



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0093791
(43) 공개일자 2018년08월22일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) G06F 15/18 (2018.01)
(52) CPC특허분류
G06F 17/30424 (2013.01)
G06F 15/18 (2018.05)
(21) 출원번호 10-2018-0005705
(22) 출원일자 2018년01월16일
심사청구일자 2018년01월16일
(30) 우선권주장
1020170020231 2017년02월14일 대한민국(KR)

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
서정연
서울특별시 서초구 서초중앙로 206, B동 203호(반포동, 삼호가든3차)
김주애
서울특별시 은평구 갈현로47길 13, 402호 (갈현동)
고영중
부산광역시 해운대구 대천로 205, 105동 1304호 (좌동, 벽산1차아파트)
(74) 대리인
유미특허법인

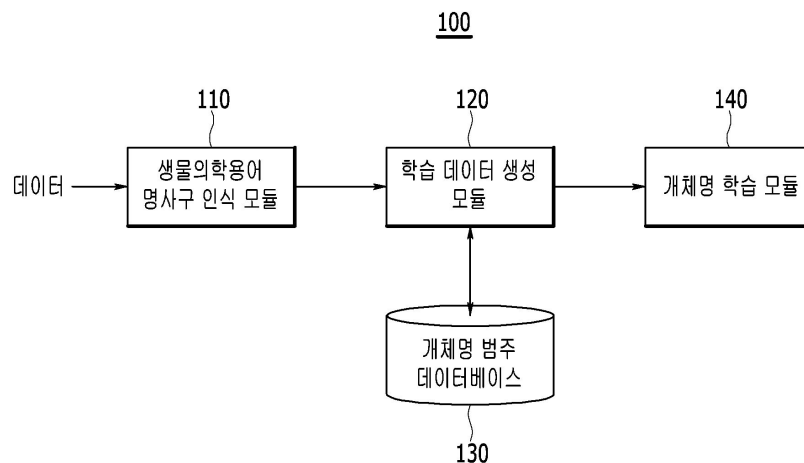
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 생물의학용어 개체명 인식 시스템 및 방법

(57) 요약

생물의학용어 개체명 인식 시스템이 생물의학용어 개체명을 인식하는 방법으로서, 임의의 입력 데이터를 수신하여 적어도 하나의 명사구를 추출하고, 추출한 명사구 각각에 대하여 저장된 개체명 범주를 확인한다. 확인한 개체명 범주의 개수를 토대로, 명사구를 형성하는 어휘 각각에 대응되는 개체명 또는 자동 생성 개체명 중 어느 하나를 태깅하고, 어휘, 상기 어휘에 대응하는 자질 그리고 태깅된 개체명을 포함하는 학습 데이터를 구성한다. 구성된 학습 데이터를 토대로 학습하여, 학습 데이터에서 자동 생성 개체명이 태깅된 어휘에 대한 개체명을 확정하여 개체명을 인식한다.

대표도



이 발명을 지원한 국가연구개발사업

과제고유번호 R2215-16-1003

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 SW중심대학지원사업

연구과제명 SW중심대학 (서강대학교)

기 여 율 1/1

주관기관 정보통신기술진흥센터

연구기간 2016.03.01~2017.02.28

명세서

청구범위

청구항 1

생물의학용어 개체명 인식 시스템이 생물의학용어 개체명을 인식하는 방법으로서,

임의의 입력 데이터를 수신하여 적어도 하나의 명사구를 추출하고, 추출한 명사구 각각에 대하여 저장된 개체명 범주를 확인하는 단계,

상기 확인한 개체명 범주의 개수를 토대로, 명사구를 형성하는 어휘 각각에 대응되는 개체명 또는 자동 생성 개체명 중 어느 하나를 태깅하고, 어휘, 상기 어휘에 대응하는 자질 그리고 태깅된 개체명을 포함하는 학습 데이터를 구성하는 단계, 그리고

상기 구성한 학습 데이터를 토대로 학습하여, 상기 학습 데이터에서 상기 자동 생성 개체명이 태깅된 어휘에 대한 개체명을 확정하는 단계

를 포함하는 생물의학용어 개체명 인식 방법.

청구항 2

제1항에 있어서,

상기 입력 데이터는 적어도 하나의 정답 생물의학용어가 포함되어 있으며, 상기 정답 생물의학용어가 어느 하나의 개체 범주로 분리되지 않은 데이터인 생물의학용어 개체명 인식 방법.

청구항 3

제2항에 있어서,

상기 개체명 범주를 확인하는 단계는,

상기 입력 데이터로부터 상기 적어도 하나의 정답 생물의학용어에 대한 명사구를 추출하는 단계,

상기 추출한 명사구 각각에 대해 미리 저장된 개체명 범주를 매칭하는 단계,

상기 입력 데이터로부터 어휘를 각각 추출하는 단계, 그리고

상기 명사구 및 상기 추출한 어휘에 대한 자질을 확인하는 단계

를 포함하는 생물의학용어 개체명 인식 방법.

청구항 4

제3항에 있어서,

상기 학습 데이터를 구성하는 단계는,

추출한 명사구에 매칭된 개체명 범주의 수가 하나인지 확인하는 단계,

명사구에 매칭된 개체명 범주의 수가 하나이면, 상기 매칭된 개체명 범주를 상기 명사구에 대한 개체명 범주로 결정하는 단계,

상기 명사구를 이루는 복수의 어휘에 상기 결정한 개체명 범주에 따른 개체명을 각각 태깅하는 단계, 그리고

상기 개체명이 태깅된 어휘가 상기 학습 데이터에 포함되도록 구성하는 단계

를 포함하는 생물의학용어 개체명 인식 방법.

청구항 5

제4항에 있어서,

상기 개체명 범주의 수가 하나인지 확인하는 단계는,

상기 추출한 명사구 또는 추출한 어휘에 매칭된 개체명 범주의 수가 하나가 아니라면, 상기 추출한 명사구를 구성하는 복수의 어휘 또는 상기 추출한 어휘에 대하여 미리 설정된 정보가 삽입된 자동 생성 개체명을 태깅하는 단계, 그리고

상기 자동 생성 개체명을 포함하는 어휘가 상기 학습 데이터에 포함되도록 구성하는 단계를 더 포함하는 생물의학용어 개체명 인식 방법.

청구항 6

제5항에 있어서,

상기 학습 데이터는 추출한 어휘에 대한 어휘원형, 상기 어휘의 자질, 태깅된 개체명으로 구성되고,

상기 개체명을 확정하는 단계는,

상기 학습 데이터의 어휘원형과 자질을 기초로 상기 개체명이 추출되는지 학습하여 새로운 학습 데이터를 생성하는 단계,

상기 학습 데이터에서 자동 생성 개체명을 가지는 어휘와 상기 학습하여 생성한 상기 새로운 학습 데이터의 개체명을 비교하는 단계,

상기 새로운 학습 데이터에서 상기 자동 생성 개체명을 가지는 어휘가 특정 개체명으로 확정되어 있으면, 상기 자동 생성 개체명을 상기 확정된 특정 개체명으로 갱신하는 단계, 그리고

상기 특정 개체명으로 갱신한 어휘가 포함되도록 학습 데이터를 구성하는 단계를 포함하는 생물의학용어 개체명 인식 방법.

청구항 7

제6항에 있어서,

상기 개체명 범주를 확정하는 단계는 부트스트래핑 기법으로 개체명 태그를 확정하는 생물의학용어 개체명 인식 방법.

청구항 8

생물의학용어 개체명을 인식하는 시스템으로서,

입력 데이터로부터 적어도 하나의 명사구를 추출하는 생물의학용어 명사구 인식 모듈,

상기 입력 데이터에 포함된 어휘의 자질을 획득하고, 상기 추출한 명사구 각각에 대한 개체명 범주를 획득하고, 상기 획득한 개체명 범주의 개수에 따라 명사구 각각을 구성하는 복수의 어휘에 특정 개체명 또는 자동 생성 개체명 중 어느 하나를 태깅하여 학습 데이터로 생성하는 학습 데이터 생성 모듈, 그리고

상기 학습 데이터를 토대로 학습하여, 상기 자동 생성 개체명이 태깅된 어휘의 개체명을 확정하는 개체명 학습 모듈

을 포함하는 생물의학용어 개체명 인식 시스템.

청구항 9

제8항에 있어서,

생물의학용어에 대한 개체명 범주 정보가 저장되어 있으며, 상기 학습 데이터 생성 모듈로 상기 명사구에 대한 개체명 범주 정보를 제공하는 개체명 범주 데이터베이스

를 포함하며,

상기 개체명 범주 데이터베이스는 메타 맵(Metamap)인 생물의학용어 개체명 인식 시스템.

청구항 10

제8항에 있어서,

상기 학습 데이터 생성 모듈은,

임의의 명사구에 대한 개체명 범주의 개수가 하나이면 상기 개체명 범주를 상기 명사구를 이루는 복수의 어휘에 대한 개체명으로 확정하여 태깅하고,

임의의 명사구에 대한 개체명 범주의 개수가 하나가 아니면, 미리 설정된 정보가 삽입되도록 자동 생성 개체명을 상기 임의의 명사구를 이루는 복수의 어휘에 각각 태깅하는 생물의학용어 개체명 인식 시스템.

청구항 11

제8항에 있어서,

상기 학습 데이터는 추출한 어휘에 대한 어휘원형, 상기 어휘의 자질, 태깅된 개체명으로 구성되고,

상기 개체명 학습 모듈은,

상기 학습 데이터의 어휘원형과 자질을 기초로 상기 개체명이 추출되는지 학습하여 새로운 학습 데이터를 생성하고,

상기 학습 데이터에서 자동 생성 개체명이 태깅된 어휘와 상기 새로운 학습 데이터에서 동일한 어휘에 대한 개체명을 비교하며,

자동 생성 개체명을 가지는 어휘가 특정 개체명으로 확정되어 있으면, 상기 자동 생성 개체명을 상기 특정 개체명으로 갱신하는 생물의학용어 개체명 인식 시스템.

청구항 12

제11항에 있어서,

상기 생물의학용어 개체명 인식 모듈은 부트스트래핑 기법으로 개체명 태그를 확정하는 생물의학용어 개체명 인식 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 생물의학용어 개체명 인식 시스템 및 방법에 관한 것이다.

배경 기술

[0002] 메타 맵(MetaMap)은 생물의학용어 개체명 인식기로 입력되는 일반적인 텍스트에 대해, 입력 텍스트에 속해있는 생물의학용어를 인식하고 미리 정의된 카테고리(예를들어, UMLS(Unified Medical Language System) 시맨틱 타입 또는 TMLS 시맨틱 그룹)를 매칭해주는 프로그램이다. 즉, UMLS 시맨틱 타입은 세계최대의 생물의학용어 사전으로, 일반 텍스트가 입력되면 어떤 어휘가 생물의학용어인지 확인하고 그 어휘가 어느 범주에 속하는지 알 수 있다. 이를 생물의학용어 개체명 인식이라 한다.

[0003] 이러한 메타 맵은 UMLS 시맨틱 타입에 할당된 어휘들을 모두 개체명으로 인식하기 때문에, 생물의학용어가 아닌 일반 용어도 추출되는 경우가 발생한다. 예를 들어, 'Can tetracycline affect tooth formation?'이라는 문장을 메타 맵에 입력하면, 메타 맵은 'affect'와 'formation'도 하나의 개체명으로 인식한다. 그러나, 이 두 개의 단어는 생물의학용어로 판단되지 않는다.

[0004] 또한, 메타 맵은 하나의 용어에 여러 개의 UMLS 시맨틱 그룹을 할당한다. 예를 들어, 'cold'라는 단어는 추위와 감기라는 두 가지 뜻을 갖는다. 따라서 메타 맵은 cold라는 개체명에 감기라는 뜻의 범주인 'Diorders'와 추위라는 뜻의 범주인 'Concepts & Ideas'라는 범주를 모두 할당하기 때문에, 중의성에 문제가 발생한다.

[0005] 또한, 메타 맵은 UMLS 용어 사전에 등재되지 않은 단어에 대해서는, 주변 문맥에 의해 생물의학용어임이 분명함에도 불구하고, 생물의학용어 개체명으로 인식하지 못하는 문제점이 있다.

발명의 내용

해결하려는 과제

[0006] 따라서, 본 발명은 부트스트래핑 기법으로 자동 레이블링한 학습 데이터를 사용하는 생물의학용어 개체명 인식 시스템 및 방법을 제공한다.

과제의 해결 수단

- [0007] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 하나의 특징인 생물의학용어 개체명 인식 시스템이 생물의학용어 개체명을 인식하는 방법으로서,
- [0008] 임의의 입력 데이터를 수신하여 적어도 하나의 명사구를 추출하고, 추출한 명사구 각각에 대하여 저장된 개체명 범주를 확인하는 단계, 상기 확인한 개체명 범주의 개수를 토대로, 명사구를 형성하는 어휘 각각에 대응되는 개체명 또는 자동 생성 개체명 중 어느 하나를 태깅하고, 어휘, 상기 어휘에 대응하는 자질 그리고 태깅된 개체명을 포함하는 학습 데이터를 구성하는 단계, 그리고 상기 구성한 학습 데이터를 토대로 학습하여, 상기 학습 데이터에서 상기 자동 생성 개체명이 태깅된 어휘에 대한 개체명을 확정하는 단계를 포함한다.
- [0009] 상기 개체명 범주를 확인하는 단계는, 상기 입력 데이터로부터 상기 적어도 하나의 정답 생물의학용어에 대한 명사구를 추출하는 단계, 상기 추출한 명사구 각각에 대해 미리 저장된 개체명 범주를 매칭하는 단계, 상기 입력 데이터로부터 어휘를 각각 추출하는 단계, 그리고 상기 명사구 및 상기 추출한 어휘에 대한 자질을 확인하는 단계를 포함할 수 있다.
- [0010] 상기 학습 데이터를 구성하는 단계는, 추출한 명사구에 매칭된 개체명 범주의 수가 하나인지 확인하는 단계, 명사구에 매칭된 개체명 범주의 수가 하나이면, 상기 매칭된 개체명 범주를 상기 명사구에 대한 개체명 범주로 결정하는 단계, 상기 명사구를 이루는 복수의 어휘에 상기 결정한 개체명 범주에 따른 개체명을 각각 태깅하는 단계, 그리고 상기 개체명이 태깅된 어휘가 상기 학습 데이터에 포함되도록 구성하는 단계를 포함할 수 있다.
- [0011] 상기 개체명 범주의 수가 하나인지 확인하는 단계는, 상기 추출한 명사구 또는 추출한 어휘에 매칭된 개체명 범주의 수가 하나가 아니라면, 상기 추출한 명사구를 구성하는 복수의 어휘 또는 상기 추출한 어휘에 대하여 미리 설정된 정보가 삽입된 자동 생성 개체명을 태깅하는 단계, 그리고 상기 자동 생성 개체명을 포함하는 어휘가 상기 학습 데이터에 포함되도록 구성하는 단계를 포함할 수 있다.
- [0012] 상기 학습 데이터는 추출한 어휘에 대한 어휘원형, 상기 어휘의 자질, 태깅된 개체명으로 구성되고,
- [0013] 상기 개체명을 확정하는 단계는, 상기 학습 데이터의 어휘원형과 자질을 기초로 상기 개체명이 추출되는지 학습하여 새로운 학습 데이터를 생성하는 단계, 상기 학습 데이터에서 자동 생성 개체명을 가지는 어휘와 상기 학습하여 생성한 상기 새로운 학습 데이터의 개체명을 비교하는 단계, 상기 새로운 학습 데이터에서 상기 자동 생성 개체명을 가지는 어휘가 특정 개체명으로 확정되어 있으면, 상기 자동 생성 개체명을 상기 확정된 특정 개체명으로 갱신하는 단계, 그리고 상기 특정 개체명으로 갱신한 어휘가 포함되도록 학습 데이터를 구성하는 단계를 포함할 수 있다.
- [0014] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 또 다른 특징인 생물의학용어 개체명을 인식하는 시스템으로서, 입력 데이터로부터 적어도 하나의 명사구를 추출하는 생물의학용어 명사구 인식 모듈, 상기 입력 데이터에 포함된 어휘의 자질을 획득하고, 상기 추출한 명사구 각각에 대한 개체명 범주를 획득하고, 상기 획득한 개체명 범주의 개수에 따라 명사구 각각을 구성하는 복수의 어휘에 특정 개체명 또는 자동 생성 개체명 중 어느 하나를 태깅하여 학습 데이터로 생성하는 학습 데이터 생성 모듈, 그리고 상기 학습 데이터를 토대로 학습하여, 상기 자동 생성 개체명이 태깅된 어휘의 개체명을 확정하는 개체명 학습 모듈을 포함한다.
- [0015] 생물의학용어에 대한 개체명 범주 정보가 저장되어 있으며, 상기 학습 데이터 생성 모듈로 상기 명사구에 대한 개체명 범주 정보를 제공하는 개체명 범주 데이터베이스를 포함할 수 있다.
- [0016] 상기 학습 데이터 생성 모듈은, 임의의 명사구에 대한 개체명 범주의 개수가 하나이면 상기 개체명 범주를 상기 명사구를 이루는 복수의 어휘에 대한 개체명으로 확정하여 태깅하고, 임의의 명사구에 대한 개체명 범주의 개수가 하나가 아니면, 미리 설정된 정보가 삽입되도록 자동 생성 개체명을 상기 임의의 명사구를 이루는 복수의 어휘에 각각 태깅할 수 있다.
- [0017] 상기 학습 데이터는 추출한 어휘에 대한 어휘원형, 상기 어휘의 자질, 태깅된 개체명으로 구성되고,
- [0018] 상기 개체명 학습 모듈은, 상기 학습 데이터의 어휘원형과 자질을 기초로 상기 개체명이 추출되는지 학습하여

새로운 학습 데이터를 생성하고, 상기 학습 데이터에서 자동 생성 개체명이 태깅된 어휘와 상기 새로운 학습 데이터에서 동일한 어휘에 대한 개체명을 비교하며, 자동 생성 개체명을 가지는 어휘가 특정 개체명으로 확정되어 있으면, 상기 자동 생성 개체명을 상기 특정 개체명으로 갱신할 수 있다.

발명의 효과

[0019] 본 발명에 따르면 언레이블드 데이터 기반 자료를 사용하여 개체명을 인식함으로써, 개체명 인식기의 성능을 향상시킬 수 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 실시예에 따른 생물의학용어 개체명 인식 시스템의 구조도이다.

도 2는 발명의 또 다른 실시예에 따른 개체명 인식 시스템의 구조도이다.

도 3은 본 발명의 실시예에 따른 생물의학용어 개체명 인식 방법에 대한 흐름도이다.

도 4는 본 발명의 실시예에 따른 생물의학용어 개체명 학습 방법에 대한 예시도이다.

도 5는 본 발명의 실시예에 따른 생물의학용어 개체명 추출 방법에 대한 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0021] 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.

[0022] 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.

[0023] 이하 도면을 참조로 하여 생물의학용어 개체명을 인식하는 시스템 및 방법에 대해 설명한다. 본 발명의 실시예에서는 설명의 편의를 위하여 개체명 인식 시스템(100)을 크게 개체명을 학습하는 학습부와, 학습된 모듈을 통해 생물의학용어의 개체명을 인식하는 개체명 인식부로 구분지어 설명한다. 먼저 학습부에 해당하는 개체명 인식 시스템에 대해 도 1을 참조로 설명하고, 개체명 인식부에 해당하는 개체명 인식 시스템에 대해 도 2를 참조로 설명한다.

[0024] 도 1은 본 발명의 실시예에 따른 생물의학용어 개체명 인식 시스템의 구조도이다.

[0025] 도 1에 도시된 바와 같이, 생물의학용어 개체명 인식 시스템(100)은 생물의학용어 명사구 인식 모듈(110), 학습 데이터 생성 모듈(120), 개체명 범주 데이터베이스(130) 그리고 개체명 학습 모듈(140)을 포함한다.

[0026] 생물의학용어 명사구 인식 모듈(110)은 대량의 학습 데이터를 생성하기 위해, 임의의 의학문서에서 추출된 데이터를 입력으로 수신한다. 여기서, 임의의 의학문서는 의학과 수의학에 관련된 데이터베이스인 메드라인(medline)에서 추출하는 것을 예로 하여 설명한다.

[0027] 그리고, 수신한 입력 데이터는 해당 데이터 속 복수의 용어들이 어느 하나의 개체 범주로 분리되지 않고 정답 생물의학용어가 태깅되어 있는 데이터인 '언레이블드 데이터(unlabeled data)'인 것을 예로 하여 설명한다. 여기서, 입력 데이터에 정답 생물의학용어를 태깅하는 방법은 어느 하나로 한정하지 않는다.

[0028] 생물의학용어 명사구 인식 모듈(110)은 수신한 입력 데이터에 포함되어 있는 적어도 하나의 정답 생물의학용어인 복수의 명사구를 추출한다. 여기서, 생물의학용어 명사구 인식 모듈(110)이 입력 데이터에 포함된 생물의학용어 명사구를 추출하는 방법은 어느 하나로 한정하지 않는다. 생물의학용어 명사구 인식 모듈(110)은 데이터와 추출한 생물의학용어 명사구를 학습 데이터 생성 모듈(120)로 제공한다.

[0029] 학습 데이터 생성 모듈(120)은 생물의학용어 명사구 인식 모듈(110)로부터 생물의학용어 명사구와 데이터를 수신하면, 데이터 내에 포함되어 있는 모든 어휘를 각각 추출한다. 학습 데이터 생성 모듈(120)이 데이터에서 복수의 명사구를 추출하거나 어휘를 추출하는 방법은 다양한 방법으로 수행할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.

- [0030] 또한, 학습 데이터 생성 모듈(120)은 추출한 어휘에 대한 자질을 확인하여, 추출한 어휘와 일대일로 매핑되게 생성한다. 여기서, 자질은 해당 어휘가 명사인지, 형용사인지, 동사인지, 명사라면 단수형인지 복수형인지 등을 나타내는 정보이다. 이를 위해, 학습 데이터 생성 모듈(120)에는 어휘에 대한 자질 정보가 저장되어 있는 것을 예로 하여 설명하나, 반드시 이와 같이 한정되는 것은 아니다.
- [0031] 학습 데이터 생성 모듈(120)은 추출한 어휘와 생물의학용어 명사구 인식 모듈(110)에서 추출한 생물의학용어 명사구를 개체명 범주 데이터베이스(130)로 전달하여 생물의학용어 명사구에 대한 개체명 범주를 획득한다.
- [0032] 즉, 개체명 범주 데이터베이스(130)는 학습 데이터 생성 모듈(120)로부터 수신한 생물의학용어 명사구를 포함한 모든 어휘에 해당하는 UMLS 시맨틱 그룹을 획득한다. 그리고 획득한 UMLS 시맨틱 그룹을 해당 명사구에 대한 개체명 범주으로써 학습 데이터 생성 모듈(120)에 알린다. 여기서, 개체명 범주 데이터베이스(130)로는 메타 맵(Metamap)을 사용하는 것을 예로 하여 설명한다.
- [0033] 학습 데이터 생성 모듈(120)이 추출한 복수의 어휘를 개체명 범주 데이터베이스(130)에 입력하여, 각 어휘에 해당하는 개체명 범주를 획득하기 위해, 먼저 학습 데이터 생성 모듈(120)은 수신한 복수의 어휘를 기본형으로 변형한다. 이는, 어휘 중 동사나 명사의 경우에는 단수형과 복수형이 있을 수 있고, 동사의 경우에는 과거형으로 기재될 수도 있기 때문이다.
- [0034] 따라서, 학습 데이터 생성 모듈(120)은 생물의학용어 명사구 인식 모듈(110)로부터 수신한 데이터에 포함된 복수의 어휘를 기본형으로 변형한다. 여기서, 어휘의 기본형은 현재 시제의 단수 형태로 하는 것을 예로 하여 설명하나, 반드시 이와 같이 한정되는 것은 아니다.
- [0035] 학습 데이터 생성 모듈(120)은 개체명 범주 데이터베이스(130)를 통해 어휘 각각에 대해 획득한 개체명 범주의 수가 하나인지 또는 복수 개 또는 개체명 범주가 없는지 확인한다. 만약, 어휘에 매칭된 개체명 범주가 하나인 경우, 학습 데이터 생성 모듈(120)은 매칭된 개체명 범주를 해당 어휘에 대한 개체명으로 결정하고, 개체명을 해당 어휘에 태깅한다.
- [0036] 그러나, 획득한 개체명 범주가 복수 개 이거나, 개체명 범주가 없는 경우 학습 데이터 생성 모듈(120)은 해당 어휘에 생물의학용어가 아님을 나타내는 자동 생성 개체명을 태깅한다. 본 발명의 실시예에서는 자동 생성 개체명으로 'O'(Out of chunk)이 삽입되는 것을 예로 하여 설명하나, 반드시 이와 같이 한정되는 것은 아니다.
- [0037] 학습 데이터 생성 모듈(120)은 개체명이 각 어휘에 태깅되면, 학습 데이터를 생성한다. 학습 데이터는 학습 데이터 생성 모듈(120)이 추출한 어휘원형, 어휘에 대응하는 자질, 그리고 태깅한 개체명을 포함한다. 여기서 개체명은 특정 개체명이 태깅되어 있을 수도 있고 자동 생성 개체명이 태깅되어 있을 수도 있다.
- [0038] 개체명 학습 모듈(140)은 학습 데이터 생성 모듈(120)이 생성한 학습 데이터를 수신하여 학습한다. 즉, 개체명 학습 모듈(140)은 학습 데이터를 수신하여, 임의의 어휘원형이 해당 자질을 가질 때, 태깅된 개체명이 추출되도록 학습한다.
- [0039] 여기서, 개체명 학습 모듈(140)은 학습 데이터에 존재하는 중의성 문제와 비인식 대상 어휘 문제를 해결하기 위해, 학습 데이터를 이용하여 생물의학용어 개체명을 재인식한다. 이를 위해, 개체명 학습 모듈(140)은 부트스트래핑 방법을 적용하여 학습 데이터에 포함된 어휘의 개체명을 인식한다.
- [0040] 즉, 개체명 학습 모듈(140)은 부트스트래핑의 여러 단계를 거치면서, 학습 데이터 생성 모듈(120)이 생성한 학습 데이터를 이용하여 각각 다양한 모델로 학습하여 성능을 측정한다. 그리고 성능이 가장 높은 모델을 최종적으로 개체명 학습 모듈(140)로 사용한다. 본 발명의 실시예에서는 개체명 학습 모듈(140)을 CRF(Conditional Random Fields) 학습기로 사용하는 것을 예로 하여 설명하나, 반드시 이와 같이 한정되는 것은 아니다.
- [0041] 개체명 학습 모듈(140)이 학습 데이터를 이용하여 학습할 때 생물의학용어 개체명 인식 성능을 향상시키기 위하여, 학습 데이터 생성 모듈(120)이 입력 데이터로부터 확인한 어휘의 자질을 사용한다. 본 발명의 실시예에서는 생물의학용어 개체명 인식 모듈(140)에서 사용하는 자질을 '언레이블드 데이터 기반 자질'이라 지칭한다.
- [0042] 도 2는 발명의 또 다른 실시예에 따른 개체명 인식 시스템(100)의 구조도이다.
- [0043] 도 2에 도시된 개체명 인식 시스템(100)은 생물의학용어에 대한 개체명을 추출하는 개체명 추출 모듈(150)을 포함한다. 여기서 개체명 추출 모듈(150)은 학습 데이터로 학습한 개체명 학습 모듈(140)으로써, 입력되는 데이터로부터 개체명을 추출하는 모듈을 더 포함한다.
- [0044] 개체명 추출 모듈(150)은 생물의학용어 명사구 인식 모듈(110)이 입력된 데이터에서 생물의학용어 명사구를 추

출하면, 명사구에 대응하는 자질을 추출한다. 개체명 추출 모듈(150)은 이미 학습되어 있기 때문에 추출한 생물의학용어 명사구와 자질을 토대로, 생물의학용어 명사구의 개체명 범주를 확인한다.

- [0045] 그리고 확인한 개체명 범주를 해당 생물의학용어 명사구의 개체명으로 확인한 후, 명사구에 태깅한다. 여기서, 개체명 추출 모듈(150)은 생물의학용어 명사구와 자질 이외에도 명사구 주변의 다른 어휘들을 이용하여 개체명 범주를 확인할 수도 있다.
- [0046] 개체명 추출 모듈(150)은 상기 도 1의 학습 데이터 생성 모듈(120)이 학습 데이터를 생성할 때 이용한 데이터와 동일한 데이터가 생물의학용어 명사구 인식 모듈(110)에 입력될 때, 학습 데이터 생성 모듈(120)에서 개체명을 태깅하지 못한 생물의학용어 명사구를 임의의 개체명으로 확정하여 태깅할 수도 있다.
- [0047] 따라서, 개체명 추출 모듈(150)은 임의의 개체명으로 확정하여 태깅된 생물의학용어 명사구, 해당 명사구의 자질 정보, 확정된 개체명을 포함하는 데이터를 학습 데이터로써 사용하여 추가로 학습할 수도 있다.
- [0048] 이상에서 설명한 생물의학용어 개체명 인식 시스템(100)이 생물의학용어 개체명을 인식하기 위해 개체명 추출 모듈(150)을 학습하고, 학습된 개체명 추출 모듈(150)로부터 개체명을 인식하는 방법에 대해 도 3을 참조로 설명한다. 본 발명의 실시예에서는 설명의 편의를 위하여, 개체명 학습 모듈(140) 대신 개체명 추출 모듈(150)을 포함하는 개체명 인식 시스템(100)이 개체명을 학습하고 인식하는 방법에 대해 설명한다.
- [0049] 도 3은 본 발명의 실시예에 따른 생물의학용어 개체명 인식 방법에 대한 흐름도이다.
- [0050] 도 3에 도시된 바와 같이, 본 발명의 실시예에서는 데이터에서부터 생물의학용어를 추출하기 위해, 생물의학 분야에 최적화된 개체명 단위를 추출하는 생물의학용어 명사구 인식 모듈(110)을 생성한다. 대량의 학습 데이터가 될 일반 텍스트를 메타 맵에 입력하고, 메타 맵에서 출력된 개체명을 바로 정답으로 확정하지 않고 따로 생물의학용어 명사구 인식 모듈(110)을 생성하는 이유는 다음 두 가지 이다.
- [0051] 첫째, UMLS 시맨틱 그룹은 의학용어가 아닌 일반적인 명사나 동사에도 할당되어있기 때문에, 생물의학 분야의 특성을 제대로 반영하기 어렵다. 즉, 메타 맵에 일반 텍스트를 입력하면, 일반적인 명사나 동사까지도 생물의학용어로 출력한다.
- [0052] 둘째, 메타 맵은 문맥이나 단어의 형태적 판단으로 충분히 의학용어임에도 불구하고, UMLS에 등재되어있지 않으면 생물의학용어로 인식하지 않는다.
- [0053] 따라서, 의미역을 결정하는 생물의학용어 명사구 인식 모듈(110)을 구축하고 주변 단어와 추출한 어휘에 대한 자질 정보를 충분히 반영하여, 메타 맵이 갖는 문제점을 해결하고자 한다. 본 발명의 실시예에서는 생물의학용어 명사구 인식 모듈(110)을 구축하기 위해, 입력 데이터를 수신한다(S100).
- [0054] 본 발명의 실시예에서는 정답 생물의학용어가 태깅된 데이터를 입력 데이터로 수신하는 것을 예로 하여 설명한다. 여기서 입력 데이터에 정답 생물의학용어가 태깅되는 방법이나 입력 데이터의 수를 어느 하나로 한정하지 않는다.
- [0055] 입력 데이터는 개체명의 범주 없이 개체명만을 결정하게끔 생성되어 있다. 따라서 입력 데이터를 생물의학용어 명사구 인식 모듈(110)에 입력하면, 개체명 범주 없이 생물의학용어 개체명에 해당하는 명사구만 추출한다(S110). 생물의학용어 명사구 인식 모듈(110)이 명사구를 추출하면, 학습 데이터 생성 모듈(120)은 추출한 명사구를 포함한 모든 어휘의 원형어휘를 확인한다. 그리고 추출한 각 어휘에 대한 자질을 추출한다(S120).
- [0056] 학습 데이터 생성 모듈(120)이 추출한 어휘에 대한 원형어휘와 자질을 추출하고 파악하면, 학습 데이터 생성 모듈(120)은 입력된 명사구를 개체명 범주 데이터베이스(130)에 전달하여, 명사구에 대한 개체명 범주를 확인한다. 개체명 범주 데이터베이스(130)로부터 각각의 명사구에 대한 개체명 범주를 수신한 학습 데이터 생성 모듈(120)은 명사구가 하나의 개체명 범주를 갖는지 확인한다(S130).
- [0057] 만약 명사구 하나가 하나의 개체명 범주를 갖는다면, 학습 데이터 생성 모듈(120)은 개체명 범주 데이터베이스(130)로부터 수신한 개체명 범주를 해당 명사구의 개체명으로 확정한다. 그리고 개체명 범주에 대한 개체명을 명사구를 이루고 있는 각 어휘에 태깅하여(S140), 학습 데이터를 구성한다(S150). 여기서 학습 데이터는 (원형어휘, 자질, 개체명)을 포함한다.
- [0058] 그러나, S130 단계에서 확인한 결과 생물의학용어 명사구의 개체명 범주가 여러 개이거나 또는 아예 존재하지 않는다면, 학습 데이터 생성 모듈(120)은 해당 명사구나 어휘가 생물의학용어가 아닌 것으로 확인한다. 따라서, 학습 데이터 생성 모듈(120)은 임의의 개체명을 자동으로 생성하여, 해당 어휘의 개체명에 태깅한다(S160). 그

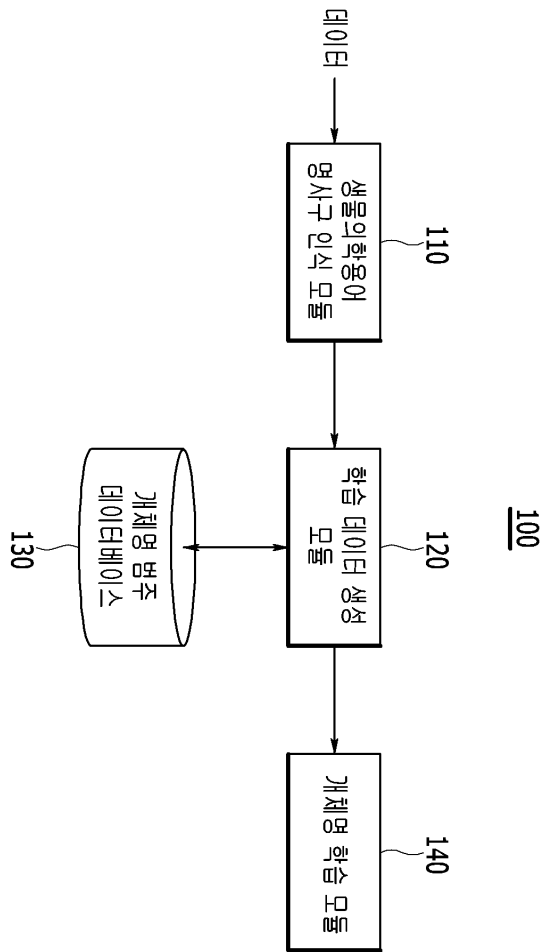
리고 특정 개체명으로 태깅된 어휘와 함께 학습 데이터로 구성한다(S150).

- [0059] S150 단계에서 생성된 학습 데이터는, 개체명이 결정된 어휘도 있지만, 생물의학용어 명사구임에도 불구하고 개체명 범주를 하나로 확정할 수 없는 어휘들이 생물의학용어로 추출되지 않은 용어들도 포함되어 있다. 그러므로 학습 데이터의 정확도는 확실한 에러를 가지게 된다.
- [0060] 이러한 어휘들에 생물의학용어 개체명을 확정하기 위해, 개체명 학습 모듈(140)은 학습 데이터로 학습한다(S170). 본 발명의 실시예에서는 개체명 학습 모듈(140)이 자가 학습 방법을 사용하여, 하나의 개체명 범주를 갖는 생물의학용어 개체명을 확정하기 위하여 부트스트래핑 기법을 적용한다.
- [0061] 여기서, 부트스트래핑은 다음과 같이 진행된다. 먼저 S150 단계에서 생성한 학습 데이터를 이용하여 개체명 학습 모듈(140)은 학습한다. 즉, 개체명 학습 모듈(140)은 수신한 학습 데이터를, S130 단계, S140 단계, 그리고 S160 단계의 절차를 반복 수행하여 학습 데이터를 생성한다(S180). 그리고, 반복 수행하여 생성한 학습 데이터의 결과와 S150 단계에서 구성한 학습 데이터를 비교한다.
- [0062] 학습 데이터에서는 개체명으로 자동 생성 개체명인 '0'이 태깅되어 있지만, S180 단계에서 학습 데이터에는 생물의학용어 개체명이 태깅되어 있다면, S150 단계에서 생성한 학습 데이터를 S180 단계에서 생성한 학습 데이터로 치환하여 새로운 학습 데이터를 생성한다. 새로운 학습 데이터를 이용하여 개체명 학습 모듈(140)은 다시 학습하고, 학습 데이터와 같은 내용의 데이터를 실험하면서 위 과정을 반복한다.
- [0063] 부트스트래핑 기법을 진행하면, 학습 데이터에서 나타나지 않았던 생물의학용어들이 자가 학습 단계를 진행할수록 복원되어 늘어난다. 자가 학습 과정에서 추출한 생물의학용어 개체명은 학습된 단어 자체뿐 아니라, 주변 단어 정보 및 학습된 단어의 자질 정보를 사용한다. 따라서, 중의성 문제와 비 인식 대상어휘 문제가 있었던 어휘들이 올바르게 하나의 개체명 범주를 갖게 되면서 생물의학용어 개체명으로 인식 된다.
- [0064] 본 발명의 실시예에서는 개체명 인식의 성능 향상을 위해 사용하는 자질로 언레이블드 데이터 기반 자질을 사용한다. 언레이블드 데이터 기반 자질은 정답이 태깅되지 않은 대량의 언레이블드 데이터에 대해, 생물의학용어 명사구 인식 모듈(110)을 적용한 결과와, 인식된 명사구들을 학습 데이터 생성 모듈(120)에 입력한 결과 개체명들 중, 하나의 UMLS 시맨틱 타입을 갖고 일정 빈도 이상으로 자주 나타나며 문서 빈도가 높지 않게 나타난 개체명들로 정의된다. 다양한 개체명들 중 언레이블드 데이터 기반 자질을 정의하는 방법에 대해 본 발명의 실시예에서는 상세한 설명을 생략한다.
- [0065] 낮은 빈도와 높지 않은 문서 빈도의 한계점은 검증 실험을 통해 결정한다. 이러한 개체명들에 해당하는 단어에는 언레이블드 데이터 기반 자질을 추가하여 학습한다.
- [0066] 이와 같은 방식으로 언레이블드 데이터 기반의 자질을 사용하는 이유는, 생물의학용어 명사구 인식 모듈(110)과 학습 데이터 생성 모듈(120)에 입력된 데이터로부터 획득한 개체명들은 완전한 정답이 아니지만, 하나의 범주를 가진 개체명으로 자주 나타난다면 이는 정답일 가능성이 높기 때문이다.
- [0067] 이상에서 설명한 방법에 대해 도 4 및 도 5를 참조로 예를 들어 설명한다.
- [0068] 도 4는 본 발명의 실시예에 따른 생물의학용어 개체명 학습 방법에 대한 예시도이고, 도 5는 본 발명의 실시예에 따른 생물의학용어 개체명 추출 방법에 대한 예시도이다.
- [0069] 도 4에 도시된 바와 같이, 데이터로 "How many disulfide bridges has the protein hepcidin got?"이 입력되면, 생물의학용어 명사구 인식 모듈(110)은 생물의학용어로 'disulfide bridges, protein hepcidin'을 추출한다. 여기서 'disulfide bridges'는 메타 맵에 개체명 범주가 'CHEM'과 'DEVI'가 설정되어 있고, 'protein hepcidin'에 대한 개체명 범주는 'CHEM'으로 설정되어 있다.
- [0070] 따라서, 학습 데이터 생성 모듈(120)에서는 개체명 범주 데이터베이스(130)에 저장된 정보를 토대로 'protein'에 대한 개체명 범주와 'hepcidin'에 대한 개체명 범주인 CHEM을 확인한다. 여기서 'protein hepcidin'에 대한 개체명 범주는 CHEM으로 하나가 설정되어 있으므로, 학습 데이터 생성 모듈(120)은 해당 어휘인 protein과 hepcidin의 개체명으로 B-CHEM과 I-CHEM을 태깅한다.
- [0071] 여기서, B-CHEM의 B는 CHEM 범주의 beginning(처음)을, I-CHEM의 I는 inside(중간 또는 끝)를 뜻한다. 이는 Proten hepcidin이 하나의 개체명을 갖기 때문에 구분을 위하여 B-CHEM과 I-CHEM으로 태깅한다. 예를 들어, "deep vein thrombosis"라는 질병의 개체명은 "B-DISO I-DISO I-DISO" 으로 태깅된다. 이와 같이 개체명이 태깅되는 것은 Inside-outside-beginning 태깅 기법을 이용한 것을 예로 하여 설명한다.

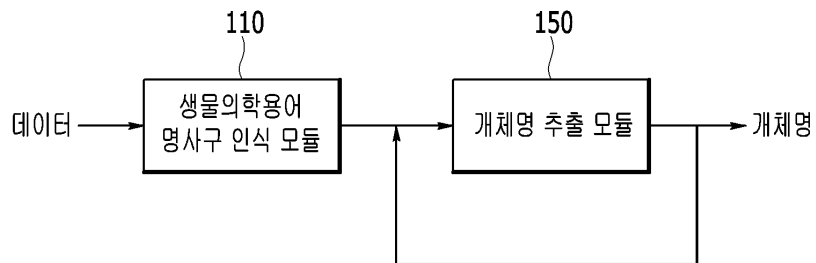
- [0072] 반면, 'disulfide bridges'의 경우에는 두 개의 개체명 범주인 CHEM과 DEVI가 설정되어 있기 때문에, 학습 데이터 생성 모듈(120)은 자동 생성 개체명으로 각 어휘에 태깅한다. 본 발명의 실시예에서는 자동 생성 개체명으로 'O'이 생성되도록 한다. 이 외에 개체명 범주 데이터베이스(130)에 개체명 범주가 정해지지 않은 'How', 'many', 'has', 'the', 'got', '?'의 태깅 정보도 자동 생성 개체명인 'O'를 태깅한다.
- [0073] 학습 데이터 생성 모듈(120) 개체명이 태깅된 데이터 즉, 어휘원형, 어휘의 자질, 태깅된 개체명을 학습 데이터로 생성한다. 본 발명의 실시예에서는 (어휘원형, 자질, 개체명)을 포함하는 데이터를 학습 데이터라 지칭할 수도 있고, 도 4에 도시한 표 자체를 학습 데이터라고 지칭할 수도 있다.
- [0074] 그리고 학습 데이터 생성 모듈(120)이 각 어휘원형에 대응하여 추출한 자질인 WRB는 wh로 시작하는 부사를, JJ는 형용사, NN은 단수 명사, NNS는 복수 명사, VBZ는 3인칭 현재형 동사, DT는 한정사, VBD는 과거 분사를 의미한다. 본 발명의 실시예에서는 설명의 편의를 위하여 자질 표현 방법을 도 4에 도시한 바와 같이 나타내었으나, 반드시 이와 같이 한정하지 않는다.
- [0075] 학습 데이터 생성 모듈(120)이 생성한 학습 데이터를 이용하여 개체명 학습 모듈(140)은 학습한다. 즉, 개체명 학습 모듈(140)은 어휘원형과 자질을 입력으로 받았을 때, 해당 어휘의 개체명이 출력되도록 학습한다.
- [0076] 이와 같이 개체명 학습 모듈(140)이 학습 데이터로 학습하여 생성된 개체명 인식 모듈(150)은 도 5에 도시된 바와 같이 데이터에 포함된 생물의학용어의 개체명을 인식한다.
- [0077] 즉, 상기 도 4의 입력 데이터와 동일한 데이터가 생물의학용어 명사구 인식 모듈(110)로 입력되면, 생물의학용어 명사구 인식 모듈(110)은 생물의학용어 명사구와 어휘들을 추출한다. 그리고 추출한 명사구와 어휘를 수신한 개체명 인식 모듈(150)은 학습된 사항을 토대로 추출한 어휘에 대한 개체명을 인식하여 각 어휘에 태깅한다.
- [0078] 이때, 상기 도 4에서 개체명이 자동 생성 개체명으로 태깅된 어휘인 'disulfie'와 'bridge'들이 도 5에 나타낸 바와 같이 특정 개체명인 'B-CHEM'과 'I-CHEM'으로 인식되어 태깅된 것을 알 수 있다. 따라서, 개체명 인식 모듈(150)은 새롭게 개체명이 태깅된 데이터를 학습 데이터로 삼아 학습할 수 있다.
- [0079] 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

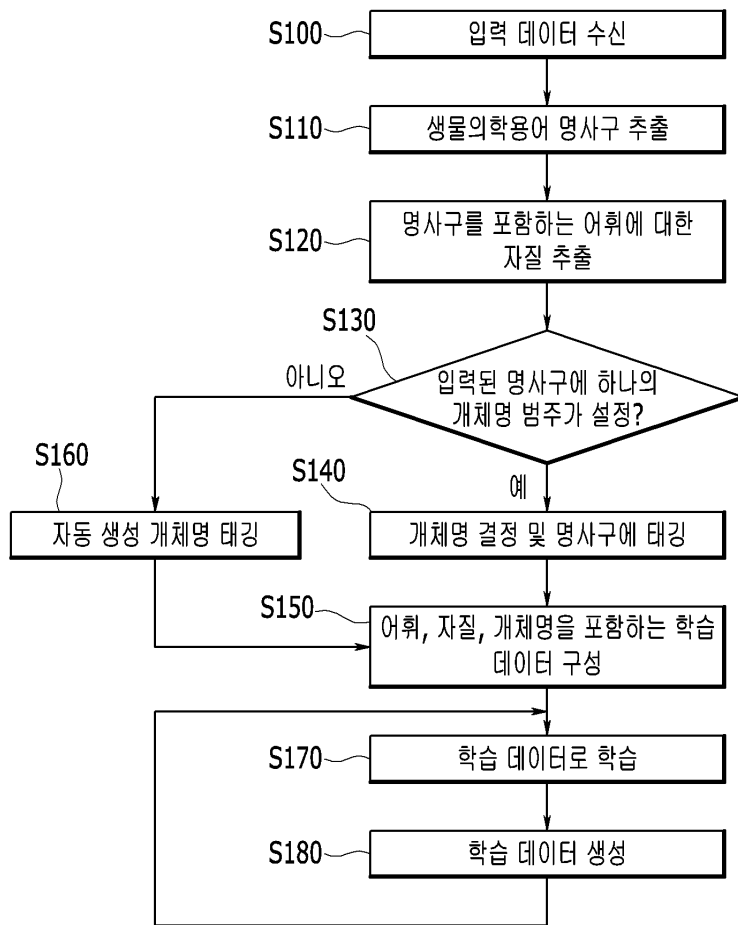
도면1



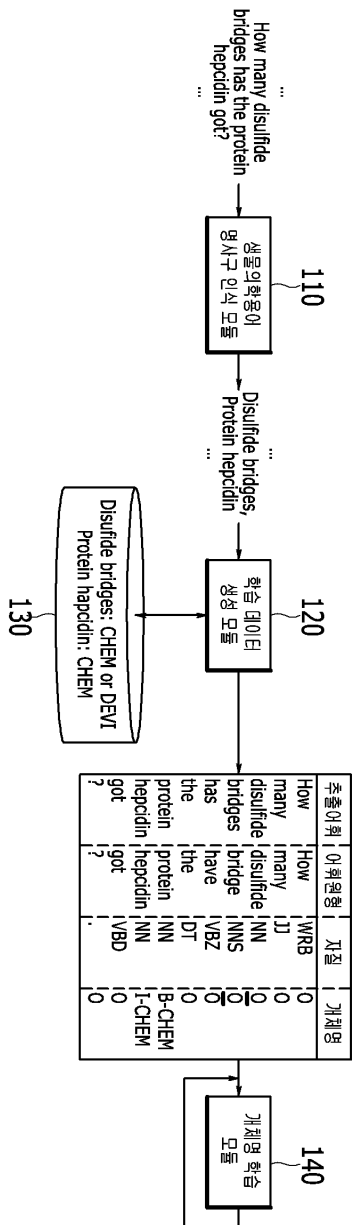
도면2



도면3



도면4



도면5

