



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0063005
(43) 공개일자 2022년05월17일

(51) 국제특허분류(Int. Cl.)

G16H 10/60 (2018.01) G06F 40/211 (2020.01)
G06F 40/253 (2020.01) G06F 40/30 (2020.01)
G06N 20/00 (2019.01) G16H 10/20 (2018.01)
G16H 50/70 (2018.01)

(52) CPC특허분류

G16H 10/60 (2021.08)
G06F 40/211 (2020.01)

(21) 출원번호 10-2020-0148833

(22) 출원일자 2020년11월09일

심사청구일자 2020년11월09일

(71) 출원인

경희대학교 산학협력단

경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)

(72) 발명자

이승룡

경기도 성남시 분당구 구미로144번길 25, 102동 102호(구미동, 삼환빌라)

후세인 무사라트

용인시 기흥구 덕영대로 1732(서천동)

(74) 대리인

유미특허법인

전체 청구항 수 : 총 13 항

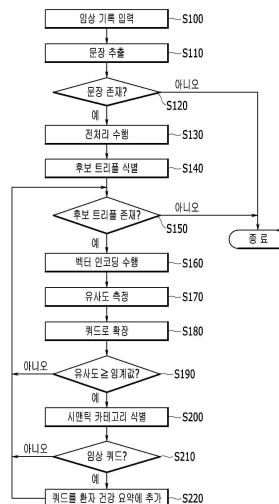
(54) 발명의 명칭 환자 건강 요약 생성 장치 및 그 방법

(57) 요약

환자 건강 요약 생성 장치 및 그 방법이 개시된다.

이 장치는 환자에 대한 임상 기록에 포함된 임상 정보의 내용으로부터 적어도 하나의 문장을 추출하고, 추출된 문장에 대해 품사 태깅을 수행하는 전처리를 수행하는 전처리부를 포함한다. 또한, 이 장치는 전처리부에 의해 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 구문을 추출한 후 미리 설정된 의료 정보를 사용하여 시맨틱 분석을 통해 환자 건강 요약을 생성하는 주처리부를 더 포함한다. 여기서, 인과성 구문 형태의 후보 구문은 전처리부에 의해 전처리가 수행된 문장으로부터 식별된 <명사, 동사, 명사> 형태의 트리플을 나타낸다.

대표도 - 도8



(52) CPC특허분류

G06F 40/253 (2020.01)

G06F 40/30 (2020.01)

G06N 20/00 (2021.08)

G16H 10/20 (2021.08)

G16H 50/70 (2018.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711103238
 과제번호 2017-0-01629-004
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 정보통신기술인력양성/대학ICT연구센터육성지원
 연구과제명 [정부] 2020 ITRC 지능형 의료 플랫폼 개발(4/4)
 기 여 율 3/10
 과제수행기관명 경희대학교 산학협력단
 연구기간 2017.06.01~2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711103071
 과제번호 2017-0-00655-004
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 디지털콘텐츠원천기술개발(R&D, 정보화)
 연구과제명 다양한 디지털 제품의 UX 평가를 위한 Lean UX 핵심기술 및 플랫폼(4/8)
 기 여 율 2/10
 과제수행기관명 경희대학교 산학협력단
 연구기간 2017.04.01~2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711099211
 과제번호 2019R1A2C2090504
 부처명 과학기술정보통신부
 과제관리(전문)기관명 한국연구재단
 연구사업명 (유형1-2)중견연구(연평균연구비 1억원~2억원 이내)
 연구과제명 Big Knowledge 관리를 위한 진화형 지식학습 기술(2/3)
 기 여 율 2/10
 과제수행기관명 경희대학교 산학협력단
 연구기간 2019.09.01~2022.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711099622
 과제번호 2016K1A3A7A03951968
 부처명 과학기술정보통신부
 과제관리(전문)기관명 한국연구재단
 연구사업명 한-EU 공동연구지원사업
 연구과제명 스마트 환경에서 치매환자를 위한 다중 센서기반 상황인지 컴퓨터 기술 개발(4/4)
 기 여 율 1/10
 과제수행기관명 경희대학교 산학협력단
 연구기간 2016.12.01~2020.11.30

이 발명을 지원한 국가연구개발사업

과제고유번호 1711120021
 과제번호 2020-0-01489-001
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 정보통신기술인력양성/Grand연구센터지원
 연구과제명 Grand ICT 연구센터(순천대학교)2
 기 여 율 2/10
 과제수행기관명 순천대학교 산학협력단
 연구기간 2020.07.01~2027.12.31

명세서

청구범위

청구항 1

환자 건강 요약 생성하는 장치로서,

환자에 대한 임상 기록에 포함된 임상 정보의 내용으로부터 적어도 하나의 문장을 추출하고, 추출된 문장에 대해 품사 태깅을 수행하는 전처리를 수행하는 전처리부, 그리고

상기 전처리부에 의해 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 구문을 추출한 후 미리 설정된 의료 정보를 사용하여 시맨틱 분석을 통해 환자 건강 요약을 생성하는 주처리부

를 포함하며,

상기 인과성 구문 형태의 후보 구문은 상기 전처리부에 의해 전처리가 수행된 문장으로부터 식별된 <명사, 동사, 명사> 형태의 트리플을 나타내는,

환자 건강 요약 생성 장치.

청구항 2

제1항에 있어서,

상기 주처리부는,

상기 전처리부에 의해 전처리가 수행된 문장에 존재하는 트리플을 식별하여 후보 트리플로 추출하는 후보 트리플 추출기,

상기 후보 트리플 추출기에 의해 추출되는 후보 트리플이, 상기 후보 트리플을 구성하는 2개의 명사가 각각 원인과 결과를 나타내는 인과 관계를 형성하는 인과성 트리플인 경우, 상기 후보 트리플을 인과성 후보로 분류하는 인과성 후보 분류기,

상기 인과성 후보 분류기에 의해 분류된 인과성 후보를 구성하는 용어의 시맨틱 유형을 결정하여 임상 도메인에 해당하는 인과성 후보를 필터링하는 시맨틱 분석기, 그리고

상기 시맨틱 분석기에 의해 필터링된 인과성 후보를 그룹화하여 상기 환자 건강 요약으로 생성하는 요약 생성기를 포함하는 환자 건강 요약 생성 장치.

청구항 3

제2항에 있어서,

상기 후보 트리플 추출기는,

상기 전처리부에 의해 전처리가 수행된 문장에 존재하는 트리플을 식별하는 후보 트리플 식별기, 그리고

상기 후보 트리플 식별기에 의해 식별된 후보 트리플에 대해 단어 임베딩 기술을 사용하여 벡터 인코딩을 수행하는 벡터 변환기

를 포함하는, 환자 건강 요약 생성 장치.

청구항 4

제2항에 있어서,

상기 인과성 후보 분류기는,

상기 후보 트리플 추출기에 의해 추출되는 후보 트리플에 대해 미리 설정된 인과성 검출 모델을 사용하여 유사도를 측정하고, 측정된 유사도를 포함하여 상기 후보 트리플을 후보 쿼드로 확장하는 유사도 식별기, 그리고

상기 유사도 식별기에 의해 측정된 유사도가 미리 설정된 임계값 이상인 후보 쿼드만을 상기 인과성 후보로 필터링하는 쿼드 필터

를 포함하며,

상기 후보 쿼드는 <명사, 동사, 명사, 유사도>의 형태로 구성되는,

환자 건강 요약 생성 장치.

청구항 5

제4항에 있어서,

상기 인과성 검출 모델은 인공지능 기술을 사용하여 <명사, 동사, 명사> 형태로 구성된 트리플에서 두 개의 명사가 각각 ‘원인’과 ‘결과’에 해당하는 인과관계를 나타내는 인과성 트리플을 포함하도록 학습을 통해 형성되는,

환자 건강 요약 생성 장치.

청구항 6

제4항에 있어서,

상기 유사도 식별기는 상기 인과성 검출 모델을 사용하여 상기 후보 트리플에 대한 유사도 측정시 코사인 유사도 방식을 사용하는,

환자 건강 요약 생성 장치.

청구항 7

제4항에 있어서,

상기 후보 쿼드에 포함되는 유사도는 상기 후보 트리플에 대해 상기 유사도 식별기에 의해 측정되는 하나 이상의 유사도 중에서 최대 값을 갖는 유사도인,

환자 건강 요약 생성 장치.

청구항 8

제2항에 있어서,

상기 시맨틱 분석기는,

상기 인과성 후보 분류기에 의해 분류된 인과성 후보에 대해 미리 설정된 의료 사전을 사용하여 용어의 시맨틱 유형을 식별하는 시맨틱 카테고리 식별기, 그리고

상기 시맨틱 카테고리 식별기에 의해 식별된 시맨틱 유형이 상기 임상 도메인에 해당하지 않는 비 임상 후보를 제거하는 의료 쿼드 필터

를 포함하는, 환자 건강 요약 생성 장치.

청구항 9

제4항에 있어서,

상기 요약 생성기는 상기 시맨틱 분석기에 의해 필터링된 후보 쿼드를 구성하는 <명사, 동사, 명사, 유사도> 형태로부터 <원인 -> 결과, 신뢰도>의 형태인 환자 건강 요약을 생성하며,

상기 원인과 결과는 상기 후보 쿼드를 구성하며 인과관계를 나타내는 2개의 명사 중에서 원인을 나타내는 명사를 상기 원인으로 하고, 결과를 나타내는 명사를 상기 결과로 설정하며, 상기 신뢰도는 상기 후보 쿼드를 구성하는 유사도로서 대체되는,

환자 건강 요약 생성 장치.

청구항 10

제1항에 있어서,
 상기 전처리부는,
 상기 임상 기록에 포함된 임상 정보의 내용으로부터 문장을 추출하는 문장 추출기,
 문장 추출기에 의해 추출된 문장에 대해 중복 텍스트, 특수 문자를 포함하는 내용을 제거하는 정규화기,
 상기 정규화기에 의해 정규화 처리된 문장에 대해 단어 모음을 토큰으로 분할하는 토큰화 처리를 수행하는 토큰화기,
 상기 토큰화기에 의해 토큰화 처리된 각각의 토큰을 기본 형태로 변환하는 표제어 추출기,
 상기 표제어 추출기에 의해 변환된 문장에서 불용어를 제거하는 불용어 필터, 그리고
 상기 불용어 필터에 의해 필터링된 후의 문장에 대해 자연어 처리 파서를 사용하여 각각의 토큰에 품사를 태깅하는 태깅부
 를 포함하는, 환자 건강 요약 생성 장치.

청구항 11

환자 건강 요약을 생성하는 방법으로서,
 환자에 대한 임상 기록에 포함된 임상 정보의 내용으로부터 문장을 추출하는 단계,
 추출된 문장에 대해 품사 태깅을 수행하는 전처리를 수행하는 단계,
 상기 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 트리플을 추출하는 단계,
 추출되는 후보 트리플에 대해 미리 설정된 인과성 검출 모델을 사용하여 유사도를 측정하고, 측정된 유사도를 포함하여 상기 후보 트리플을 후보 쿼드로 확장하는 단계, 그리고
 상기 측정된 유사도가 미리 설정된 임계값 이상인 경우, 미리 설정된 의료 정보를 사용하여 시맨틱 분석을 통해 환자 건강 요약을 생성하는 단계
 를 포함하며,
 상기 후보 트리플은 <명사, 동사, 명사> 형태로 구성되고, 상기 후보 쿼드는 <명사, 동사, 명사, 유사도> 형태로 구성되는,
 환자 건강 요약 생성 방법.

청구항 12

제11항에 있어서,
 상기 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 트리플을 추출하는 단계는,
 상기 후보 트리플에 대해 단어 임베딩 기술을 사용하여 벡터 인코딩을 수행하는 단계를 포함하는,
 환자 건강 요약 생성 방법.

청구항 13

제11항에 있어서,
 상기 환자 건강 요약은 <원인 -> 결과, 신뢰도> 형태로 구성되며, 상기 원인과 결과는 상기 후보 쿼드를 구성하며 인과관계를 나타내는 2개의 명사 중에서 원인을 나타내는 명사를 상기 원인으로 하고, 결과를 나타내는 명사를 상기 결과로 설정하며, 상기 신뢰도는 상기 후보 쿼드를 구성하는 유사도로서 대체되는,
 환자 건강 요약 생성 방법.

발명의 설명

기술 분야

[0001] 본 발명은 환자 건강 요약 생성 장치 및 그 방법에 관한 것이다.

배경 기술

[0002] 임상 기록은 임상적 만남 동안 환자와 의료 서비스 제공자, 예를 들어 의사와의 상호 작용을 표현한다. 이러한 임상 기록에는 보고된 증상, 임상 검사, 결과, 진단 테스트 및 그 결과, 처방된 치료, 및 가족력과 사회적 및 행동 측면에 의해 보완되는 기타 필수 개입이 포함된다.

[0003] 임상 기록의 주요 목적은 시간이 지남에 따라 환자의 건강 상태를 추적하는 것이다. 환자가 의사를 방문할 때마다, 의사는 필요한 조치나 결정을 내리기 전에 먼저 환자의 건강에 대한 완전한 세부 정보를 얻기 위해 환자의 임상 기록을 철저히 확인한다.

[0004] 환자와 관련된 임상 기록의 수량이 인간 전문가 수작업 큐레이션의 임계값을 초과하므로, 현재 시대의 고급 텍스트 마이닝 기술을 사용하는 시스템이 필요하며, 최근까지 임상 기록 마이닝을 위한 다양한 기술이 개발되었다.

[0005] 그러나, 기존의 애플리케이션의 대부분은 질병 상태, 실험실 테스트 결과 또는 약물 관련 정보와 같은 일부 특정 유형의 정보에만 초점을 맞춘다. 반면에, 의사들은 증상, 실험실 테스트, 진단된 질병 및 치료 계획을 포함한 포괄적인 프로파일이 필요하다.

[0006] 따라서, 의사의 기대를 충족하고 환자 건강 프로파일로서 필요한 정보 목록을 그들에게 제공할 수 있는 방안이 요구된다.

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자 하는 과제는 환자 건강 상태를 한 눈에 쉽게 파악할 수 있고 의료 전문가의 시간을 절약할 수 있는 환자 건강 요약 생성 장치 및 그 방법을 제공하는 것이다.

과제의 해결 수단

[0008] 상기한 바와 같은 본 발명의 과제를 달성하고, 후술하는 본 발명의 특징적인 효과를 실현하기 위한, 본 발명의 특징적인 구성은 하기와 같다.

[0009] 본 발명의 일 측면에 따르면, 환자 건강 요약 생성 장치가 제공되며, 이 장치는,

[0010] 환자에 대한 임상 기록에 포함된 임상 정보의 내용으로부터 적어도 하나의 문장을 추출하고, 추출된 문장에 대해 품사 태깅을 수행하는 전처리를 수행하는 전처리부, 그리고 상기 전처리부에 의해 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 구문을 추출한 후 미리 설정된 의료 정보를 사용하여 시맨틱 분석을 통해 환자 건강 요약을 생성하는 주처리부를 포함하며, 상기 인과성 구문 형태의 후보 구문은 상기 전처리부에 의해 전처리가 수행된 문장으로부터 식별된 <명사, 동사, 명사> 형태의 트리플을 나타낸다.

[0011] 본 발명의 다른 측면에 따르면, 환자 건강 요약 생성 방법이 제공되며, 이 방법은,

[0012] 환자에 대한 임상 기록에 포함된 임상 정보의 내용으로부터 적어도 하나의 문장을 추출하는 단계, 추출된 문장에 대해 품사 태깅을 수행하는 전처리를 수행하는 단계, 상기 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 트리플을 추출하는 단계, 추출되는 후보 트리플에 대해 미리 설정된 인과성 검출 모델을 사용하여 유사도를 측정하고, 측정된 유사도를 포함하여 상기 후보 트리플을 후보 쿼드로 확장하는 단계, 그리고 상기 측정된 유사도가 미리 설정된 임계값 이상인 경우, 미리 설정된 의료 정보를 사용하여 시맨틱 분석을 통해 환자 건강 요약을 생성하는 단계를 포함하며, 상기 후보 트리플은 <명사, 동사, 명사> 형태로 구성되고, 상기 후보 쿼드는 <명사, 동사, 명사, 유사도> 형태로 구성된다.

발명의 효과

[0013] 본 발명에 따르면, 의료 서비스 제공자가 환자 건강 상태를 쉽게 파악할 수 있다.

- [0014] 또한, 의료 서비스 제공자가 환자의 임상 기록을 처리하는 시간이 단축됨으로써 의료 서비스 제공자의 시간을 절약할 수 있다.
- [0015] 또한, 환자의 완전한 건강 상태를 제공하여 의료 평등과 효율성을 높일 수 있다.
- [0016] 또한, 다양한 정보를 바탕으로 임상 의사 결정을 지원함으로써 환자 간병 능력을 증가시킬 수 있다.

도면의 간단한 설명

- [0017] 도 1은 본 발명의 실시예에 따른 환자 건강 요약 생성 장치의 개략적인 구성 블록도이다.
- 도 2는 도 1에 도시된 전처리부의 구체적인 구성 블록도이다.
- 도 3은 도 1에 도시된 주처리부의 구체적인 구성 블록도이다.
- 도 4는 본 발명의 실시예에 따라 생성되는 건강 요약의 구성 형태의 예를 도시한 도면이다.
- 도 5는 도 3에 도시된 후보 트리플 추출기의 구체적인 구성 블록도이다.
- 도 6은 도 3에 도시된 인과성 후보 분류기의 구체적인 구성 블록도이다.
- 도 7은 도 3에 도시된 시맨틱 분석기의 구체적인 구성 블록도이다.
- 도 8은 본 발명의 실시예에 따른 환자 건강 요약 생성 방법의 개략적인 흐름도이다.
- 도 9는 본 발명의 다른 실시예에 따른 환자 건강 요약 생성 장치의 개략적인 구성을 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0018] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0019] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.
- [0020] 본 발명에서 설명하는 장치들은 적어도 하나의 프로세서, 메모리 장치, 통신 장치 등을 포함하는 하드웨어로 구성되고, 지정된 장소에 하드웨어와 결합되어 실행되는 프로그램이 저장된다. 하드웨어는 본 발명의 방법을 실행할 수 있는 구성과 성능을 가진다. 프로그램은 도면들을 참고로 설명한 본 발명의 동작 방법을 구현한 명령어(instructions)를 포함하고, 프로세서와 메모리 장치 등의 하드웨어와 결합하여 본 발명을 실행한다.
- [0021] 도 1은 본 발명의 실시예에 따른 환자 건강 요약 생성 장치의 개략적인 구성 블록도이다.
- [0022] 도 1을 참조하면, 본 발명의 실시예에 따른 환자 건강 요약 생성 장치(100)는 저장부(110), 입력부(120), 전처리부(130) 및 주처리부(140)를 포함한다.
- [0023] 저장부(110)는 한 명 이상의 환자 각각에 대한 임상 기록을 저장한다. 이러한 임상 기록은 미리 설정되어 있는 일정한 형식의 데이터로서 기록될 수 있다. 또한, 임상 기록은 환자의 병원 검진 횟수나 병증에 따라 그 양이 다를 수 있다.
- [0024] 또한, 저장부(110)는 주처리부(140)에서 출력되는 환자 건강 요약을 저장할 수 있다.
- [0025] 입력부(120)는 환자 건강 요약을 생성할 대상 환자의 임상 기록을 입력받아서 전처리부(130)로 제공한다.
- [0026] 입력부(120)는 저장부(110)에 저장되어 있는 환자의 임상 기록을 입력받을 수 있지만, 이외에도 외부로부터 환자의 임상 기록을 입력받을 수도 있다.
- [0027] 입력부(120)는 외부로부터 환자의 임상 기록이 입력되는 경우에는 입력된 환자의 임상 기록을 저장부(110)에 저장할 수 있다.

- [0028] 전처리부(130)는 입력부(120)로부터 입력되는 환자의 임상 기록에 포함된 원시 임상 텍스트를 후속 처리가 가능한 포맷으로 변환하는 전처리 작업을 수행한다. 예를 들어, 전처리부(130)에 의해 수행되는 전처리 작업에는 문장 추출 작업, 정규화 작업, 토큰화 작업 등이 포함될 수 있다. 전처리부(130)의 구성에 대해서는 추후 구체적으로 설명한다.
- [0029] 주처리부(140)는 전처리부(130)에 의해 전처리 작업이 수행된 임상 기록으로부터 인과성 구문 형태의 후보 구문을 추출한 후 기존의 의료 정보를 사용한 시맨틱(semantic) 분석을 통해 환자 건강 요약물 생성한다.
- [0030] 주처리부(140)에 의해 생성되는 환자 건강 요약물은 출력부(150)를 통해 외부로 출력된다.
- [0031] 또한, 주처리부(140)는 생성된 환자 건강 요약물을 저장부(110)에 저장할 수 있다.
- [0032] 출력부(150)는 주처리부(140)에서 생성된 환자 건강 요약물을 미리 설정되어 있는 형태로 외부로 출력하거나 또는 외부의 장치로 전송할 수 있다.
- [0033] 도 2는 도 1에 도시된 전처리부(130)의 구체적인 구성 블록도이다.
- [0034] 도 2에 도시된 바와 같이, 전처리부(130)는 문장 추출기(131), 정규화기(132), 토큰화기(133), 표제어 추출기(134), 불용어 필터(135) 및 태깅부(136)를 포함한다.
- [0035] 문장 추출기(131)는 입력부(120)로부터 입력되는 임상 기록에 포함된 임상 텍스트를 판독하여 문장으로 추출한다. 입력부(120)로부터 입력되는 임상 기록에는 집단적 의미를 가진 단어 조합으로 표현된 환자에 대한 완전한 정보가 포함된다. 따라서, 문장 추출기(131)는 임상 기록의 단어 조합을 문장이라고 하는 작은 정보 덩어리로 분할하여 추출한다.
- [0036] 정규화기(132)는 문장 추출기(131)에 의해 추출된 문장들에 대해 중복 텍스트, 관련없는 정보(예를 들어, 환자에게 유용하지만 상황 설정에는 필요하지 않은 괄호 안의 설명 등), 특수 문자(-, + 등)와 같은 구문을 제거한다. 문장 추출기(131)에 의해 추출된 문장은 관련없는 정보를 포함할 수 있으므로, 정규화기(132)에 의해 관련 있는 정보로만 구성되도록 하는 처리가 수행되는 것이다.
- [0037] 토큰화기(133)는 정규화기(132)에 의해 정규화 처리된 문장들에 대해 단어 모음을 토큰으로 분할하는 토큰화 처리를 수행한다. 이러한 토큰화 처리에 대해서는 이미 잘 알려져 있으므로 여기에서는 구체적인 설명을 생략한다.
- [0038] 표제어 추출기(134)는 토큰화기(133)에 의해 토큰화 처리된 각각의 토큰을 기본 형태로 변환한다. 일반적으로, 자연어에서, 단어는 여러 표현을 가질 수 있으며 이는 용어 처리의 복잡성을 증가시킨다. 그러나, 용어의 기본 의미는 동일하게 유지되므로 복잡성을 줄이기 위해 표제어 추출(Lemmatization)이 수행되는 것에 대해서는 이미 잘 알려져 있으므로 여기에서는 구체적인 설명을 생략한다.
- [0039] 불용어 필터(135)는 표제어 추출기(134)에 의해 표제어 추출된 문장에서 부차적인 역할을 하는 중요하지 않은 단어, 즉 불용어(stop word)를 필터링한다.
- [0040] 태깅부(136)는 불용어 필터(135)에 의해 불용어가 필터링된 문장들에 대해 자연어를 처리하는 구문 분석기인 자연어 처리 파서(Parser)를 사용하여 각각의 토큰에 품사를 붙이는 품사(Part of Speech, POS) 태깅을 수행한다.
- [0041] 이와 같이, 전술한 전처리부(130)에 의해 처리된 내용은 환자 건강 요약물의 생성을 위해 다른 컴포넌트에 의해 추가로 처리될 수 있는 문장의 의미론적으로 강화된 토큰이다.
- [0042] 한편, 전처리부(130)는 표제어 추출기(134)에 의해 표제어 추출이 수행된 문장들에 대해 특정 문자 형태로 변환하는 변환기(도시되지 않음)를 더 포함할 수 있다. 예를 들어, 임상 기록에 기재된 임상 텍스트들이 영어로 기재되어 있는 경우 표제어 추출기(134)에 의해 출력되는 문장들에 포함된 모든 토큰을 소문자로 통합하는 변환기가 더 포함될 수 있다.
- [0043] 도 3은 도 1에 도시된 주처리부(140)의 구체적인 구성 블록도이다.
- [0044] 도 3에 도시된 바와 같이, 주처리부(140)는 후보 트리플 추출기(141), 인과성 후보 분류기(142), 시맨틱 분석기(143) 및 요약 생성기(144)를 포함한다.
- [0045] 임상 기록에서 사용된 일부 용어는 환자 건강과 관련하여 다른 용어보다 더 중요할 수 있다. 따라서, 환자 건강 관련 용어는 다른 용어와 구분될 필요가 있으며, 임상 기록의 문장 내에서 가능한 모든 후보 용어를 찾기 위

해 후보 트리플 추출기(141)가 사용된다. 임상 기록에서 가장 중요한 정보는 명사구와 그 뒤를 따르는 동사구와 명사구의 조합으로 제공된다.

- [0046] 따라서, 후보 트리플 추출기(141)는 전처리부(130)에서 전처리되어 출력되는 임상 기록의 문장들에 존재하는 <명사, 동사, 명사> 구문의 형태인 후보 트리플을 추출한다.
- [0047] 인과성 후보 분류기(142)는 미리 설정되어 있는 인과성 검출 모델(145)을 사용하여 후보 트리플 추출기(141)에서 추출된 후보 트리플의 유사도를 측정된 후, 측정된 유사도가 미리 설정된 임계값 이상인 후보 트리플에 대해서만 트리플과 유사도를 포함하는 쿼드(quad)인 인과성 후보로서 분류한다. 예를 들어, 쿼드는 <명사, 동사, 명사, 유사도>의 형태를 포함할 수 있다.
- [0048] 한편, 전술한 인과성 검출 모델(145)은 저장부(110)에 저장되어 있을 수 있으며, 잘 알려져 있는 인공지능 기술, 예를 들어 딥 러닝 기술에 의해 <명사, 동사, 명사>로 구성된 트리플에서 두 개의 명사가 각각 ‘원인’과 ‘결과’를 나타내는 트리플들을 학습을 통해 형성할 수 있다.
- [0049] 인과성 후보 분류기(142)에 의해 분류된 인과성 후보인 쿼드 중 일부는 임상 도메인과 관련이 없을 수 있으며 일반 정보를 나타낼 수 있으므로, 임상 도메인과 관련이 없는 쿼드가 필터링되어야 한다.
- [0050] 시맨틱 분석기(143)는 기존의 의료 사전을 사용하여 쿼드를 구성하는 각각의 용어의 시맨틱 유형을 결정하여 임상 도메인과 관련이 없는 비 임상 쿼드를 필터링한다. 즉, 시맨틱 분석기(143)는 환자 건강 요약에 식별된 쿼드가 제공하는 정보를 의미론적으로 강화한다. 이러한 시맨틱 분석은 각각의 쿼드의 도메인 정보를 보여준다.
- [0051] 요약 생성기(144)는 시맨틱 분석기(143)에 의해 의미론적으로 강화된 임상 쿼드를 그룹화하여 환자 건강 요약으로 생성하여 출력부(150) 및 저장부(110)로 제공한다. 이 때, 요약 생성기(144)는 환자 건강 요약에 쿼드 정보를 모두 포함할 수 있으나, 보다 요약되고 단순화된 형태의 환자 건강 요약을 제공하기 위해, 예를 들어 도 4에 도시된 바와 같이, 쿼드의 정보 중에서 ‘원인’과 ‘결과’를 나타내는 명사들과 유사도에 해당되는 ‘신뢰도’로 구성된 형태로써 임상 정보의 요약을 포함하는 환자 건강 요약을 생성한다. 즉, ‘원인’에 대응하는 ‘결과’와 이에 대한 신뢰도 정보로서 환자 건강 요약이 구성될 수 있다.
- [0052] 이와 같이, 본 발명의 실시예에 따른 환자 건강 요약 생성 장치에 따라 환자별 임상 정보를 포함하는 임상 기록을 대응되는 환자 건강 요약으로 제공함으로써, 환자를 진료하는 의사 등이 환자 건강 상태를 한 눈에 파악하는데 도움이 되므로, 의사를 포함한 의료 전문가의 시간을 절약하고 환자 이해를 도울 수 있다.
- [0053] 도 5는 도 3에 도시된 후보 트리플 추출기(141)의 구체적인 구성 블록도이다.
- [0054] 도 5에 도시된 바와 같이, 후보 트리플 추출기(141)는 후보 트리플 식별기(1411) 및 벡터 변환기(1412)를 포함한다.
- [0055] 후보 트리플 식별기(1411)는 전처리부(130)에서 전처리되어 출력되는 임상 기록의 문장들에 존재하는 트리플, 즉 <명사, 동사, 명사> 구문의 형태를 후보 트리플로 식별한다.
- [0056] 한편, 후보 트리플 식별기(1411)에서 식별된 후보 트리플은 기계에서 직접 처리될 수 없으며 기계에서 이해될 수 있는 인코딩이 필요하다.
- [0057] 벡터 변환기(1412)는 단어 임베딩 기술을 사용하여 후보 트리플 식별기(1411)에서 식별된 트리플을 벡터 형태로 인코딩한다. 여기서, 단어 임베딩 기술은 단어를 밀집한 표현(dense representation)으로 변환하는 방법으로, 자연어 처리에서 주로 사용되는 기술로서 잘 알려져 있으므로 여기에서는 구체적인 설명을 생략한다.
- [0058] 도 6은 도 3에 도시된 인과성 후보 분류기(142)의 구체적인 구성 블록도이다.
- [0059] 도 6에 도시된 바와 같이, 인과성 후보 분류기(142)는 유사도 식별기(1421) 및 쿼드 필터(1422)를 포함한다.
- [0060] 유사도 식별기(1421)는 후보 트리플 추출기(141)에서 추출된 후보 트리플 각각에 대해 인과성 검출 모델(145)을 사용하여 유사도를 측정한다. 이러한 유사도 측정을 사용하여 이미 알려진 용어에 대한 후보의 근접성 및 관련성이 확인될 수 있다. 본 발명의 실시예에서, 유사도 식별기(1421)는 코사인 유사도를 사용하여 후보 트리플 각각에 대한 유사도를 측정하였지만, 이것으로만 한정되는 것은 아니다. 여기서, 코사인 유사도는 텍스트 데이터의 유사도를 구하는 방법 중 하나로, 두 벡터 사이의 코사인 각도를 구해 서로의 유사도를 구하는 방식이다. 코사인 유사도는 데이터의 크기에 관계없이 각도로만 유사도를 판단하기에, 데이터 셋의 길이 차이가 심한 상황

일 때도 데이터들의 유사도를 판단할 수 있다.

- [0061] 또한, 유사도 식별기(1421)는 측정된 최대 유사도를 후보 트리플에 추가하여 <명사, 동사, 명사, 유사도> 형태의 쿼드로 확장한다. 여기서, 최대 유사도는, 후보 트리플에 대해 인과성 검출 모델(145)에 포함된 트리플들과의 유사도가 계산되므로, 계산된 유사도 중에서 최대값을 갖는 유사도를 의미한다.
- [0062] 쿼드 필터(1422)는 유사도 식별기(1421)에서 각각의 후보 트리플에 대해 측정된 유사도가 미리 설정된 임계값 이상인 후보 트리플에 대해서만 트리플과 유사도를 포함하는 쿼드를 인과성 후보로서 분류한다. 여기서, 임계값은 그 값이 높을수록 정확도가 높아지고, 요약에 포함되는 쿼드 수를 제한할 수 있으므로, 통계적이거나 실험적인 방식, 또는 기타 방식을 통해서 적절한 값으로 설정될 수 있다.
- [0063] 도 7은 도 3에 도시된 시맨틱 분석기(143)의 구체적인 구성 블록도이다.
- [0064] 도 7에 도시된 바와 같이, 시맨틱 분석기(143)는 시맨틱 카테고리 식별기(1431) 및 의료 쿼드 필터(1432)를 포함한다.
- [0065] 시맨틱 카테고리 식별기(1431)는 인과성 후보 분류기(142)에서 식별된 후보 쿼드들에 대해 기존의 의료 사전을 사용하여 각각의 용어의 시맨틱 유형을 결정한다. 여기서, 기존의 의료 사전으로는 UMLS(Unified Medical Language System)가 사용될 수 있으나 이것으로만 제한되는 것은 아니다. 여기서, UMLS는 60여개의 의학 관련 시소러스, 분류표 등에 수록된 개념을 연계한 통합 개념 체계로서, 기본 엔트리로 단어, 용어가 아닌 개념을 이용하고 있고, 생물 의학 분야의 다양한 정보 시스템 간의 정보 검색과 통합을 촉진시키기 위한 지식 소스이다.
- [0066] 의료 쿼드 필터(1432)는 시맨틱 카테고리 식별기(1431)에 의해 시맨틱 유형이 결정된 후보 트리플들의 정보가 임상 프로파일로 제한되도록 하기 위해 임상 쿼드가 아닌 비 임상 쿼드를 제거하는 필터링을 수행한다. 즉, 의료 쿼드 필터(1432)는 임상 도메인과 관련있는 임상 쿼드만을 출력한다.
- [0067] 이하, 전술한 환자 건강 요약 생성 장치(100)에서의 설명을 참조하여 본 발명의 실시예에 따른 환자 건강 요약 생성 방법에 대해 설명한다.
- [0068] 도 8은 본 발명의 실시예에 따른 환자 건강 요약 생성 방법의 개략적인 흐름도이다.
- [0069] 설명 전에, 예를 들어, 인공지능 기술을 사용하여 임상 프로파일에서 사용되는 용어들을 사용하여 <명사, 동사, 명사>로 구성된 트리플에서 두 개의 명사가 각각 ‘원인’ 과 ‘결과’ 를 나타내는 트리플을 학습을 통해 판별하여 인과성 검출 모델(145)로서 미리 형성한 것으로 가정한다. 또한, 본 발명의 실시예에 따른 환자 건강 요약 방법은 도 1 내지 도 7을 참조하여 설명한 환자 건강 요약 생성 장치(100)에 의해 수행될 수 있다.
- [0070] 도 8을 참조하면, 먼저 특정 환자에 해당하는 임상 기록이 입력된다(S100). 전술한 바와 같이, 이러한 환자에 대한 임상 기록은 저장부(110)에 저장되어 있거나 외부로부터 입력될 수 있다.
- [0071] 다음, 입력된 임상 기록에 포함된 임상 텍스트를 판독하여 해당되는 문장들을 추출하고(S110), 추출되는 문장이 존재하는지가 판단된다(S120).
- [0072] 만약 추출되는 문장이 없으면, 본 발명의 실시예에 따른 환자 건강 요약 생성 방법이 종료된다. 그러나, 적어도 하나의 문장이라도 존재하면, 해당 문장에 대한 전처리가 수행된다(S130). 이러한 전처리에는 추출된 문장에 대해 중복 텍스트, 관련없는 정보(예를 들어, 환자에게 유용하지만 상황 설정에는 필요하지 않은 괄호 안의 설명 등), 특수 문자(-, + 등)와 같은 구문을 제거하는 정규화 처리, 정규화 처리된 문장들에 대해 단어 모음을 토큰으로 분할하는 토큰화 처리, 토큰화 처리된 각각의 토큰을 기본 형태로 변환하는 표제어 추출 처리, 표제어 추출된 문장에서 부차적인 역할을 하는 중요하지 않은 단어, 즉 불용어를 필터링하는 불용어 필터링 처리, 불용어가 필터링된 문장들에 대해 자연어를 처리하는 구문 분석기인 자연어 처리 파서를 사용하여 각각의 토큰에 품사를 붙이는 품사(Part of Speech, POS) 태깅 처리 등이 포함되며, 이에 대해서는 위에서 이미 설명하였으므로 여기에서는 구체적인 설명을 생략한다.
- [0073] 다음, 전처리된 문장들에 존재하는 <명사, 동사, 명사> 구문의 형태인 트리플을 후보 트리플로 식별한다(S140).
- [0074] 그 후, 식별된 후보 트리플이 존재하는지가 판단되고(S150), 만약 후보 트리플이 존재하지 않으면 본 발명의 실시예에 따른 환자 건강 요약 생성 방법이 종료된다.
- [0075] 그러나, 후보 트리플이 적어도 하나라도 존재하는 것으로 판단되면, 먼저 하나의 후보 트리플에 대해 단어 임베딩 기술을 사용하여 벡터 형태로 인코딩을 수행한다(S160).

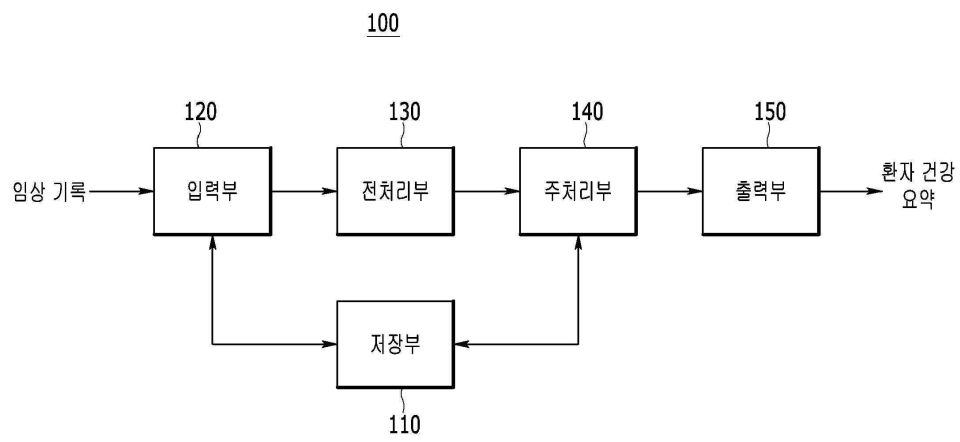
- [0076] 계속해서, 벡터 인코딩된 후보 트리플에 대해 인과성 검출 모델(145)을 사용하여 유사도를 측정한다(S170). 이 때, 측정되는 유사도는 하나의 후보 트리플에 대해 다수로 측정될 수 있다.
- [0077] 그 후, 측정된 유사도 중에서 최대의 값을 갖는 유사도를 포함하도록 후보 트리플이 쿼드로 확장된다(S180). 여기서, 쿼드는 트리플에 최대의 유사도를 추가한 형태, 즉 <명사, 동사, 명사, 유사도>로 확장될 수 있다.
- [0078] 다음, 상기 단계(S170)에서 측정된 유사도 중에서 최대 유사도가 미리 설정된 임계값 이상인지가 판단된다(S190).
- [0079] 만약 유사도, 즉 최대 유사도가 미리 설정된 임계값 이상인 경우에는 인과성 후보로 확장된 쿼드에 대해 UMLS(Unified Medical Language System) 등의 의료 사전을 사용하여 시맨틱 카테고리를 식별하고(S200), 식별된 시맨틱 카테고리가 임상 도메인과 관련있는 임상 쿼드인지가 판단된다(S210).
- [0080] 만약 식별된 시맨틱 카테고리를 통해 임상 도메인과 관련있는 임상 쿼드인 것으로 판단되면, 해당되는 후보 쿼드가 환자 건강 요약으로 추가된다(S220). 이 때, <명사, 동사, 명사, 유사도> 형태의 쿼드의 정보는 도 4에 도시되어 있는 바와 같이 <원인 -> 결과, 신뢰도>의 형태로써 환자 건강 요약에 추가된다. 여기서, '원인'과 '결과'는 이미 전술한 단계(S170)에서 인과성 여부가 판단되었으므로, 두 개의 명사 중에서 '원인'에 해당하는 명사와 '결과'에 해당하는 명사로서 대체되며, '신뢰도'는 쿼드의 유사도로 대체된다.
- [0081] 이와 같이, 상기 단계(140)에서 식별된 하나의 후보 트리플에 대해 대응하는 환자 건강 요약으로의 추가가 완료되면 다른 후보 트리플에 대해서도 마찬가지로의 과정을 통해 환자 건강 요약으로의 추가가 수행될 수 있도록 상기 단계(S150 ~ S220)가 반복 수행된다.
- [0082] 한편, 전술한 단계(S190)에서 유사도, 즉 최대 유사도가 미리 설정된 임계값보다 작은 경우와, 전술한 단계(S210)에서 임상 쿼드가 아닌 경우에도 마찬가지로 다른 후보 트리플에 대해 상기 단계(S150 ~ S220)가 수행될 수 있도록 절차가 진행된다.
- [0083] 이와 같이 모든 후보 트리플에 대해 전술한 단계(S150 ~ S220)가 수행되면 환자 건강 요약에는 임상 쿼드에 대응하는 환자 건강 요약만이 추가되어 있으므로, 최종의 환자 건강 요약이 저장부(110)에 저장되고 또한 의사 등에게 제공될 수 있다.
- [0084] 도 9는 본 발명의 다른 실시예에 따른 환자 건강 요약 생성 장치의 개략적인 구성을 나타내는 도면이다.
- [0085] 도 9에 도시된 바와 같이, 본 발명의 다른 실시예에 따른 환자 건강 요약 생성 장치(200)는 적어도 하나의 프로세서(210), 메모리(220), 통신기(230), 입출력기(240) 및 통신 버스(250)를 포함한다.
- [0086] 프로세서(210)는 범용 CPU(Central Processing Unit), 마이크로프로세서, ASIC(Application-Specific Integrated Circuit), 또는 본 출원의 해결수단에서 프로그램 실행을 제어하기 위한 하나 이상의 집적 회로일 수 있다.
- [0087] 메모리(220)는 본 발명의 실시예에 따른 환자 건강 요약 생성과 관련된 정보를 저장한다.
- [0088] 구체적으로, 메모리(220)는 코드의 집합을 저장하도록 추가로 구성되고, 그 코드는 다음과 같은 프로세스를 실행하기 위해 프로세서(210)를 제어하는 데 사용된다. 이러한 프로세스는, 환자에 대한 임상 기록에 포함된 임상 정보의 내용으로부터 적어도 하나의 문장을 추출하는 프로세스, 추출된 문장에 대해 품사 태깅을 수행하는 전처리를 수행하는 프로세스, 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 트리플을 추출하는 프로세스, 추출되는 후보 트리플에 대해 미리 설정된 인과성 검출 모델을 사용하여 유사도를 측정하는 프로세스, 측정된 유사도를 포함하여 후보 트리플을 후보 쿼드로 확장하는 프로세스, 그리고 측정된 유사도가 미리 설정된 임계값 이상인 경우, 미리 설정된 의료 정보를 사용하여 시맨틱 분석을 통해 환자 건강 요약을 생성하는 프로세스를 포함한다.
- [0089] 여기서, 후보 트리플은 <명사, 동사, 명사> 형태로 구성되고, 후보 쿼드는 <명사, 동사, 명사, 유사도> 형태로 구성된다.
- [0090] 또한, 환자 건강 요약은 <원인 -> 결과, 신뢰도> 형태로 구성되며, 원인과 결과는 후보 쿼드를 구성하며 인과관계를 나타내는 2개의 명사 중에서 원인을 나타내는 명사를 상기 원인으로 하고, 결과를 나타내는 명사를 상기 결과로 설정하며, 신뢰도는 후보 쿼드를 구성하는 유사도로서 대체된다.
- [0091] 또한, 이러한 프로세스는 입출력기(240)를 통해 외부로부터 임상 기록을 입력받는 프로세스, 및 생성된 환자 건강

강 요약을 외부로 출력하는 프로세스를 더 포함한다.

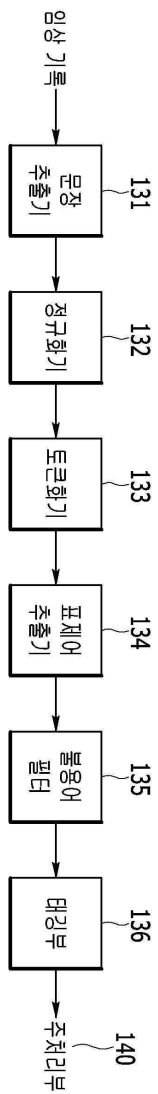
- [0092] 또한, 이러한 프로세스는 통신기(230)를 통해 외부로부터 임상 기록을 입력받는 프로세스, 및 생성된 환자 건강 요약을 외부로 출력하는 프로세스를 더 포함한다.
- [0093] 또한, 전처리가 수행된 문장으로부터 인과성 구문 형태의 후보 트리플을 추출하는 프로세스는, 후보 트리플에 대해 단어 임베딩 기술을 사용하여 벡터 인코딩을 수행하는 프로세스를 포함한다.
- [0094] 메모리(220)는 ROM(Read-Only Memory) 또는 명령을 저장할 수 있는 다른 유형의 정적 저장 장치, 또는 RAM(Random Access Memory) 또는 정보 및 명령을 저장할 수 있는 다른 유형의 동적 저장 장치일 수 있거나, 또는 EEPROM(Electrically Erasable Programmable Read-Only Memory), CD-ROM(Compact Disc Read-Only Memory) 또는 다른 컴팩트 디스크 저장 장치 또는 광 디스크 저장 장치(압축 광 디스크, 레이저 디스크, 광 디스크, 디지털 다용도 디스크, 블루레이 디스크 등을 포함함), 자기 디스크 저장 매체 또는 다른 자기 저장 장치, 또는 명령 또는 데이터 구조의 형태로 예상 프로그램 코드를 운반하거나 저장할 수 있으면서 컴퓨터에 의해 액세스될 수 있는 임의의 다른 매체일 수 있으며, 이것은 제한되지 않는다. 메모리(220)는 독립적으로 존재할 수 있으며, 통신 버스(250)에 의해 프로세서(210)에 연결된다.
- [0095] 통신기(230)는 다른 장치 또는 통신 네트워크와 통신을 수행하며, 다양한 통신 기술로 구현될 수 있다. 즉, 와이파이(WiFi), WCDMA(Wideband CDMA), HSDPA(High Speed Downlink Packet Access), HSUPA(High Speed Uplink Packet Access), HSPA(High Speed Packet Access), 모바일 와이맥스(Mobile WiMAX), 와이브로(WiBro), LTE(Long Term Evolution), 블루투스(bluetooth), 적외선통신(IrDA, infrared data association), NFC(Near Field Communication), 지그비(Zigbee), 무선랜 기술, USB(Universal Serial Bus) 등이 적용될 수 있다. 또한, 인터넷과 연결되어 서비스를 제공하는 경우 인터넷에서 정보 전송을 위한 표준 프로토콜인 TCP/IP를 따를 수 있다.
- [0096] 입출력기(240)는 구체적으로는 입력 장치(241)와 출력 장치(242)로 구성되며, 입력 장치(241)는 프로세서(210)와 통신하고, 복수의 방식으로 사용자의 입력을 수신할 수 있다. 예를 들어, 입력 장치(241)는 마우스, 키보드, 터치 스크린 또는 센싱 장치일 수 있다. 출력 장치(242)는 프로세서(210)와 통신하고, 복수의 방식으로 정보를 디스플레이하거나 음성을 출력할 수 있다. 예를 들어, 출력 장치(242)는 LCD(Liquid Crystal Display, LCD), LED(Light Emitting Diode, LED) 디스플레이, OLED(Organic Light Emitting Diode) 디스플레이, 스피커 등일 수 있다.
- [0097] 통신 버스(250)는 환자 건강 요약 생성 장치(200)의 모든 컴포넌트들, 즉 프로세서(210), 메모리(220), 통신기(230) 및 입출력기(240)를 결합하도록 구성된다.
- [0098] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.
- [0099] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

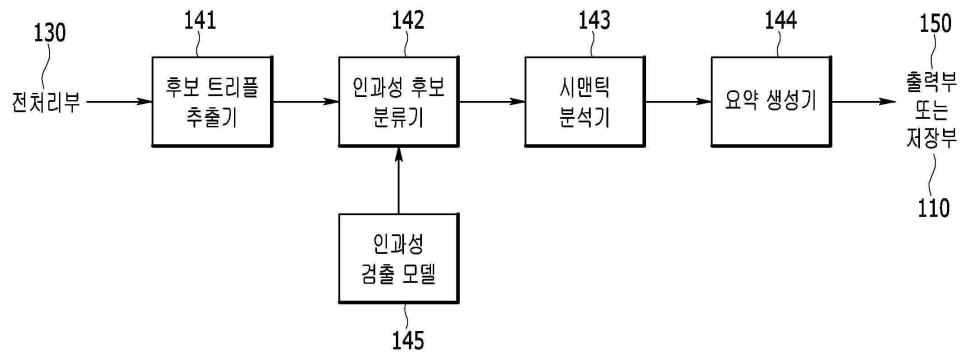
도면1



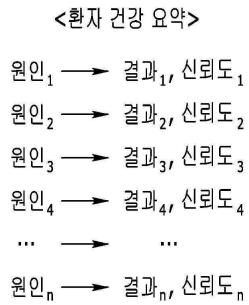
도면2



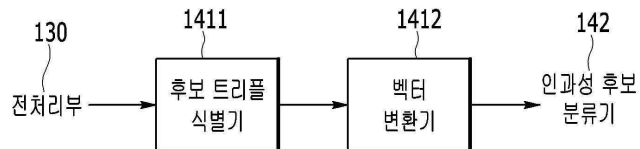
도면3



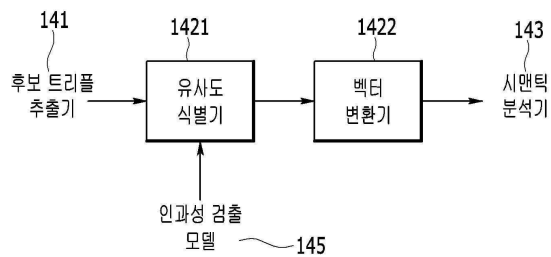
도면4



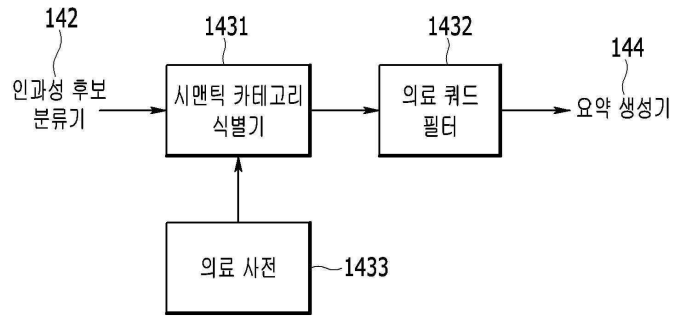
도면5



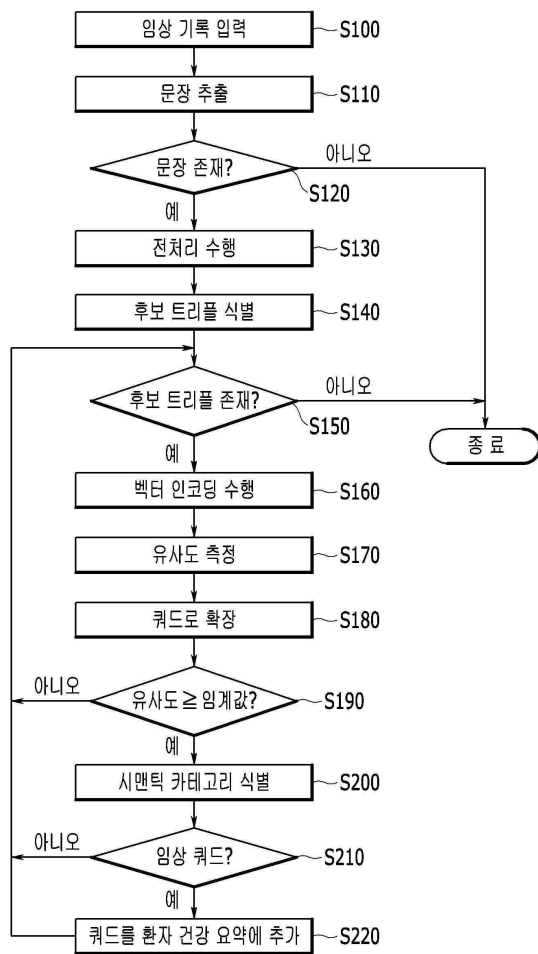
도면6



도면7



도면8



도면9

