



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0080131
(43) 공개일자 2023년06월07일

(51) 국제특허분류(Int. Cl.)
G06F 40/295 (2020.01) G06F 40/284 (2020.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 40/295 (2020.01)
G06F 40/284 (2020.01)
(21) 출원번호 10-2021-0167578
(22) 출원일자 2021년11월29일
심사청구일자 2021년11월29일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암
동5가)
(72) 발명자
김진성
서울특별시 성북구 안암로9가길 85, 204호
김경민
서울특별시 종로구 평창길 31, B동 203호
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애
기능생활관 311호 NLP&AI연구실
(74) 대리인
김홍석

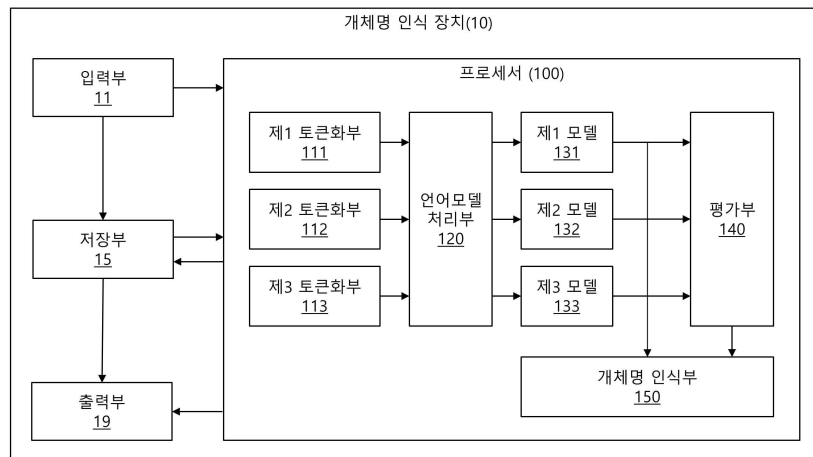
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 개체명 인식 장치 및 방법

(57) 요약

개체명 인식 장치 및 방법에 관한 것으로, 개체명 인식 장치는 대상 텍스트를 획득하는 입력부 및 상기 대상 텍스트에 대해 둘 이상의 토큰화 기법을 적용하여 둘 이상의 토큰화 기법 각각에 대응하는 둘 이상의 토큰화 결과를 획득하고, 상기 둘 이상의 토큰화 결과에 대응하는 둘 이상의 언어 모델을 획득하고, 상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 프로세서를 포함할 수 있다.

대표도 - 도1



(52) CPC특허분류

G06N 20/00 (2021.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126407
과제번호	2018-0-01405-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명세서

청구범위

청구항 1

대상 텍스트를 획득하는 입력부;

상기 대상 텍스트에 대해 둘 이상의 토큰화 기법을 적용하여 둘 이상의 토큰화 기법 각각에 대응하는 둘 이상의 토큰화 결과를 획득하고, 상기 둘 이상의 토큰화 결과에 대응하는 둘 이상의 언어 모델을 획득하고, 상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 프로세서;를 포함하는 개체명 인식 장치.

청구항 2

제1항에 있어서,

상기 둘 이상의 토큰화 기법 중 적어도 하나는, 하나의 단어를 더 분할하여 획득된 작은 단위의 단어인 서브 워드를 이용하여 토큰화를 수행하는 개체명 인식 장치.

청구항 3

제1항에 있어서,

상기 둘 이상의 토큰화 기법은, 바이트 페어 인코딩(Byte Pair Encoding), 워드피스(Wordpiece) 및 센텐스피스(Sentencepiece) 중 적어도 하나를 포함하는 개체명 인식 장치.

청구항 4

제1항에 있어서,

상기 언어 모델은, 사전 학습 언어 모델을 포함하고, 상기 사전 학습 언어 모델은, 버트(BERT), 로버타(RoBERTa: Robustly Optimized BERT Approach), 스팸버트(SpanBERT), 알버트(ALBERT) 및 코버트(KoBERT) 중 적어도 하나를 포함하는 개체명 인식 장치.

청구항 5

제1항에 있어서,

상기 프로세서는, 상기 둘 이상의 언어 모델 각각에 대한 복잡도를 연산하고, 연산된 복잡도가 상대적으로 작은 언어 모델을 상기 최적 언어 모델로 결정하는 개체명 인식 장치.

청구항 6

제5항에 있어서,

상기 복잡도는, 단어 시퀀스 내의 제(i-1) 단어에 대한 제i 단어의 조건부 확률의 역수들의 곱의 N 제곱근을 연산하여 획득되는 개체명 인식 장치.

청구항 7

제1항에 있어서,

상기 프로세서는, 상기 최적 모델을 기반으로 개체명 인식을 수행하는 개체명 인식 장치.

청구항 8

대상 텍스트에 대해 둘 이상의 토큰화 기법을 적용하여 둘 이상의 토큰화 기법 각각에 대응하는 둘 이상의 토큰화 결과를 획득하는 단계;

상기 둘 이상의 토큰화 결과에 대응하는 둘 이상의 언어 모델을 획득하는 단계; 및

상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 단계;를 포함하는 개체명 인식 방법.

청구항 9

제8항에 있어서,

상기 둘 이상의 토큰화 기법 중 적어도 하나는, 하나의 단어를 더 분할하여 획득된 작은 단위의 단어인 서브 워드를 이용하여 토큰화를 수행하는 토큰화 기법을 포함하는 개체명 인식 방법.

청구항 10

제8항에 있어서,

상기 둘 이상의 토큰화 기법은, 바이트 페어, 워드피스 및 센텐스피스 중 적어도 하나를 포함하는 개체명 인식 방법.

청구항 11

제8항에 있어서,

상기 둘 이상의 언어 모델은, 사전 학습 언어 모델을 포함하고, 상기 사전 학습 언어 모델은, 버트, 로버타, 스 판버트, 알버트 및 코버트 중 적어도 하나를 포함하는 개체명 인식 방법.

청구항 12

제8항에 있어서,

상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 단계는,

상기 둘 이상의 언어 모델 각각에 대한 복잡도를 연산하는 단계; 및

연산된 상기 복잡도가 상대적으로 작은 언어 모델을 상기 최적 언어 모델로 결정하는 단계;를 포함하는 개체명 인식 방법.

청구항 13

제12항에 있어서,

상기 둘 이상의 언어 모델 각각에 대한 복잡도를 연산하는 단계는,

상기 둘 이상의 언어 모델 각각에 대해 단어 시퀀스 내의 제(i-1) 단어에 대한 제i 단어의 조건부 확률의 역수들의 곱의 N 제곱근을 연산하여 상기 복잡도를 획득하는 단계;를 포함하는 개체명 인식 방법.

청구항 14

제8항에 있어서,

상기 최적 모델을 기반으로 개체명 인식을 수행하는 단계;를 더 포함하는 개체명 인식 방법.

발명의 설명

기술 분야

[0001]

개체명 인식 장치 및 방법에 관한 것이다.

배경 기술

[0002]

문장의 토큰화(tokenization, 분절)는, 주어진 문장의 이해에 상당히 많은 영향을 미치는 요소이다. 문장의 토큰화가 적절하게 수행되지 않는 경우에는, 해당 문장은 오해석될 가능성이 높으며, 이는 언어 모델이 문장의 자연어 추론, 감성 분석 또는 의미 판단 등의 성능 저하를 야기하게 된다. 그런데, 각국의 언어(영어나 한국어 등)나 이의 표기 방법(알파벳이나 한글 등)은 모두 동일한 것은 아니며, 상당한 차이점을 가지고 있는 경우가 많다. 예를 들어, 한국어는 라틴 계열 언어(영어 등)와 같은 굴절어와 다르게 교착어적 특징을 갖는데, 라틴 계열 언어를 기반으로 제작된 토큰화를 그대로 한국어 기반 문장에 적용하는 경우, 개체명(NE: Named Entity) 인

식에 있어서 어휘 부재(OOV: Out Of Vocabulary)와 같은 문제점이 발생하게 된다. 따라서, 특정한 언어에 대한 토큰화 기법은 다른 언어에 대해서는 그대로 적용하기 어려운 경우가 빈번하며, 이는 토큰화 과정에서 해당 언어의 특징을 고려해야 할 필요성을 낳는다. 그러나, 효과적인 토큰화 기법을 개발 및 구현하는 것은 결코 용이한 것만은 아니다. 특히, 전세계적으로 서로 상이한 다양한 언어 체계 및 표기 방법의 존재는, 이러한 토큰화 기법의 연구 개발을 더욱 더 어렵게 하고 있다.

발명의 내용

해결하려는 과제

[0003] 다양한 토큰화 기법 중에서 처리 대상이 되는 텍스트에 가장 적합한 토큰화 기법을 선택하고, 이를 기반으로 최적의 토큰화를 수행할 수 있는 개체명 인식 장치 및 방법을 제공하는 것을 해결하고자 하는 과제로 한다.

과제의 해결 수단

- [0004] 상술한 과제를 해결하기 위하여 개체명 인식 장치 및 방법이 제공된다.
- [0005] 개체명 인식 장치는, 대상 텍스트를 획득하는 입력부 및 상기 대상 텍스트에 대해 둘 이상의 토큰화 기법을 적용하여 둘 이상의 토큰화 기법 각각에 대응하는 둘 이상의 토큰화 결과를 획득하고, 상기 둘 이상의 토큰화 결과에 대응하는 둘 이상의 언어 모델을 획득하고, 상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 프로세서를 포함할 수 있다.
- [0006] 상기 둘 이상의 토큰화 기법 중 적어도 하나는, 하나의 단어를 더 분할하여 획득된 작은 단위의 단어인 서브 워드를 이용하여 토큰화를 수행하도록 마련된 것일 수 있다.
- [0007] 상기 둘 이상의 토큰화 기법은, 바이트 페어 인코딩(Byte Pair Encoding), 워드피스(Wordpiece) 및 센텐스피스(Sentencepiece) 중 적어도 하나를 포함할 수도 있다.
- [0008] 상기 언어 모델은, 사전 학습 언어 모델을 포함하고, 상기 사전 학습 언어 모델은, 버트(BERT), 로버타(RoBERTa: Robustly Optimized BERT Approach), 스팸버트(SpanBERT), 알버트(ALBERT) 및 코버트(KoBERT) 중 적어도 하나를 포함할 수도 있다.
- [0009] 상기 프로세서는, 상기 둘 이상의 언어 모델 각각에 대한 복잡도를 연산하고, 연산된 복잡도가 상대적으로 작은 언어 모델을 상기 최적 언어 모델로 결정하는 것도 가능하다.
- [0010] 상기 복잡도는, 단어 시퀀스 내의 제(i-1) 단어에 대한 제i 단어의 조건부 확률의 역수들의 곱의 N 제곱근을 연산하여 획득되는 것일 수 있다.
- [0011] 상기 프로세서는, 상기 최적 모델을 기반으로 개체명 인식을 수행하는 것도 가능하다.
- [0012] 개체명 인식 방법은, 대상 텍스트에 대해 둘 이상의 토큰화 기법을 적용하여 둘 이상의 토큰화 기법 각각에 대응하는 둘 이상의 토큰화 결과를 획득하는 단계, 상기 둘 이상의 토큰화 결과에 대응하는 둘 이상의 언어 모델을 획득하는 단계 및 상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 단계를 포함할 수도 있다.
- [0013] 상기 둘 이상의 토큰화 기법 중 적어도 하나는, 하나의 단어를 더 분할하여 획득된 작은 단위의 단어인 서브 워드를 이용하여 토큰화를 수행하는 토큰화 기법을 포함할 수도 있다.
- [0014] 상기 둘 이상의 토큰화 기법은, 바이트 페어, 워드피스 및 센텐스피스 중 적어도 하나를 포함할 수도 있다.
- [0015] 상기 둘 이상의 언어 모델은, 사전 학습 언어 모델을 포함하고, 상기 사전 학습 언어 모델은, 버트, 로버타, 스팸버트, 알버트 및 코버트 중 적어도 하나를 포함할 수도 있다.
- [0016] 상기 둘 이상의 언어 모델을 평가하여 최적 모델을 결정하는 단계는, 상기 둘 이상의 언어 모델 각각에 대한 복잡도를 연산하는 단계 및 연산된 상기 복잡도가 상대적으로 작은 언어 모델을 상기 최적 언어 모델로 결정하는 단계를 포함할 수도 있다.
- [0017] 상기 둘 이상의 언어 모델 각각에 대한 복잡도를 연산하는 단계는, 상기 둘 이상의 언어 모델 각각에 대해 단어 시퀀스 내의 제(i-1) 단어에 대한 제i 단어의 조건부 확률의 역수들의 곱의 N 제곱근을 연산하여 상기 복잡도를 획득하는 단계를 포함할 수도 있다.

[0018] 개체명 인식 방법은, 상기 최적 모델을 기반으로 개체명 인식을 수행하는 단계를 더 포함할 수도 있다.

발명의 효과

[0019] 상술한 개체명 인식 장치 및 방법에 의하면, 다양한 토큰화 기법 중에서 처리 대상이 되는 텍스트에 가장 적합한 토큰화 기법을 선택하고, 이를 기반으로 최적의 토큰화를 수행할 수 있는 효과를 얻을 수 있다.

[0020] 상술한 개체명 인식 장치 및 방법에 의하면, 기존에 사전 학습이 진행된 모델을 이용하지 않고도 다양한 토큰화 기법을 기반으로 사전학습 언어 모델을 생성할 수 있으며, 또한 생성된 언어 모델의 성능을 객관적으로 비교할 수 있게 되는 효과도 얻을 수 있다.

[0021] 상술한 개체명 인식 장치 및 방법에 의하면, 최적의 토큰화 기법을 기반으로 하는 사전학습 언어모델을 획득할 수 있어, 개체명 인식에 있어서 어휘 부재 문제를 해결하고, 개체명 인식 성능의 향상을 도모할 수 있게 되는 장점도 얻을 수 있다.

[0022] 상술한 개체명 인식 장치 및 방법에 의하면, 서로 문법적 구조가 상이한 언어나 문자 기반 텍스트, 일례로 알파벳 기반 텍스트 등이나 한글 기반 텍스트 등에 대해 가장 적절한 토큰화 기법을 결정할 수 있으며, 이에 따라 한글 기반 텍스트에 대한 최적의 사전학습 언어 모델을 생성할 수 있는 장점도 획득할 수 있다.

도면의 간단한 설명

[0023] 도 1은 개체명 인식 장치의 일 실시예에 대한 블록도이다.

도 2는 한국어 및 영어에 있어서의 분절화의 차이점을 설명하기 위한 도면이다.

도 3은 서브 워드의 일례를 설명하기 위한 도면이다.

도 4는 제1 내지 제3 토큰화부의 처리 결과의 일례를 설명하기 위한 도면이다.

도 5는 개체명 인식 방법의 일 실시예에 대한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0024] 이하 명세서 전체에서 동일 참조 부호는 특별한 사정이 없는 한 동일 구성요소를 지칭한다. 이하에서 사용되는 '부'가 부가된 용어는, 소프트웨어 및/또는 하드웨어로 구현될 수 있으며, 실시예에 따라 하나의 '부'가 하나의 물리적 또는 논리적 부품으로 구현되거나, 복수의 '부'가 하나의 물리적 또는 논리적 부품으로 구현되거나, 하나의 '부'가 복수의 물리적 또는 논리적 부품들로 구현되는 것도 가능하다. 명세서 전체에서 어떤 부분이 다른 부분과 연결되어 있다고 할 때, 이는 어떤 부분과 다른 부분이 상호 간에 물리적으로 연결되었음을 의미할 수도 있고, 및/또는 전기적으로 연결되었음을 의미할 수도 있다. 또한, 어떤 부분이 다른 부분을 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 부분 이외의 또 다른 부분을 제외하는 것이 아니며, 설계자의 선택에 따라서 또 다른 부분을 더 포함할 수 있음을 의미한다. 제1 내지 제N(N은 1 이상의 자연수) 등의 표현은, 적어도 하나의 부분(들)을 다른 부분(들)으로부터 구분하기 위한 것으로, 특별한 기재가 없는 이상 이들이 순차적임을 반드시 의미하지는 않는다. 또한 단수의 표현은 문맥상 명백하게 예외가 있지 않는 한, 복수의 표현을 포함할 수 있다.

[0025] 이하 도 1 내지 도 4를 참조하여 개체명 인식 장치의 일 실시예에 대해 설명하도록 한다.

[0026] 도 1은 개체명 인식 장치의 일 실시예에 대한 블록도이다.

[0027] 도 1에 도시된 바에 의하면, 개체명 인식 장치(10)는, 일 실시예에 있어서, 프로세서(100)를 포함할 수 있으며, 입력부(11), 저장부(15) 및 출력부(19) 중 적어도 하나를 포함할 수 있다. 입력부(11), 저장부(15), 출력부(19) 및 프로세서(100) 중 적어도 둘은 상호 전기적으로 연결되어, 데이터나 명령 등을 일방으로 또는 쌍방으로 전달 가능하게 마련된다. 입력부(11), 저장부(15) 및 출력부(19) 중 적어도 하나는, 실시예에 따라, 생략 가능하다.

[0028] 입력부(11)는, 개체명 인식 장치(10)의 동작에 필요한 명령이나, 데이터나, 프로그램(앱, 애플리케이션 또는 소프트웨어 등으로 지칭 가능함) 등을 사용자로부터 또는 다른 외부의 장치(미도시)로부터 수신하여 획득하고, 획득한 데이터나 프로그램 등을 저장부(15)나 프로세서(100) 등으로 전달할 수 있다. 예를 들어, 입력부(11)는 프로세서(100)에 의해 토큰화되거나 및/또는 개체명 인식이 수행될 적어도 하나의 텍스트(적어도 하나의 문장을 포함 가능함, 이하 대상 텍스트)을 수신할 수 있다. 여기서, 적어도 하나의 텍스트는 사용자나 설계자(이하 사

용자 등)이 요구하는 적어도 하나의 언어(일례로 한국어, 영어, 중국어 또는 일본어 등) 및/또는 문자(일례로 한글, 라틴 문자, 그리스 문자, 한자 또는 히라가나 등) 등을 단독으로 또는 조합 이용하여 작성된 것일 수 있다. 또한, 입력부(11)는 프로세서(100)의 제1 내지 제3 토큰화부(111 내지 113)가 이용할 적어도 하나의 토큰화 알고리즘을 입력 받거나, 언어모델 처리부(120)가 이용할 적어도 하나의 언어모델을 입력 받을 수도 있고, 프로세서(100)의 동작 개시, 일례로 토큰화 동작 개시나 개체명 인식 동작의 개시 등과 같은 명령을 사용자 등으로부터 입력 받을 수도 있다. 입력부(11)는, 실시예에 따라서, 개체명 인식 장치(10)와 일체형으로 마련된 것일 수도 있고, 또는 물리적으로 분리 가능하게 마련된 것일 수도 있다. 입력부(11)는, 예를 들어, 키보드, 마우스, 태블릿, 터치 스크린, 터치 패드, 트랙 볼, 트랙패드, 스캐너 장치, 영상 촬영 모듈, 초음파 스캐너, 동작 감지 센서, 진동 센서, 수광 센서, 감압 센서, 근접 센서 및/또는 마이크로 폰 등을 포함할 수 있다. 또한 실시예에 따라, 입력부(11)는, 외부의 다른 장치(일례로 휴대용 메모리 장치 등)로부터 데이터 등의 수신이 가능한 데이터 입출력 단자나, 외부의 다른 장치(일례로 서버용 하드웨어 장치, 스마트폰 또는 데스크톱 컴퓨터 등) 등과 유무선 통신 네트워크를 통해 연결되는 통신 모듈(일례로 랜 카드, 근거리 통신 모듈 또는 이동통신 모듈 등) 등을 포함할 수도 있다.

[0029] 저장부(15)는, 프로세서(100) 등과 같이 개체명 인식 장치(10)의 동작에 필요한 데이터나 프로그램 등을 일시적 또는 비일시적으로 저장할 수 있다. 예를 들어, 저장부(15)는, 입력부(11)를 통해 전달 받은 적어도 하나의 대상 텍스트를 저장할 수도 있고, 프로세서(110)의 처리 과정에서 획득된 적어도 하나의 데이터, 각각의 언어 모델(131 내지 133), 평가 결과 및 개체명 인식 결과 중 적어도 하나를 저장할 수 있다. 또한, 저장부(15)는 제1 내지 제3 토큰화부(111 내지 113)에 대한 적어도 하나의 토큰화 알고리즘이나, 언어모델 처리부(120)에 대한 적어도 하나의 언어 모델 등을 저장하는 것도 가능하다. 저장부(15)는 저장한 데이터나 알고리즘 등을 미리 정의된 설정이나 프로세서(100)의 호출에 따라 프로세서(100)에 제공할 수 있다. 저장부(15)에 저장된 프로그램은 프로그래머 등의 설계자에 의해 직접 작성되거나 수정된 후 저장부(15)에 저장된 것일 수도 있고, 다른 물리적 기록 매체(외장 메모리 장치나 콤팩트 디스크(CD) 등)으로부터 전달받아 저장된 것일 수도 있으며, 및/또는 유무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망을 통하여 획득 또는 갱신된 것일 수도 있다. 저장부(15)는, 예를 들어, 주기억장치(롬이나 램 등) 및 보조기억장치(플래시 메모리 장치(솔리드 스테이트 드라이브(SSD) 등), 하드디스크 드라이브(HDD), SD(Secure Digital) 카드, 콤팩트 디스크(CD), 디브이디(DVD) 또는 플로피 디스크 등) 중 적어도 하나를 포함할 수 있다.

[0030] 출력부(19)는 프로세서(100)의 처리 결과나 저장부(15)에 저장된 데이터나 모델(131 내지 133) 등을 사용자에게 시각적 또는 청각적으로 제공하거나 및/또는 외부의 다른 장치로 직접 또는 유무선 통신 네트워크 등을 통해 전달할 수 있다. 예를 들어, 출력부(19)는 모든 언어 모델(131 내지 133) 또는 최적으로 선택된 언어 모델(131 내지 133 중 적어도 하나)나, 각각의 언어 모델(131 내지 133)의 평가 결과나, 개체명 인식 결과 등을 사용자에게 제공하거나 다른 장치로 전달할 수 있다. 출력부(19)는, 실시예에 따라, 개체명 인식 장치(10)와 일체형으로 마련되거나 물리적으로 분리 가능하게 마련될 수 있다. 출력부(10)는, 예를 들어, 디스플레이, 프린터 장치, 스피커 장치, 영상 출력 단자, 데이터 입출력 단자 및/또는 통신 모듈 등을 포함할 수 있다.

[0031] 프로세서(100)는 대상 텍스트에 대해 적어도 하나의 방법으로 토큰화를 수행하고, 각각의 토큰화 결과를 기반으로 각각의 토큰화 결과에 대응하는 적어도 하나의 모델(131 내지 133)을 생성할 수 있다. 또한, 프로세서(100)는 생성된 적어도 하나의 모델(131 내지 133)에 대한 평가를 수행할 수도 있고, 평가 결과 최적으로 판단된 적어도 하나의 모델(131 내지 133 중 적어도 하나)을 이용하여 대상 텍스트에 대한 개체명 인식을 수행할 수도 있다. 프로세서(100)는, 상술한 동작의 수행을 위해서 저장부(15)에 저장된 프로그램을 실행시켜, 미리 정의된 연산, 판단, 처리 및/또는 제어 동작 등을 수행함으로써, 상술한 동작을 수행할 수도 있다. 프로세서(100)는, 예를 들어, 중앙 처리 장치(CPU: Central Processing Unit), 그래픽 처리 장치(GPU: Graphic Processing Unit), 마이크로 컨트롤러 유닛(MCU: Micro Controller Unit), 애플리케이션 프로세서(AP: Application Processor), 전자 제어 유닛(ECU: Electronic Controlling Unit), 기본보드 관리 컨트롤러(BMC: Baseboard Management Controller), 마이컴(Micom: Micro Processor) 및/또는 이외 각종 연산 및 제어 처리를 수행할 수 있는 적어도 하나의 전자 장치 등을 이용하여 구현될 수 있다. 이들 처리 또는 제어 장치는, 예를 들어, 하나 이상의 반도체 칩, 회로 및/또는 관련 부품 등을 포함할 수 있다.

[0032] 도 2는 한국어 및 영어에 있어서의 분절화의 차이점을 설명하기 위한 도면이다.

[0033] 토큰화는 주어진 텍스트를 소정의 작은 단위(즉, 토큰)로 나누는 것을 의미한다. 예를 들어, 토큰화는, 도 2에 도시된 영문장(31)과 같이 단어 사이의 공백(blank)을 기준으로 수행될 수도 있다. 이 경우, 대상 텍스트 내의 공백에 의해 구분된 각각의 단어(들)은 토큰화에 따라 각각 토큰(들)(31a)으로 분절되고, 이에 따라 적어도 하

나의 토큰(들)이 획득될 수 있게 된다. 개체명 인식은 이러한 토큰화 결과를 기반으로 수행된다. 개체명은 특정한 분류(일례로 인명, 지명, 기관명, 동물, 식물, 사건, 시간 또는 단위 등)로 구분될 수 있는 단어이고, 개체명 인식은 대상 텍스트의 문맥적 의미를 파악하여, 대상 텍스트로부터 개체명(들)을 인식하여 검출하고 검출된 개체명(들)에 해당 개체명(들)의 분류를 나타내는 메타 데이터(일례로 태그나 라벨 등)를 부가(태깅 또는 라벨링으로 표현 가능함)하는 것을 의미한다. 영문장(31)은 보통 단어가 공백을 기준으로 나뉘어져 있어, 공백 기반 토큰화를 통해서도 개체명 인식을 수행할 수 있다. 그런데, 한국어는 교착어로 접사의 역할을 하는 조사 또는 어미가 어근과 결합하여 어절의 의미를 형성하고, 조사나 어미의 변화에 따라 의미가 다양하게 변화하므로, 한국어 문장(32)을 단순히 공백으로 구분하는 경우, 토큰화 결과(32a)가 개체명 인식에 충분하지 않다. 예를 들어, 도 2의 [조선에서]라는 어절이 어휘 집합(vocabulary)에 기록되어 있어 개체명으로 인식될 수 있다고 하더라도, [조선은]과 같이 조사만 바뀐 어절은 어휘 집합에 기록되어 있지 않으면, 이러한 어절은 통상 불명(unknown)으로 처리되어 어휘 부재(OOV)의 문제가 발생할 수 있다. 이는 어휘 집합이 [조선], [에서] 또는 [은]와 같은 단어를 포함한다고 하더라도 마찬가지이다. 다시 말해서, 공백을 기준으로 한 토큰화만으로는 한국어 등의 교착어에 있어서 최적의 토큰화 결과(32b)를 획득할 수 없으며, 이는 이들 언어에 대한 적절한 처리를 어렵게 한다. 프로세서(100)는, 이와 같이 교착어 등에 있어서도 적절하게 토큰화를 수행하고, 이를 기반으로 해당 언어에 부합하는 언어 모델을 획득하도록 할 수 있다.

[0034] 프로세서(100)는, 일 실시예에 따르면, 도 1에 도시된 바와 같이, 적어도 하나의 토큰화부(예를 들어, 제1 토큰화부(111), 제2 토큰화부(112) 및 제3 토큰화부(113) 등), 언어모델처리부(120), 평가부(140) 및 개체명 인식부(150)를 포함할 수 있다. 적어도 하나의 토큰화부(111 내지 113), 언어모델처리부(120), 평가부(140) 및 개체명 인식부(150) 중 적어도 둘은, 실시예에 따라서 논리적으로 구분되는 것일 수도 있고 또는 물리적으로 구분되는 것일 수도 있다. 예를 들어, 적어도 하나의 토큰화부(111 내지 113), 언어모델처리부(120), 평가부(140) 및 개체명 인식부(150)는 하나의 반도체 칩을 이용하여 구현된 것일 수도 있고, 또는 둘 이상의 반도체 칩을 이용하여 구현된 것일 수도 있다. 실시예에 따라, 개체명 인식부(150)는 생략될 수도 있으며, 필요에 따라 개체명 인식 장치(10) 이외의 다른 장치에 포함된 것일 수도 있다.

[0035] 제1 토큰화부 내지 제3 토큰화부(111 내지 113)는, 입력부(11) 또는 저장부(15)로부터 대상 텍스트를 수신하고, 각각 독립적으로 또는 의존적으로 수신한 대상 텍스트에 대한 토큰화를 수행할 수 있다.

[0036] 도 3은 서브 워드의 일례를 설명하기 위한 도면이다.

[0037] 일 실시예에 있어서, 제1 토큰화부(111), 제2 토큰화부(112) 및 제3 토큰화부(113) 중 적어도 하나는 서브 워드(sub word)를 이용하여 토큰화를 수행할 수도 있다. 여기서, 서브 워드는, 어떠한 의미를 갖는 하나의 단어를 더 분할하여 획득된 작은 단위의 단어(또는 문자)를 의미한다. 이들 서브 워드 역시 의미를 가질 수 있다. 서브 워드 각각의 의미는 분할 전의 단어의 의미에 대응될 수도 있으나, 반드시 분할 전의 단어의 의미에 대응하는 것은 아니다. 서브 워드를 이용하는 경우, 제1 내지 제3 토큰화부(111 내지 113) 중 적어도 하나는, 예를 들어, 바이트 페어 인코딩(BPE: Byte Pair Encoding) 기법을 이용하여 서브 워드를 분리할 수도 있다. 이 경우, 제1 내지 제3 토큰화부(111 내지 113) 중 적어도 하나는, 도 3에 도시된 바와 같이, 대상 텍스트를 공백을 통해 분할하는 것 외에도 각각의 단어를 서브 워드로 더 분할함으로써 토큰화를 수행할 수 있다. 보다 구체적으로 예를 들어, 제1 내지 제3 토큰화부(111 내지 113) 중 적어도 하나는, 대상 텍스트 내의 하나 이상의 단어(들)([eleven], [twelve], [thirteen] 및 [fourteen])가 대상 텍스트 내에서 각각 2번, 3번, 4번 및 3번 출현하는 경우, 모든 단어를 단위 글자(43)로 분리하여 하나의 집합을 생성하고, 대상 텍스트 내에서 출현 빈도가 가장 높은 적어도 한 쌍의 단위 글자, 일례로 [en]을 통합하여 적어도 하나의 단위 글자(44)로 생성하고 생성된 단위 글자(44)([en])을 하나의 집합에 부가할 수 있다. 이어서 해당 시점의 단위 글자(43, 44)를 기준으로 출현 빈도가 가장 높은 다른 적어도 한 쌍의 단위 글자, 일례로 [tw]를 통합하여 다른 적어도 하나의 단위 글자(45)를 생성하고 생성된 단위 글자(45)([tw])를 상술한 집합에 더 부가할 수 있다. 이 경우, 다른 적어도 하나의 단위 글자(45)는 단위 글자(43) 간의 조합에 의해 또는 단위 글자(43)와 통합된 단위 글자(44) 간의 조합에 의해 생성된 것일 수 있다. 또한, 계속해서 해당 시점의 단위 글자(43, 44 및/또는 45)를 기준으로 출현 빈도가 가장 높은 또 다른 적어도 한 쌍의 단위 글자, 일례로 [el]을 통합하여 또 다른 적어도 하나의 단위 글자(46)를 새로 생성한 후, 이를 상술한 집합에 더 부가할 수 있다. 이와 같은 과정은 미리 정의된 바에 따라 일 회 이상 반복 수행될 수 있다. 이에 따라, 다수의 단위 글자(43 내지 46), 일례로 [e, l, v, n, t, w, h, l, r, f, o, u, en, tw, el, ...] 등을 포함하는 적어도 하나의 서브 워드가 획득될 수 있으며, 제1 내지 제3 토큰화부(111 내지 113) 중 적어도 하나는 이와 같이 획득된 서브 워드를 기반으로 토큰화를 수행할 수 있다.

- [0038] 도 4는 제1 내지 제3 토큰화부의 처리 결과의 일례를 설명하기 위한 도면이다.
- [0039] 일 실시예에 따르면, 제1 토큰화부(111), 제2 토큰화부(112) 및 제3 토큰화부(113)는, 각각 서로 상이한 토큰화 기법을 이용하여 대상 텍스트에 대한 토큰화를 수행하도록 마련된 것일 수도 있다. 예를 들어, 제1 토큰화부(111)는 상술한 바이트 페어 인코딩 기법을 기반으로 대상 텍스트(51)에 대한 토큰화를 수행하여 분절화된 텍스트(52), 즉 적어도 하나의 토큰(들)(52)을 획득하고, 제2 토큰화부(112)는 다른 토큰화 기법, 일례로 워드피스 모델(WPM: Wordpiece Model)을 기반으로 대상 텍스트(51)에 대한 토큰화를 수행하여 적어도 하나의 다른 토큰(들)(53)을 획득하며, 제3 토큰화부(113)는 또 다른 토큰화 기법, 일례로 센텐스피스 모델(Sentencepiece Model)을 기반으로 대상 텍스트(51)에 대한 토큰화를 수행하여 적어도 하나의 또 다른 토큰(들)(54)을 획득할 수도 있다. 여기서, 워드 피스 모델은, 상술한 바이트 페어 인코딩 기법과 상이하게 출현 빈도가 아니라 말뭉치의 우도(Likelihood)를 최대로 하는 단위 글자 쌍을 통합하여 서브 워드에 대한 토큰화를 수행하는 방법으로, 우도의 최대화를 위해 소정의 언어 모델을 이용한다. 센텐스피스 모델은 원시 문장(Raw sentence)에 빈번하게 등장하는 단어(들)을 형태소로 인식하여 모델이 스스로 형태소를 분리하여 검출할 수 있도록 훈련된 비지도 학습 모델이다. 이와 같이 제1 토큰화부(111), 제2 토큰화부(112) 및 제3 토큰화부(113)는 각각 서로 상이한 토큰화 기법을 이용하므로, 각각 별도의 토큰화 결과(즉, 제1 내지 제3 토큰화부(111 내지 113) 각각에 대응하는 제1 내지 제3 토큰화 결과)를 출력한다. 이 경우, 제1 토큰화부(111), 제2 토큰화부(112) 및 제3 토큰화부(113) 각각이 출력하는 토큰화 결과(즉, 토큰(들))의 전부 또는 일부는, 대상 텍스트(51)에 따라 동일할 수도 있고 상이할 수도 있다. 제1 토큰화부(111)의 제1 토큰화 결과, 제2 토큰화부(112)의 제2 토큰화 결과 및 제3 토큰화부(113)의 제3 토큰화 결과 중 적어도 하나는 언어모델 처리부(120)로 전달될 수 있다.
- [0040] 실시예에 따라서, 프로세서(100)는, 제3 토큰화부(113)를 포함하지 않을 수도 있다. 또한, 프로세서(100)는 상술한 토큰화부(111 내지 113) 외에도 적어도 하나의 다른 토큰화부(미도시)를 더 포함하는 것도 가능하다. 이 경우, 적어도 하나의 다른 토큰화부는, 상술한 제1 내지 제3 토큰화부(111 내지 113) 각각이 이용하는 토큰화 기법과 상이한 토큰화 기법을 이용하여 대상 텍스트에 대한 토큰화를 수행하도록 마련된 것일 수 있다.
- [0041] 언어 모델 처리부(120)는 제1 토큰화부(111)의 제1 토큰화 결과, 제2 토큰화부(112)의 제2 토큰화 결과 및/또는 제3 토큰화부(113)의 제3 토큰화 결과를 수신하고, 제1 토큰화 결과, 제2 토큰화 결과 및/또는 제3 토큰화 결과를 기반으로 학습을 수행하여, 제1 토큰화 결과에 대응하는 언어 모델(131, 이하 제1 모델), 제2 토큰화 결과에 대응하는 언어 모델(132, 이하 제2 모델) 및/또는 제3 토큰화 결과에 대응하는 언어 모델(133, 이하 제3 모델)을 획득할 수 있다. 실시예에 따라, 제3 토큰화부(113)가 생략된 경우에는, 언어 모델 처리부(120)는 제1 토큰화 결과에 대응하는 제1 모델(131) 및 제2 토큰화 결과에 대응하는 제2 모델(132)만 획득할 수 있다. 또한, 제1 내지 제3 토큰화부(111 내지 113) 이외에 다른 제k 토큰화부(k는 4 이상의 자연수)가 더 마련된 경우에는, 언어 모델 처리부(120)는 제1 내지 제3 모델(131 내지 133) 이외에 제k 토큰화부의 제k 토큰화 결과에 대응하는 또 다른 제k 모델을 더 획득할 수도 있다.
- [0042] 일 실시예에 있어, 언어 모델 처리부(120)는, 언어 모델링(Language Modeling)을 위한 언어 모델을 이용하여 구현될 수도 있다. 여기서, 언어 모델링은, 사전 단어(previous word)를 이용하여 미지의 다음 단어(next word)를 예측하는 것을 포함할 수 있다. 이 경우, 언어 모델 처리부(120)는 사전 학습 언어 모델(PLM: Pre-trained Language Model)을 이용할 수도 있으며, 예를 들어, 버트(BERT), 로버타(RoBERTa: Robustly Optimized BERT Approach), 스팸버트(SpanBERT), 알버트(ALBERT) 또는 코버트(KoBERT) 등을 이용할 수도 있다. 로버타는, 버트 모델과는 상이하게 마스크(Mask)의 위치를 동적으로 변경하면서 학습을 수행하도록 마련된 것으로(다이나믹 마스크, Dynamic Masking), 통상 버트 모델보다 매우 많은 데이터와 자원을 이용하여 학습을 수행하도록 마련된 것이다. 이외에도 언어 모델 처리부(120)는, 실시예에 따라서, 적어도 일 종류의 다른 학습 모델을 이용하여 학습을 수행함으로써, 제1 모델(131), 제2 모델(132) 및 제3 모델(133)을 획득할 수도 있다. 예를 들어, 언어 모델 처리부(120)는, 다층 퍼셉트론(Multi-layer Perceptron), 심층 신경망(DNN: Deep Neural Network), 컨볼루션 신경망(CNN: Convolutional Neural Network) 또는 장단기 메모리(LSTM: Long short term memory) 등을 이용하여 학습을 수행하여, 제1 모델(131), 제2 모델(132) 및 제3 모델(133)을 획득할 수도 있다.
- [0043] 필요에 따라서, 언어 모델 처리부(120)는 제1 토큰화부(111)의 토큰화 결과를 이용하여 획득된 모델, 제2 토큰화부(112)의 토큰화 결과를 이용하여 획득된 모델 및 제3 토큰화부(113)의 토큰화 결과를 이용하여 획득된 모델 중 적어도 하나에 대해 적어도 하나의 말뭉치를 이용하여 미세 조정(파인 튜닝, fine-tuning)을 더 수행함으로써, 제1 모델 내지 제3 모델(131 내지 133) 중 적어도 하나를 더 획득하는 것도 가능하다.
- [0044] 평가부(140)는 제1 내지 제3 모델(131 내지 133)에 대한 평가를 수행하여, 제1 내지 제3 모델(131 내지 133) 중

에서 해당 대상 텍스트에 대한 최적의 모델을 결정할 수 있다. 일 실시예에 의하면, 평가부(140)는 확률 모델(probability)을 기반으로 한 복잡도(perplexity)를 이용하여 최적의 모델을 결정할 수도 있다. 이 경우, 평가부(140)는, 하기의 수학적 식 1을 이용하여 복잡도를 연산할 수 있다.

[수학적 식 1]

$$PPL(W) = P(\omega_1, \omega_2, \dots, \omega_N)^{-N} = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(\omega_i|\omega_{i-1})}}$$

여기서, PPL(W)는 복잡도이고, W는 전체 문장(일례로, 대상 텍스트의 전부 또는 일부일 수도 있고, 대상 텍스트를 포함하는 문장일 수도 있음)이며, (ω_1 내지 ω_N)은 단어 시퀀스이고, ω_1 내지 ω_N 각각은 단어를 의미한다. N은 단어 시퀀스의 길이(즉, 단어의 개수)를 의미하고, P()는 확률을 의미한다. 수학적 식 1에서 단어 시퀀스에 대한 확률 $P(\omega_1, \omega_2, \dots, \omega_N)$ 에 대해 정규화된 데이터에 대한 확률 P()에 대해 역수를 취한 후, 체인 룰(chain rule, 연쇄 법칙)을 적용하면, 복잡도 PPL(W)은 수학적 식 1의 맨 우변과 같이 주어지게 된다. 즉, 복잡도 PPL(W)은 제(i-1) 단어에 관한 제i 단어의 조건부 확률의 역수들의 곱에 대한 N 제곱근을 연산하는 것으로 주어질 수 있다. 수학적 식 1과 같은 복잡도 PPL(W)는 그 연산 결과가 작으면 작을수록 해당 모델의 성능이 상대적으로 우수함을 의미한다.

평가부(140)는 각각의 모델(131 내지 133)에 대해서 복잡도를 수학적 식 1을 이용하여 각각 연산하고, 각각의 연산 결과를 비교한 후, 복잡도의 연산 결과가 가장 작은 모델(131 내지 133) 중 어느 하나를 결정함으로써, 어느 토큰화부(111 내지 113)에 대응하는 모델(131 내지 133)이 해당 대상 텍스트를 처리하기 위한 최적 모델인지를 결정할 수 있다. 예를 들어, 제1 모델(131)의 복잡도가 가장 낮은 경우, 평가부(140)는 제1 토큰화부(111)에 의해 토큰화된 결과를 기반으로 생성된 제1 모델(131)을 최적 모델로 판단할 수 있다.

개체명 인식부(150)는 평가부(140)에 의해 최적의 모델로 판단된 모델, 즉 제1 내지 제3 모델(131 내지 133) 중 적어도 하나를 기반으로 대상 텍스트 및 대상 텍스트 이외의 다른 텍스트 중 적어도 하나의 텍스트로부터 하나 이상의 개체명을 인식하여 추출할 수 있다. 실시예에 따라서, 개체명 인식부(150)는 대상 텍스트 등이 입력된 최적 모델(131 내지 133)에 대해 소정의 말뭉치를 이용하여 미세 조정을 더 수행하고, 미세 조정된 모델을 이용하여 개체명 인식 및 추출을 수행하는 것도 있다. 텍스트에 따라서 개체명 인식부(150)는 텍스트로부터 어떠한 개체명도 추출하지 못할 수도 있다. 추출된 개체명은 저장부(15) 및 출력부(19) 중 적어도 하나로 전달될 수 있다.

상술한 개체명 인식 장치(10)는, 상술한 대상 텍스트에 대한 적어도 하나의 토큰화, 적어도 하나의 모델(131 내지 133)의 생성, 모델(131 내지 133)에 대한 평가 수행 및/또는 대상 텍스트에 대한 개체명 인식 등의 수행을 위해 특별히 고안된 장치를 이용하여 구현될 수도 있고, 또는 상술한 동작을 수행 가능한 하나 이상의 정보처리 장치를 단독으로 이용하거나 조합 이용하여 구현될 수도 있다. 여기서, 하나 이상의 정보처리장치는, 예를 들어, 데스크톱 컴퓨터, 랩톱 컴퓨터, 서버용 하드웨어 장치, 스마트 폰, 태블릿 피씨, 스마트 시계, 스마트 태그, 스마트 밴드, 두부 장착형 디스플레이(HMD: Head Mounted Display) 장치, 휴대용 게임기, 개인용 디지털 보조기(PDA: Personal Digital Assistant), 내비게이션 장치, 프린터 장치, 삼차원(3D) 프린터 장치, 스캐너 장치, 리모트 컨트롤러(리모컨), 디지털 텔레비전, 셋 톱 박스, 디지털 미디어 플레이어 장치, 미디어 스트리밍 장치, 디브이디 재생 장치, 콤팩트 디스크 재생 장치, 음향 재생 장치(인공 지능 스피커 등), 가전 기기(일례로 냉장고 등), 유인 또는 무인 이동체(일례로 차량(승용차, 버스나 이륜차 등), 이동성 로봇, 무선 모형 차량 또는 로봇 청소기 등), 유인 또는 무인 비행체(일례로 항공기나, 헬리콥터나, 드론, 모형 비행기, 모형 헬리콥터 등), 의료 기기, 가정용, 산업용 또는 군사용 로봇, 산업용 또는 군사용 기계, 또는 전자 광고판 등을 포함할 수 있으나 이에 한정되는 것은 아니다. 설계자나 사용자 등은 상황이나 조건에 따라서 상술한 정보처리장치 이외에도 정보의 연산 처리 및 제어가 다양한 장치 중 적어도 하나를 상술한 개체명 인식 장치(10)로 고려할 수 있다.

이하 도 5를 참조하여 개체명 인식 방법의 일 실시예를 설명한다.

도 5는 개체명 인식 방법의 일 실시예에 대한 흐름도이다.

도 5에 도시된 바에 의하면, 먼저 대상 텍스트가 획득될 수 있다(20). 대상 텍스트는 적어도 하나의 언어나 문자 등을 포함할 수 있다. 대상 텍스트는 사용자 등에 의해 직접 입력된 것일 수도 있고 또는 외부의 다른 장치

(일례로 메모리 장치나 휴대용 단말 장치 등)로부터 전달된 것일 수도 있다.

- [0054] 이어서 대상 텍스트에 대해서 적어도 하나 이상의 방법으로 토큰화가 수행될 수 있다(22). 예를 들어, 대상 텍스트는 세 개의 토큰화 기법을 이용하여 토큰화가 수행될 수 있으며, 이에 따라 이용된 토큰화 기법의 개수에 대응하는 개수의 토큰화 결과, 일례로 세 개의 토큰화 결과가 획득될 수 있다. 여기서, 토큰화 기법은 서브 워드를 이용하는 토큰화 기법을 포함할 수 있으며, 예를 들어, 바이트 페어 인코딩 기법, 워드피스 모델 및 센텐스피스 모델 중 적어도 하나를 포함할 수 있다. 즉, 바이트 페어 인코딩 기법에 의한 대상 텍스트의 토큰화, 워드피스 모델에 의한 대상 텍스트의 토큰화 및 센텐스피스 모델에 의한 대상 텍스트의 토큰화 중 적어도 하나가 수행될 수 있으며, 이에 따라 바이트 페어 인코딩 기법에 의해 획득된 제1 토큰화 결과, 워드피스 모델에 의해 획득된 제2 토큰화 결과 및 센텐스피스 모델에 의해 획득된 제3 토큰화 결과 중 적어도 하나가 획득될 수 있다. 실시예에 따라서, 이용되는 토큰화 기법의 개수는 3 미만일 수도 있고, 4 이상일 수도 있다.
- [0055] 서로 상이한 기법을 이용하여 획득된 토큰화 결과가 획득되면, 각각의 토큰화 결과를 이용하여 학습을 수행함으로써 각각의 토큰화 결과에 대응하는 적어도 하나의 언어 모델이 획득될 수 있다(24). 예를 들어, 제1 토큰화 결과 내지 제3 토큰화 결과가 획득되면, 각각의 제1 토큰화 결과 내지 제3 토큰화 결과에 대응하는 제1 모델 내지 제3 모델이 획득될 수 있다. 여기서, 각각의 제1 내지 제3 토큰화 결과를 입력으로 하는 학습의 수행을 위해 사전 학습 언어 모델이 이용될 수 있다. 사전 학습 언어 모델은, 예를 들어, 버트, 로버타, 스판버트, 알버트 또는 코버트 등을 포함할 수 있다. 실시예에 따라서, 제1 모델 내지 제3 모델 중 적어도 하나에 대해서 미세 조정이 더 수행될 수도 있다.
- [0056] 적어도 하나의 토큰화 결과 각각에 대한 적어도 하나의 언어 모델이 획득되면, 적어도 하나의 언어 모델에 대한 평가가 수행될 수 있으며, 평가 결과에 따라서 적어도 하나의 언어 모델 중에서 최적의 모델이 선택될 수 있다(26). 평가의 수행 및 최적 모델의 선택은, 확률을 기반으로 연산되는 복잡도를 이용하여 수행될 수도 있다. 구체적으로 예를 들어, 평가의 수행 및 최적 모델의 선택은, 제1 모델 내지 제3 모델 각각에 대한 복잡도를 연산하고, 연산 결과 복잡도가 가장 낮은 적어도 하나의 모델, 일례로 제1 모델을 최적의 언어 모델로 선택함으로써 수행될 수도 있다. 여기서, 복잡도는 상술한 수학적 식 1과 같이 제($i-1$) 단어에 대한 제 i 단어의 조건부 확률의 역수들의 곱에 대한 N 제곱근으로 주어질 수도 있다.
- [0057] 일 실시예에 따르면, 최적 모델이 결정되면, 최적 모델을 이용하여 개체명 인식이 수행될 수 있다(28). 개체명 인식은 상술한 최적 모델의 결정을 수행한 장치에 의해 수행될 수도 있고, 또는 다른 장치에 의해 수행될 수도 있다. 후자의 경우, 최적 모델은 유무선 통신 네트워크나 휴대용 메모리 장치 등을 이용하여 최적 모델 결정을 수행한 장치에서 다른 장치로 전달될 수도 있다. 필요에 따라 개체명의 인식 과정 이전에 최적 모델에 대한 미세 조정이 더 수행되는 것도 가능하다.
- [0058] 상술한 실시예에 따른 개체명 인식 방법은, 컴퓨터 장치에 의해 구동될 수 있는 프로그램의 형태로 구현될 수 있다. 프로그램은, 명령어, 라이브러리, 데이터 파일 및/또는 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며, 기계어 코드나 고급 언어 코드를 이용하여 설계 및 제작된 것일 수 있다. 프로그램은 상술한 개체명 인식 방법을 구현하기 위하여 특별히 설계된 것일 수도 있고, 컴퓨터 소프트웨어 분야에서 통상의 기술자에게 공지되어 사용 가능한 각종 함수나 정의를 이용하여 구현된 것일 수도 있다. 또한, 여기서, 컴퓨터 장치는, 프로그램의 기능을 실현 가능하게 하는 프로세서나 메모리 등을 포함하여 구현된 것일 수 있으며, 필요에 따라 통신 장치를 더 포함할 수도 있다. 상술한 개체명 인식 방법을 구현하기 위한 프로그램은, 컴퓨터에 의해 판독 가능한 기록 매체에 기록될 수 있다. 컴퓨터에 의해 판독 가능한 기록 매체는, 예를 들어, 솔리드 스테이트 드라이브(SSD), 롬, 램 또는 플래시 메모리 등과 같은 반도체 저장 매체나, 하드 디스크나 플로피 디스크 등과 같은 자기 디스크 저장 매체나, 콤팩트 디스크나 디브이디 등과 같은 광 기록 매체나, 또는 플롭티컬 디스크 등과 같은 자기-광 기록 매체 등 컴퓨터 등의 호출에 따라 실행되는 하나 이상의 프로그램을 일시적 또는 비일시적으로 저장 가능한 적어도 한 종류의 물리적 저장 매체를 포함할 수 있다.
- [0059] 이상 개체명 인식 장치 및 개체명 인식 방법의 일 실시예에 대해 설명하였으나, 장치 또는 방법은 오직 상술한 실시예에 한정되는 것은 아니다. 해당 기술 분야에서 통상의 지식을 가진 자가 상술한 실시예를 기초로 수정 및 변형하여 구현할 수 있는 다른 다양한 장치나 방법 역시 상술한 개체명 인식 장치 또는 개체명 인식 방법의 일 실시예가 될 수 있다. 예를 들어, 설명된 방법(들)이 설명된 바와 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소(들)가 설명된 바와 다른 형태로 결합, 연결 또는 조합되거나 다른 구성 요소 또는 균등물 등에 의하여 대체 또는 치환되더라도, 상술한 개체명 인식 장치 및/또는 개체명 인식 방법의

일 실시예가 될 수 있다.

부호의 설명

[0060]

10: 개체명 인식 장치 100: 프로세서

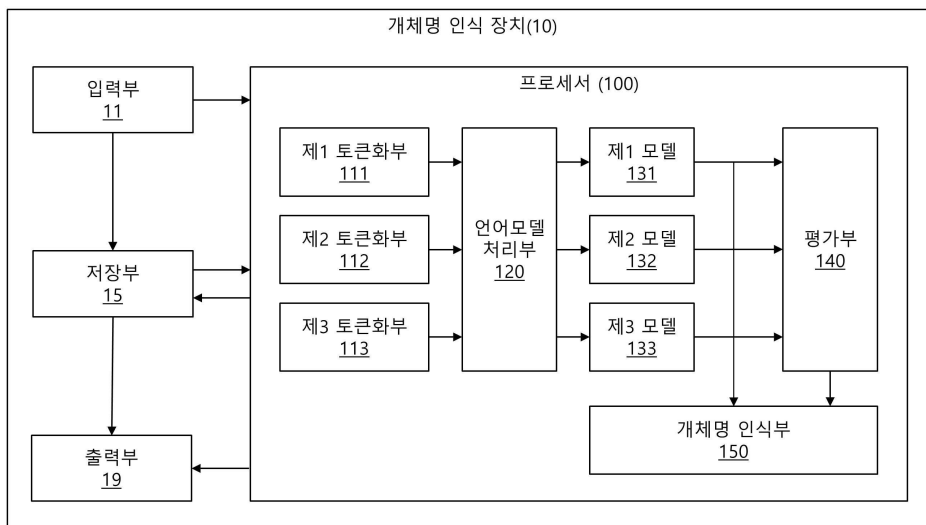
111: 제1 토큰화부 112: 제2 토큰화부

113: 제3 토큰화부 120: 언어모델처리부

140: 평가부 150: 개체명 인식부

도면

도면1

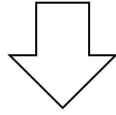


도면2

	문장	분절 결과		
31 ...	영문장(Admiral Yi born in Chosun ...)	⌞ Admiral/Yi<PS>/born/in/Chosun<LC>		... 31a
32 ...	한국어문장(조선에서 태어난 이순신은)	⌞ 조선에서/태어난/이순신은	(X)	... 32a
		⌞ 조선<LC>/에서/태어/난/이순신<PS>/은	(O)	... 32b

도면3

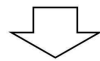
41 ...	eleven	twelve	thirteen	twenty
42 ...	2	3	4	3



43 ...	e, l, v, n, t, w, h, l, r, o, u
44 ...	en
45 ...	tw
46 ...	el
	:

도면4

51 ... 세종은 조선의 제 4대 국왕이다



	알고리즘	분절 결과
52 ...	BPE	세/종은/조선의/제/4대/국왕/이다/
53 ...	Wordpiece	세종/##은/조선의/제/4대/국왕이/##다/
54 ...	Sentencepiece	_세종/은/_조선의/_제/_4/대/_국왕/이다/

도면5

