈 단어 데이터로부터 일반 노이즈 필터를 생성하고, 노이즈 사전에 저장된 진료 특징 기반 노이즈 단어 데이터로부터 진료 특징 기반 노이즈 필터를 생성할 수 있다.

[0059] 여기서, 프로세서(110)는, 진료 특징 기반 노이즈 필터를 생성할 때, 처방 정보 기반 노이즈 필터와 환자 정보 노이즈 필터를 포함하는 진료 특징 기반 노이즈 필터를 생성할 수 있다.

[0060] 그리고, 프로세서(110)는, 새로운 의료 텍스트가 생성되면 새로운 의료 텍스트의 각 문장에 상응하는 노이즈 필터를 기반으로 의료 텍스트의 각 문장에 포함되는 노이즈 단어를 제거하고, 노이즈 단어가 제거된 의료 텍스트를 재구성할 수 있다.

[0061] 여기서, 프로세서(110)는, 새로운 의료 텍스트의 문장이 일반 정보 관련 문장이면 일반 노이즈 필터를 기반으로 의료 텍스트의 일반 정보 관련 문장에 포함되는 노이즈 단어를 제거하고, 새로운 의료 텍스트의 문장이 진료 특징 관련 문장이면 진료 특징 기반 노이즈 필터를 기반으로 의료 텍스트의 진료 특징 문장에 포함되는 노이즈 단어를 제거할 수 있다.

[0062] 경우에 따라, 프로세서(110)는, 새로운 의료 텍스트의 문장이 처방 정보 문장이면 처방 정보 기반 노이즈 필터를 기반으로 의료 텍스트의 처방 정보 문장에 포함되는 노이즈 단어를 제거하고, 새로운 의료 텍스트의 문장이 환자 정보 문장이면 환자 정보 기반 노이즈 필터를 기반으로 의료 텍스트의 환자 정보 문장에 포함되는 노이즈 단어를 제거할 수도 있다.

[0063] 본 발명의 일 실시예에 따르면, 프로세서(110)는, 하나 이상의 코어로 구성될 수 있으며, 컴퓨팅 장치의 중앙처리 장치(CPU: central processing unit), 범용 그래픽 처리 장치 (GPGPU: general purpose graphics processing unit), 텐서 처리 장치(TPU: tensor processing unit) 등의 데이터 분석, 딥러닝을 위한 프로세서를 포함할 수 있다. 프로세서(110)는, 메모리(130)에 저장된 컴퓨터 프로그램을 판독하여 본 발명의 일 실시예에 따른 기계 학습을 위한 데이터 처리를 수행할 수 있다. 본 발명의 일실시예에 따라 프로세서(110)는, 신경망의 학습을 위한 연산을 수행할 수 있다. 프로세서(110)는, 딥러닝(DL: deep learning)에서 학습을 위한 입력 데이터의 처리, 입력 데이터에서의 피처 추출, 오차 계산, 역전파(backpropagation)를 이용한 신경망의 가중치 업데이트 등의 신경망의 학습을 위한 계산을 수행할 수 있다. 프로세서(110)의 CPU, GPGPU, 및 TPU 중 적어도 하나가 네트워크 함수의 학습을 처리할 수 있다. 예를 들어, CPU 와 GPGPU가 함께 네트워크 함수의 학습, 네트워크 함수를 이용한 데이터 분류를 처리할 수 있다. 또한, 본 발명의 일 실시예에서 복수의 컴퓨팅 장치의 프로세서를 함께 사용하여 네트워크 함수의 학습, 네트워크 함수를 이용한 데이터 분류를 처리할 수 있다. 또한, 본 발명의 일 실시예에 따른 컴퓨팅 장치에서 수행되는 컴퓨터 프로그램은, CPU, GPGPU 또는 TPU 실행가능 프로그램일 수 있다.

[0064] 본 발명의 일 실시예에 따르면, 메모리(130)는, 의료 텍스트의 노이즈 데이터 필터링 방법을 수행하기 위한 컴퓨터 프로그램을 저장할 수 있으며, 저장된 컴퓨터 프로그램은 프로세서(120)에 의하여 판독되어 구동될 수 있다. 메모리(130)는, 프로세서(110)가 생성하거나 결정한 임의의 형태의 정보 및 네트워크부(150)가 수신한 임의의 형태의 정보를 저장할 수 있다.

[0065] 본 발명의 일 실시예에 따르면, 메모리(130)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. 컴퓨팅 장치(100)는 인터넷(internet) 상에서 상기 메모리(130)의 저장 기능을 수행하는 웹 스토리지(web storage)와 관련되어 동작할 수도 있다. 전술한 메모리에 대한 기재는 예시일 뿐, 이에 제한되지 않는다.

[0066] 본 발명의 일 실시예에 따른 네트워크부(150)는, 의료 텍스트의 노이즈 데이터 필터링 방법 결과 정보 등을 다른 컴퓨팅 장치, 서버 등과 송수신할 수 있다. 또한, 네트워크부(150)는, 복수의 컴퓨팅 장치 사이의 통신을 가능하게 하여 복수의 컴퓨팅 장치 각각에서 의료 텍스트의 노이즈 데이터 필터링 또는 모델의 학습을 위한 동작들이 분산 수행되도록 할 수 있다. 네트워크부(150)는, 복수의 컴퓨팅 장치 사이의 통신을 가능하게 하여 의료 텍스트의 노이즈 데이터 필터링 또는 네트워크 함수를 사용한 모델 학습을 위한 연산을 분산 처리하도록 할 수 있다.

[0067] 본 발명의 일 실시예에 따른 네트워크부(150)는, 근거리(단거리), 원거리, 유선 및 무선 등과 같은 현재 사용 및 구현되는 임의의 형태의 유무선 통신 기술에 기반하여 동작할 수 있으며, 다른 네트워크들에서도 사용될 수 있다.

[0068] 본 발명의 컴퓨팅 장치(100)는, 출력부 및 입력부를 더 포함할 수도 있다.

[0069] 본 발명의 일 실시예에 따른 출력부는, 의료 텍스트의 노이즈 데이터 필터링 방법을 수행하기 위한 사용자 인터페이스(UI, user interface)를 표시할 수 있다. 출력부는, 프로세서(110)가 생성하거나 결정한 임의의 형태의 정보 및 네트워크부(150)가 수신한 임의의 형태의 정보를 출력할 수 있다.

[0070] 본 발명의 일 실시예에서, 출력부는, 액정 디스플레이(liquid crystal display, LCD), 박막 트랜지스터 액정 디스플레이(thin film transistor-liquid crystal display, TFT LCD), 유기 발광 다이오드(organic light-emitting diode, OLED), 플렉시블 디스플레이(flexible display), 3차원 디스플레이(3D display) 중에서 적어도 하나를 포함할 수 있다. 이들 중 일부 디스플레이 모듈은, 그를 통해 외부를 볼 수 있도록 투명형 또는 광투과형으로 구성될 수 있다. 이는 투명 디스플레이 모듈이라 지칭될 수 있는데, 상기 투명 디스플레이 모듈의 대표적인 예로는 TOLED(Transparent OLED) 등이 있다.

[0071] 본 발명의 일 실시예에 따른 입력부는, 사용자 입력을 수신할 수 있다. 입력부는, 사용자 입력을 수신받기 위한 사용자 인터페이스 상의 키 및/또는 버튼들, 또는 물리적인 키 및/또는 버튼들을 구비할 수 있다. 입력부를 통한 사용자 입력에 따라 본 발명의 실시예들에 따른 디스플레이를 제어하기 위한 컴퓨터 프로그램이 실행될 수 있다.

[0072] 본 발명의 실시예들에 따른 입력부는, 사용자의 버튼 조작 또는 터치 입력을 감지하여 신호를 수신하거나, 카메라 또는 마이크로폰을 통하여 사용자 등의 음성 또는 동작을 수신하여 이를 입력 신호로 변환할 수도 있다. 이를 위해 음성 인식(Speech Recognition) 기술 또는 동작 인식(Motion Recognition) 기술들이 사용될 수 있다.

[0073] 본 발명의 실시예들에 따른 입력부는, 컴퓨팅 장치(100)와 연결된 외부 입력 장비로서 구현될 수도 있다. 예를 들어, 입력 장비는 사용자 입력을 수신하기 위한 터치 패드, 터치 펜, 키보드 또는 마우스 중 적어도 하나일 수 있으나, 이는 예시일 뿐이며 이에 제한되는 것은 아니다.

[0074] 본 발명의 일 실시예에 따른 입력부는, 사용자 터치 입력을 인식할 수 있다. 본 발명의 일 실시예에 따른 입력부는, 출력부와 동일한 구성일 수도 있다. 입력부는, 사용자의 선택 입력을 수신하도록 구현되는 터치 스크린으로 구성될 수 있다. 터치 스크린은, 접촉식 정전용량 방식, 적외선 광 감지 방식, 표면 초음파(SAW) 방식, 압전 방식, 저항막 방식 중 어느 하나의 방식이 사용될 수 있다. 전술한 터치 스크린에 대한 자세한 기재는, 본 발명의 일 실시예에 따른 예시일 뿐이며, 다양한 터치 스크린 패널이 컴퓨팅 장치(100)에 채용될 수 있다. 터치 스크린으로 구성된 입력부는, 터치 센서를 포함할 수 있다. 터치 센서는, 입력부의 특정 부위에 가해진 압력 또는

입력부의 특정 부위에 발생하는 정전 용량 등의 변화를 전기적인 입력신호로 변환하도록 구성될 수 있다. 터치 센서는, 터치 되는 위치 및 면적뿐만 아니라, 터치 시의 압력까지도 검출할 수 있도록 구성될 수 있다. 터치 센서에 대한 터치입력이 있는 경우, 그에 대응하는 신호(들)는 터치 제어기로 보내진다. 터치 제어기는, 그 신호(들)를 처리한 다음 대응하는 데이터를 프로세서(110)로 전송할 수 있다. 이로써, 프로세서(110)는 입력부의 어느 영역이 터치 되었는지 여부 등을 인식할 수 있게 된다.

[0075] 본 발명의 일 실시예에서, 서버는, 서버의 서버 환경을 수행하기 위한 다른 구성들이 포함될 수도 있다. 서버는 임의의 형태의 장치는 모두 포함할 수 있다. 서버는, 디지털 기기로서, 랩탑 컴퓨터, 노트북 컴퓨터, 데스크톱 컴퓨터, 웹 패드, 이동 전화기와 같이 프로세서를 탑재하고 메모리를 구비한 연산 능력을 갖춘 디지털 기기일 수 있다.

[0076] 본 발명의 일 실시예에 따른 의료 텍스트의 노이즈 데이터 필터링 결과를 표시하는 사용자 인터페이스를 사용자 단말로 제공하기 위한 동작을 수행하는 서버(미도시)는, 네트워크부, 프로세서 및 메모리를 포함할 수 있다.

[0077] 서버는, 본 발명의 실시예들에 따른 사용자 인터페이스를 생성할 수 있다. 서버는, 클라이언트(예를 들어, 사용자 단말)에게 네트워크를 통해 정보를 제공하는 컴퓨팅 시스템일 수 있다. 서버는, 생성한 사용자 인터페이스를 사용자 단말로 전송할 수 있다. 이러한 경우, 사용자 단말은, 서버에 액세스할 수 있는 임의의 형태의 컴퓨팅 장치(100)일 수 있다. 서버의 프로세서는, 네트워크부를 통해 사용자 단말로 사용자 인터페이스를 전송할 수 있다. 본 발명의 실시예들에 따른 서버는 예를 들어, 클라우드 서버일 수 있다. 서버는 서비스를 처리하는 웹 서버일 수 있다. 전술한 서버의 종류는 예시일 뿐이며 이에 제한되지 않는다.

[0078] 이와 같이, 본 발명은, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환한 의료 텍스트를 단어별로 임베딩하고 군집화하여 노이즈 단어 데이터를 식별하고, 식별한 노이즈 단어 데이터를 기반으로 새로운 텍스트 데이터를 필터링함으로써, 정확성 및 신뢰성이 높은 의료 텍스트를 제공할 수 있다.

[0079] 도 2 내지 도 7은, 본 발명의 일 실시예에 따라, 의료 텍스트의 노이즈 데이터 필터링 방법을 설명하기 위한 개념도이다.

[0080] 도 2 내지 도 7에 도시된 바와 같이, 본 발명은, 의료 텍스트 보정을 위한 방법에 관한 기술이다.

[0081] 본 발명의 의료 텍스트는, 원격 진료 과정에서 추출된 데이터를 포함할 수 있다.

[0082] 여기서, 원격 진료는, 모바일 상에서 웹이나 앱 형태로 진행되며, STT(Speech-to-Text)를 통해 진료 중 대화 내용을 텍스트 형식으로 추출할 수 있다.

[0083] 이때, 추출된 텍스트 데이터의 정확도는, STT 기술에 의존적이므로, STT 기술을 활용하여 음성 데이터로부터 텍스트 데이터를 추출할 경우, STT 기술 자체에 문제가 존재한다면 부정확한 텍스트 데이터가 추출될 수 있다.

[0084] 따라서, STT 자체를 개발하여 학습하지 않는 한 추출된 텍스트 데이터는, 기존 기술에 의존적이므로, 본 발명은, 기존 STT 기술을 사용하면서 추가적으로 추출된 텍스트를 보정하여 텍스트 데이터의 정확도를 제고하는 방법이다.

[0085] 여기서, 정확도란, 음성 데이터와 추출된 텍스트 데이터 사이에 존재하는 오차의 정도가 아닌 음성 데이터의 문맥상 의미를 추출된 텍스트 데이터가 포함하고 있는 정도를 의미한다.

[0086] 예를 들어, 음성 데이터 '가나다'를 통해 추출된 텍스트 데이터를 '가냐다'라고 가정할 경우, 음성 데이터 '가나다'를 v1, '가냐다'를 t1이라고 하면, 두 데이터 사이의 오차의 정도를 에디트 디스턴스(edit distance)인 dist(v1, t1)으로 표현할 수 있다.

[0087] 이러한 방식은, 텍스트의 차이에 따라 명확한 오차 값이 존재하는 반면에 문맥상의 차이에서는 두 데이터의 문맥상 유사도(유사도는, cosine simularity로 가정) sim(v1, t1)을 계산하여 그 결과가 높은 경우, 올바른 추출로 가정할 수 있다.

[0088] 따라서, 본 발명은, 추출된 텍스트 데이터 사이의 문맥상 유사도를 도출하고, 이 정보를 활용하여 의미 있는 단어와 무의미한 단어를 구분할 수 있다.

[0089] 그리고, 본 발명은, 이러한 분류를 활용하여 무의미한 단어를 필터링하는 방법이다.

[0090] 도 2와 같이, 본 발명은, 텍스트 임베딩, 텍스트 군집화 및 분류, 그리고 노이즈 데이터 필터인 아웃라이어(outlier) 구성의 3 단계를 통해 진행될 수 있다.

[0091] 본 발명은, 텍스트 임베딩 단계로서, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성하고, 의료 텍스트를 하나의 문장마다 단어별로 임베딩할 수 있다.

[0092] 다음, 본 발명은, 텍스트 군집화 및 분류 단계로서, 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별할 수 있다.

[0093] 이어, 본 발명은, 아웃라이어 구성 단계로서, 식별한 노이즈 단어 데이터를 노이즈 사전에 저장하고, 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성할 수 있다.

[0094] 그리고, 본 발명은, 새로운 의료 텍스트가 입력되면 노이즈 필터를 통해 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성할 수 있다.

[0095] 한편, 텍스트 임베딩 단계는, 원격 진료 대화에 상응하는 음성 데이터를 STT(Speech-to-Text)를 통해 텍스트 데이터로 변환 처리하여 의료 텍스트를 생성하고, 의료 텍스트의 문장 데이터를 미리 학습된 뉴럴 네트워크 모델에 입력하여 입력 문장 데이터에 대해 단어(word)별로 임베딩할 수 있다.

[0096] 일 예로, 뉴럴 네트워크 모델은, 워드투벡터(Word2Vec) 모델의 스킵-그램(skip-gram) 알고리즘을 포함할 수 있다.

[0097] 도 3은, 스킵-그램(skip-gram)을 적용하기 위해 텍스트 데이터를 원-핫-벡터(one-hot-vector)로 변환하는 과정을 보여주고 있다.

[0098] 여기서, 도 3과 같이, 본 발명은, 의료 텍스트에서 하나의 문장 데이터가 n개의 단어로 구성되면 각 단어에 대해 문장 데이터 내의 위치를 기반으로 원-핫-벡터(one-hot-vector)를 진행할 수 있다.

[0099] 예를 들어, n개의 단어가 있는 문장에서 w1이 문장의 첫 번째 위치한 단어라면, w1에 대한 원-핫-벡터(one-hot-vector)는, 첫 번째 위치만 1이고 나머지는 0인 n차원의 벡터일 수 있다.

[0100] 또한, 도 4는, 스킵-그램(skip-gram)을 이용하여 입력 문장에 대해 단어(word)별로 임베딩을 진행하는 과정을 보여주고 있다.

[0101] 즉, 첫 단어와 나머지 단어들의 관계를 도 4와 같은 구조로 학습함으로써, 임베딩을 진행할 수 있다.

[0102] 여기서, 본 발명은, 도 4와 같이, 중심 단어에 상응하는 하나의 원-핫-벡터가 프로젝션 레이어(projection layer)를 거쳐 주변 단어에 상응하는 다수의 원-핫-벡터로 출력되고, 각 출력 데이터를 소프트맥스(softmax) 알고리즘을 활용하여 변환하며, 변환된 출력 데이터와 실제 데이터(real data) 사이의 오차를 크로스 엔트로피(cross-entropy) 알고리즘을 활용하여 산출할 수 있다.

[0103] 여기서, 프로세서(110)는, 그레디언트 디센트(gradient descent) 알고리즘을 활용하여 출력 데이터와 실제 데이터 사이의 오차를 최소화할 수 있다.

[0104] 이어, 텍스트 군집화 및 분류 단계는, 임베딩된 단어들의 위치 정보를 기반으로 k-평균 클러스터링(k-means clustering) 알고리즘을 통해 단어들을 군집화하여 다수의 클러스터들을 생성하고, 생성된 클러스터들에 속하지 않는 단어 데이터가 존재하면 해당하는 단어 데이터를 노이즈 단어 데이터로 간주할 수 있다.

[0105] 즉, 본 발명은, 모든 단어(word) 단위의 텍스트 데이터를 벡터(vector)로 표현할 수 있고, 각 단어 벡터(word vector)를 기반으로 단어(word) 군집화가 가능ㅎ하다.

[0106] 도 5의 (1)은, 2차원 좌표 평면 상에 표현된 단어 벡터를 보여주는 일 예이고, 도 5의 (2)는, 도 5의 (1)과 같이 단어 위치 정보를 기반으로 k-평균 클러스터링 알고리즘을 적용한 결과의 일 예를 보여주고 있다.

[0107] 도 5의 (2)에서는, 크게 2개의 클러스터로 구성되며, 3개의 단어 데이터는 어느 클러스터에도 속하지 않는 노이즈 단어 데이터로 간주할 수 있다.

[0108] 여기서, 본 발명은, 생성된 클러스터들 중 최소의 단어 수인 k개 미만의 단어로 구성된 클러스터가 존재하면 해당 클러스터에 포함되는 단어들을 노이즈 단어 데이터로 간주할 수 있다.

[0109] 또한, 본 발명은, 단어들을 군집화하여 다수의 클러스터들이 생성되면 다수의 클러스터들을 진료 특징을 기반으로 재분류하고, 재분류한 진료 특징 기반 클러스터들로부터 노이즈 단어 데이터를 식별하며, 식별한 진료 특징 기반 노이즈 단어 데이터를 노이즈 사전에 저장할 수 있다.

[0110] 여기서, 본 발명은, 진료 특징 중 처방 정보를 기반으로 클러스터들을 재분류하고, 재분류한 클러스터들로부터

처방 정보 기반 노이즈 단어 데이터를 식별하여 노이즈 사전에 저장할 수 있다.

[0111] 경우에 따라, 본 발명은, 진료 특징 중 환자 정보를 기반으로 클러스터들을 재분류하고, 재분류한 클러스터들로부터 환자 정보 기반 노이즈 단어 데이터를 식별하여 노이즈 사전에 저장할 수도 있다.

[0112] 다음, 아웃라이어 구성 단계는, 노이즈 사전에 저장된 단어들을 아웃라이어(outlier)로 가정할 수 있다.

[0113] 그리고, 본 발명은, STT를 통해 새롭게 입력으로 들어온 의료 텍스트의 단어들을 먼저 노이즈 사전을 이용하여 필터링할 수 있다.

[0114] 도 5는, 노이즈 사전을 통해 구성된 노이즈 필터인 아웃라이어(Outlier)를 이용하여 새로운 의료 텍스트 입력의 노이즈 데이터를 필터링하는 예를 보여주고 있다.

[0115] 도 5와 같이, 문장 k(sentence_k)는, STT를 통해 새롭게 도출된 텍스트 데이터의 일 예로서, 문장 k(sentence_k) 내에 포함된 노이즈 단어 데이터 w_15와 w_17을 제거한 후에 문장 k(sentence_k)를 재구성할 수 있다.

[0116] 여기서, 재구성된 문장 k는, 노이즈 단어 데이터인 아웃라이어(outlier) 단어들이 제거된 문장으로 가정할 수 있다.

[0117] 본 발명과 같이, Word2vec과 클러스터링을 통해 아웃라이어(outlier)를 선택할 때의 이점은, 텍스트 데이터 변환 중 우연히 혹은 잘못된 형태로 추출된 단어들을 검출할 수 있다는 것이다.

[0118] 이로 인해, 우연히 변환되거나 잘못 변환된 단어들은, 그렇지 않은 단어들에 비해 빈도수가 적을 것이고, 다양한 문장 내에 위치할 확률도 줄어들게 된다.

[0119] 따라서, 이러한 단어들을 아웃라이어(outlier)로 선택할 경우, 향후 잘못된 단어들을 검출할 수 있는 가능성이 높아지게 된다.

[0120] 또한, 본 발명은, 분류 정보 기반으로 아웃라이어(outlier)를 구성할 수 있다.

[0121] 즉, 본 발명은, 진료 특징으로 분류된 정보를 기반으로 노이즈 단어 데이터를 식별하면 각 분류 정보별로 노이즈 사전을 도출할 수 있다.

[0122] 그리고, 각 노이즈 사전을 기반으로 원격 진료에 적합한 아웃라이어(outlier) 검출이 가능하다.

[0123] 도 6은, 분류 정보 기반 아웃라이어(outlier) 구성의 예시를 보여주고 있다.

[0124] 도 6은, 처방 A에 대한 분류 군집을 나타내며, 문장 k(sentence_k)는, 처방 A에 대한 입력문장의 단어 단위 집합이다.

[0125] 즉, 이는, STT를 통해 입력받은 문장 중 처방 A로 분류된 문장을 의미한다.

[0126] 그리고, 해당 문장(sentence)에 대해 아웃라이어(outlier) 검출을 진행할 수 있다.

[0127] 따라서, 본 발명은, 노이즈 사전에 저장된 일반 정보 기반 노이즈 단어 데이터로부터 일반 노이즈 필터를 생성하고, 노이즈 사전에 저장된 진료 특징 기반 노이즈 단어 데이터로부터 진료 특징 기반 노이즈 필터를 생성할 수 있다.

[0128] 여기서, 본 발명은, 진료 특징 기반 노이즈 필터를 생성할 때, 처방 정보 기반 노이즈 필터와 환자 정보 노이즈 필터를 포함하는 진료 특징 기반 노이즈 필터를 생성할 수 있다.

[0129] 그리고, 본 발명은, 새로운 의료 텍스트가 생성되면 새로운 의료 텍스트의 각 문장에 상응하는 노이즈 필터를 기반으로 의료 텍스트의 각 문장에 포함되는 노이즈 단어를 제거하고, 노이즈 단어가 제거된 의료 텍스트를 재구성할 수 있다.

[0130] 여기서, 본 발명은, 새로운 의료 텍스트의 문장이 일반 정보 관련 문장이면 일반 노이즈 필터를 기반으로 의료 텍스트의 일반 정보 관련 문장에 포함되는 노이즈 단어를 제거하고, 새로운 의료 텍스트의 문장이 진료 특징 관련 문장이면 진료 특징 기반 노이즈 필터를 기반으로 의료 텍스트의 진료 특징 문장에 포함되는 노이즈 단어를 제거할 수 있다.

[0131] 일 예로, 본 발명은, 새로운 의료 텍스트의 문장이 처방 정보 문장이면 처방 정보 기반 노이즈 필터를 기반으로 의료 텍스트의 처방 정보 문장에 포함되는 노이즈 단어를 제거하고, 새로운 의료 텍스트의 문장이 환자 정보 문

장이면 환자 정보 기반 노이즈 필터를 기반으로 의료 텍스트의 환자 정보 문장에 포함되는 노이즈 단어를 제거할 수도 있다.

[0132] 도 8은, 본 발명의 일 실시예에 따라, 의료 텍스트의 노이즈 데이터 필터링 방법을 설명하기 위한 흐름도이다.

[0133] 도 8에 도시된 바와 같이, 본 발명은, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환하여 의료 텍스트를 생성할 수 있다(S10).

[0134] 이어, 본 발명은, 의료 텍스트를 하나의 문장마다 단어별로 임베딩할 수 있다(S20).

[0135] 다음, 본 발명은, 임베딩된 단어들을 군집화하여 노이즈 단어 데이터를 식별하고, 식별한 노이즈 단어 데이터를 노이즈 사전에 저장하며, 노이즈 사전에 저장된 노이즈 단어 데이터를 기반으로 노이즈 필터를 생성할 수 있다(S30).

[0136] 여기서, 본 발명은, 임베딩된 단어들의 위치 정보를 기반으로 k-평균 클러스터링(k-means clustering) 알고리즘을 통해 단어들을 군집화하여 다수의 클러스터들을 생성하고, 생성된 클러스터들에 속하지 않는 단어 데이터가 존재하면 해당하는 단어 데이터를 노이즈 단어 데이터로 간주할 수 있다.

[0137] 또한, 본 발명은, 다수의 클러스터들을 진료 특징을 기반으로 재분류하고, 재분류한 진료 특징 기반 클러스터들로부터 노이즈 단어 데이터를 식별하며, 식별한 진료 특징 기반 노이즈 단어 데이터를 노이즈 사전에 저장할 수 있다.

[0138] 또한, 본 발명은, 노이즈 사전에 저장된 일반 정보 기반 노이즈 단어 데이터로부터 일반 노이즈 필터를 생성하고, 노이즈 사전에 저장된 진료 특징 기반 노이즈 단어 데이터로부터 진료 특징 기반 노이즈 필터를 생성할 수 있다.

[0139] 그리고, 본 발명은, 원격 진료 대화에 상응하는 새로운 의료 텍스트가 생성할 수 있다(S40).

[0140] 이어, 본 발명은, 새로운 의료 텍스트가 생성되면 노이즈 필터를 통해 새로운 의료 텍스트를 필터링하여 노이즈 단어 데이터가 제거된 의료 텍스트로 재구성할 수 있다(S50).

[0141] 여기서, 본 발명은, 새로운 의료 텍스트가 생성되면 새로운 의료 텍스트의 각 문장에 상응하는 노이즈 필터를 기반으로 의료 텍스트의 각 문장에 포함되는 노이즈 단어를 제거하고, 노이즈 단어가 제거된 의료 텍스트를 재구성할 수 있다.

[0142] 일 예로, 본 발명은, 새로운 의료 텍스트의 문장이 일반 정보 관련 문장이면 일반 노이즈 필터를 기반으로 의료 텍스트의 일반 정보 관련 문장에 포함되는 노이즈 단어를 제거하고, 새로운 의료 텍스트의 문장이 진료 특징 관련 문장이면 진료 특징 기반 노이즈 필터를 기반으로 의료 텍스트의 진료 특징 문장에 포함되는 노이즈 단어를 제거할 수 있다.

[0143] 이와 같이, 본 발명은, 원격 진료 대화에 상응하는 음성 데이터를 텍스트 데이터로 변환한 의료 텍스트를 단어별로 임베딩하고 군집화하여 노이즈 단어 데이터를 식별하고, 식별한 노이즈 단어 데이터를 기반으로 새로운 텍스트 데이터를 필터링함으로써, 정확성 및 신뢰성이 높은 의료 텍스트를 제공할 수 있다.

[0144] 이상에서 전술한 본 발명의 일 실시예에 따른 방법은, 하드웨어인 서버와 결합되어 실행되기 위해 프로그램(또는 어플리케이션)으로 구현되어 매체에 저장될 수 있다.

[0145] 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 필요한 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 메모리의 어느 위치(주소 번지)에서 참조되어야 하는지에 대한 메모리 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.

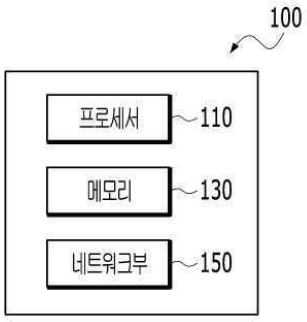
[0146] 상기 저장되는 매체는, 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상기 저장되는 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등이 있지만, 이에 제한되지 않는다. 즉, 상기 프로그램은 상기 컴퓨터가 접속할 수 있는 다양한 서버 상의 다양한 기록매체 또는 사용자의 상기 컴퓨터상의 다양한 기록매체에 저장될 수 있다. 또한, 상기 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장될 수 있다.

[0147] 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 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

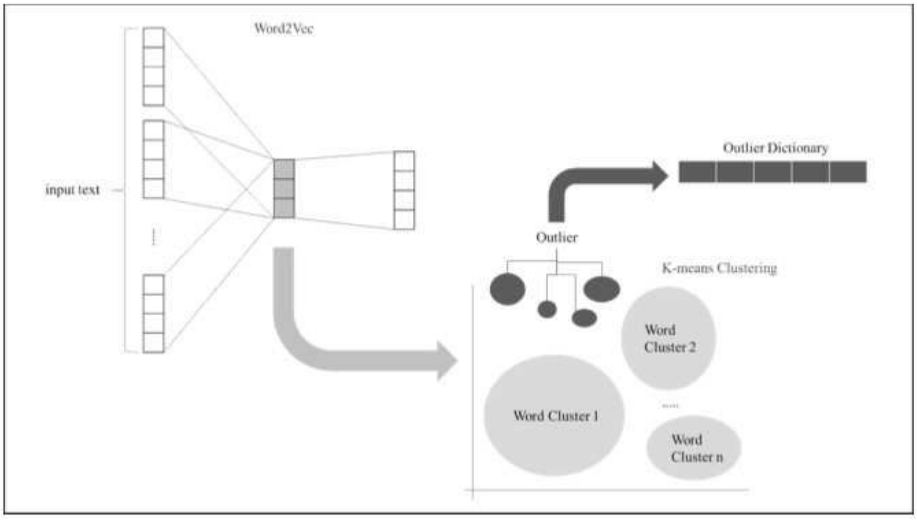
[0148] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

도면

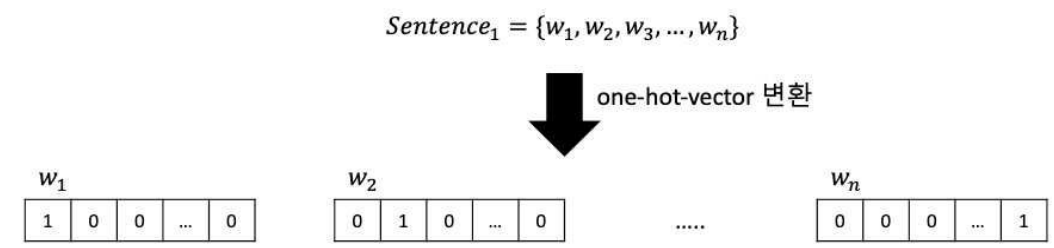
도면1



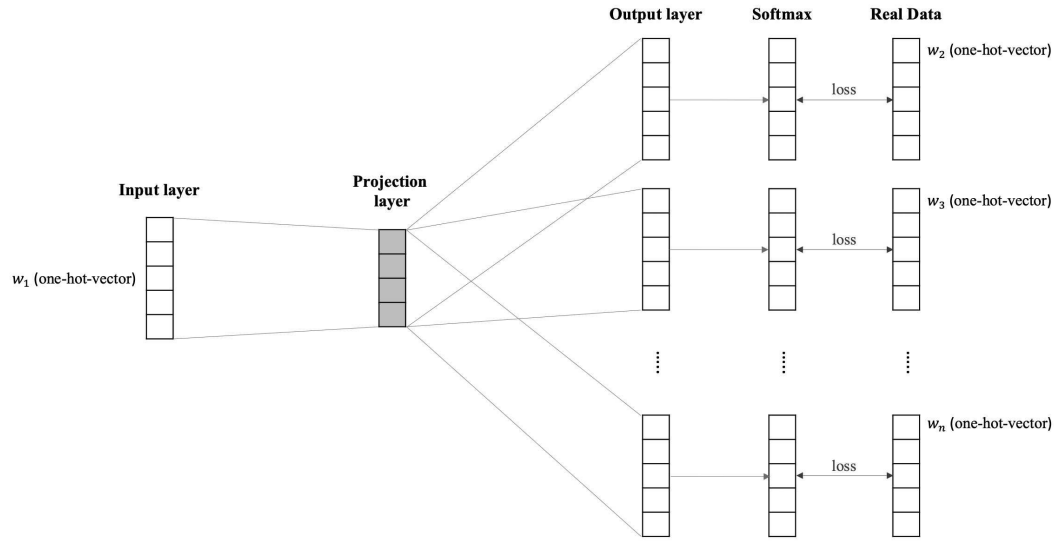
도면2



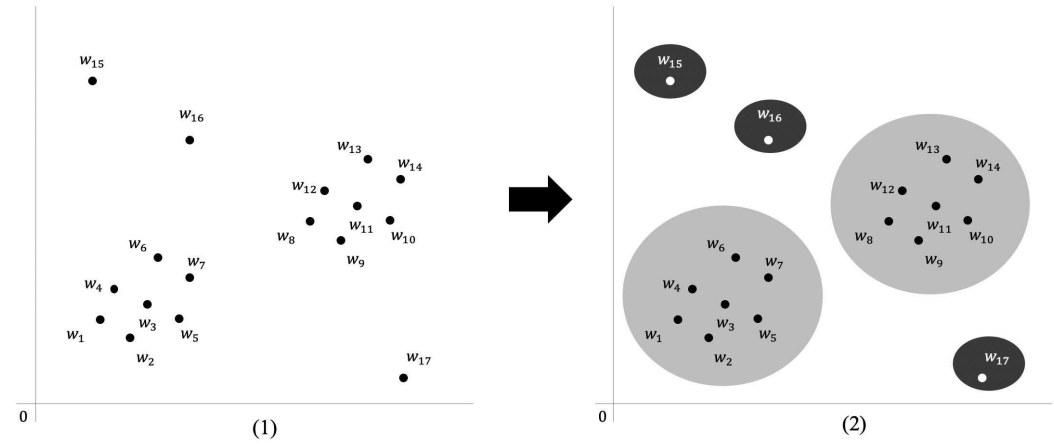
도면3


$Sentence_1 = \{w_1, w_2, w_3, \ldots, w_n\}$
one-hot-vector 변환
w_1
1 0 0 ... 0
w_2
0 1 0 ... 0
.....
w_n
0 0 0 ... 1

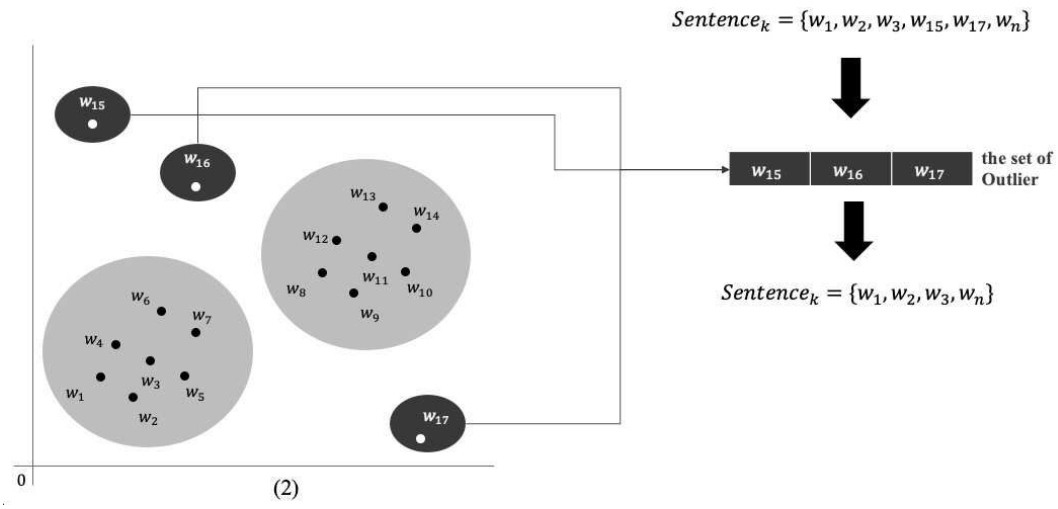
도면4


Output layer
Softmax
Real Data
Input layer
Projection layer
w_1 (one-hot-vector)
w_2 (one-hot-vector)
w_3 (one-hot-vector)
w_n (one-hot-vector)
loss
loss
loss

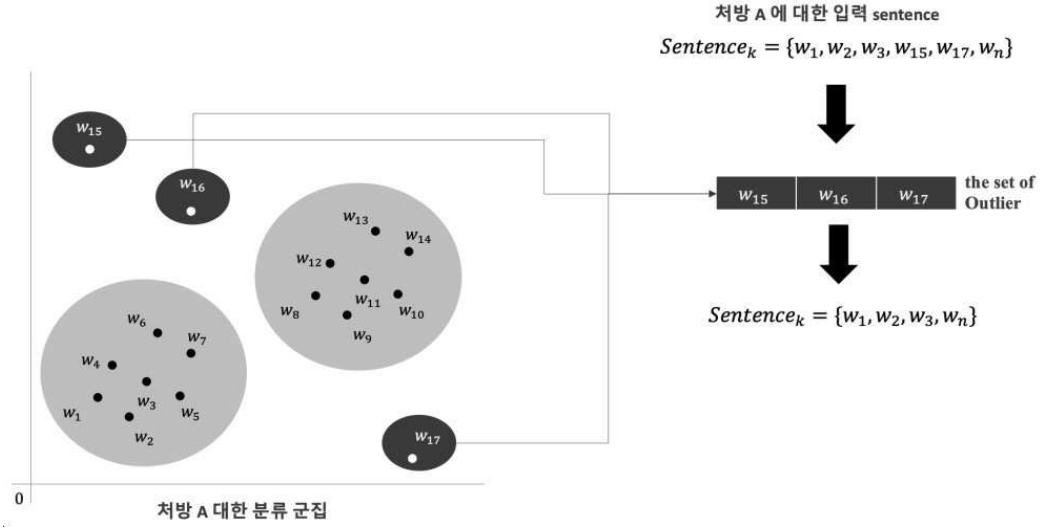
도면5


w_{15}
w_{16}
w_{13}
w_{14}
w_{12}
w_{11}
w_8
w_{10}
w_9
w_6
w_7
w_4
w_3
w_1
w_5
w_2
w_{17}
0
(1)
(2)

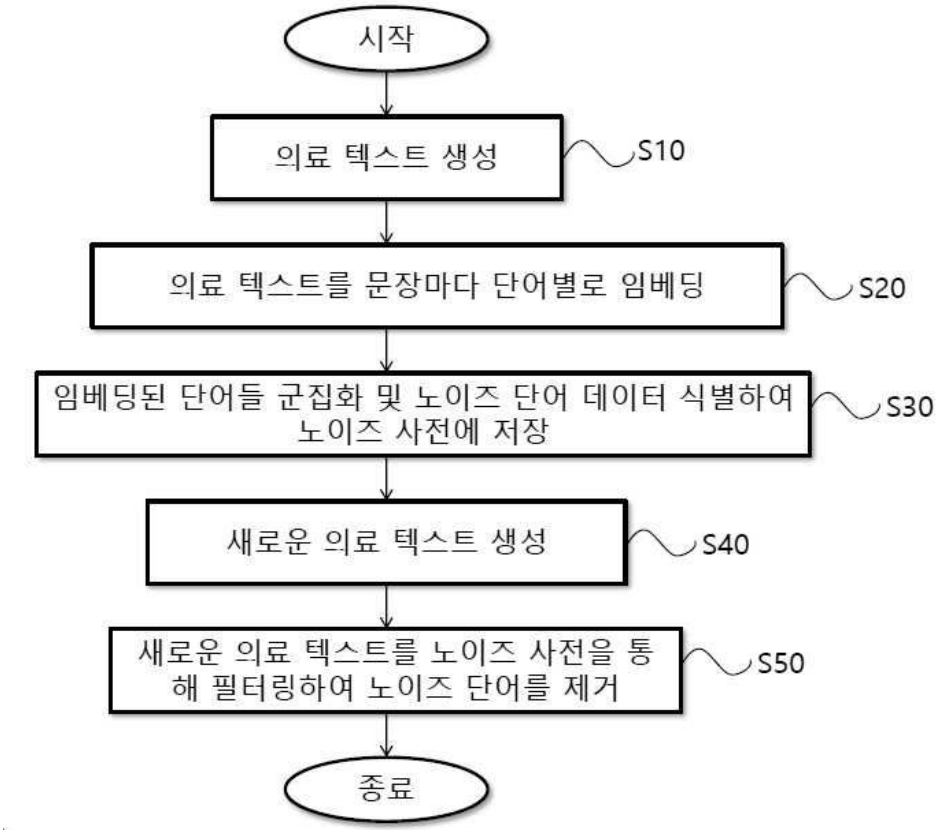
도면6


$Sentence_k = \{w_1, w_2, w_3, w_{15}, w_{17}, w_n\}$
w_{15} w_{16} w_{17}
the set of Outlier
$Sentence_k = \{w_1, w_2, w_3, w_n\}$
0
(2)

도면7


처방 A 에 대한 입력 sentence
$Sentence_k = \{w_1, w_2, w_3, w_{15}, w_{17}, w_n\}$
w_{15} w_{16} w_{17}
the set of Outlier
$Sentence_k = \{w_1, w_2, w_3, w_n\}$
0
처방 A 대한 분류 군집

도면8


시작
의료 텍스트 생성
S10
의료 텍스트를 문장마다 단어별로 임베딩
S20
임베딩된 단어들 군집화 및 노이즈 단어 데이터 식별하여 노이즈 사전에 저장
S30
새로운 의료 텍스트 생성
S40
새로운 의료 텍스트를 노이즈 사전을 통해 필터링하여 노이즈 단어를 제거
S50
종료