(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl.[7]
G06F 17/20

(11) 공개번호 10-2005-0027931
(43) 공개일자 2005년03월21일

(21) 출원번호 10-2004-0073392
(22) 출원일자 2004년09월14일

(30) 우선권주장 10/662,602 2003년09월15일 미국(US)

(71) 출원인 마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이

(72) 발명자 후앙,창-닝
중국 100081 베이징 종구안춘 난다지에 넘버. 40 빌딩 에이-1 슈트 307

가오,지앙펭
중국 100080 베이징 하이딘안 디스트릭트 지춘 로드 로주앙 웨스트 빌딩 1 룸 302

리,무
중국 베이징 샤오앙 디스트릭트 케스에유아나닐 스트리트 빌딩 602 룸207

창,아셀리엑스.
미국 98029 워싱턴주 이사쿠아 253번 시티. 사우스이스트 3527

(74) 대리인 주성민
백만기
이중희

심사청구 : 없음

(54) 중국어 단어 분절

요약

본 발명은 언어 모델을 트레이닝하는데 사용되는 언어자료(corpus)에 관한 것이다. 상기 언어자료는 복수의 문자열 및 상기 복수의 문자열과 관련된 복수의 형태론적 태그(morphological tag)를 포함한다. 상기 복수의 형태론적 태그는 관련 문자열의 형태론적 타입(morphological type) 및 형태론적 서브타입(morphological subtype)을 형성하는 성분 조합(combination of parts)을 나타낸다.

대표도

도 2

색인어

언어 처리 시스템, 언어 모델, 단어 분절, 형태론적 태그, 언어자료

명세서

도면의 간단한 설명

도 1은 본 발명이 이용될 수 있는 일반적인 컴퓨팅 환경의 블록도.

도 2는 언어 처리 시스템의 블록도.

도 3은 주석달린 언어자료(annotated corpus)를 개발하기 위한 방법의 흐름도.

도 4는 언어 모델의 성능을 생성하고 평가하는 흐름도.

도 5는 형태론적으로 유도된 단어의 타입과 서브타입의 블록도.

<도면의 주요 부분에 대한 부호의 설명>

200: 언어 처리 시스템

202: 입력

204: 출력

206: 언어 모델

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 자연어 처리 분야에 관한 것이다. 보다 구체적으로는, 본 발명의 단어 분절(word segmentation)에 관한 것이다.

단어 분절은 텍스트와 같은 언어의 표현을 구성하는 개별 단어를 식별하는 과정을 의미한다. 단어 분절은 철자와 문법을 검사하고, 텍스트에서 음성을 합성하며, 자연어 파싱과 습득을 수행하는데 유용하며, 이들 모두는 개별 단어의 식별로부터 이득을 얻는다.

띄어쓰기와 구두점이 통상 텍스트 내의 개별 단어의 경계를 정하기 때문에, 영문 텍스트의 단어 분절을 수행하는 것은 다소 수월하다. 아래 표 1의 영어 문장을 고려하자.

표 1

The motion was then tabled--that is, removed

indefinitely from consideration.

각각의 인접한 띄어쓰기 및/또는 구두점의 시퀀스를 시퀀스 이전 단어의 끝으로 식별함으로써, 표 1에서의 영어 문장은 아래 표 2에 나타낸 바와 같이 수월하게 분절된다.

표 2

The motion was then tabled--that is, removed

indefinitely from consideration.

중국어 텍스트에서, 단어 경계는 명시적이기보다는 묵시적이다. "위원회는 어제 정오 부에노스아이레스에서 이 문제를 논의했다"를 의미하는 아래 표 3의 문장을 고려하자.

표 3

昨天下午委员会在布宜诺斯艾利斯讨论了这个问题。

이 문장에서 구두점과 띄어쓰기가 없음에도 불구하고, 중국어 독자는 표 3의 문장을 아래 표 4에 개별적으로 밑줄친 단어들로 이루어짐을 인식할 수 있다.

표 4

昨 天 下 午 委 员 会 在 布 宜 诺 斯 艾 利 斯 讨 论 了 这 个 问 题。

다수의 방법 및 시스템이 중국어와 일본어 등의 언어에서 단어 분절을 제공하도록 안출되어 왔다. 몇몇 시스템에서, 분절된 텍스트의 언어자료(corpus)에 기초하여 모델이 트레이닝된다. 이 모델은 텍스트 문자열에서 발생하는 다양한 분절의 가능성을 나타내며, 이를 가리키는 출력을 제공한다. 언어자료를 구현하여 모델을 트레이닝하는 것은 시간과 비용을 요구한다. 많은 경우, 관련 단어 분절 시스템의 출력 품질은 상기 모델을 트레이닝하는데 사용되는 언어자료의 품질에 상당히 의존하게 된다. 그 결과, 언어자료를 평가하여 구현하는 방법은 우수한 단어 분절을 제공하는 것을 지원할 수 있다.

발명이 이루고자 하는 기술적 과제

본 발명은 언어 모델을 트레이닝하는데 사용하기 위한 언어자료에 관한 것이다. 이 언어자료는 복수의 문자열과 복수의 문자열에 관련되는 복수의 형태론적 태그를 포함한다. 상기 복수의 형태론적 태그는 관련 문자열의 형태론적 타입 및 형태론적 서브타입을 형성하는 성분의 조합을 나타낸다.

다른 양태에서, 단어 분절을 수행하기 위한 명령들을 갖는 컴퓨터 판독가능 매체가 제공된다. 상기 명령들은 비분절 텍스트의 입력을 수신하는 단계와 언어 모델에 액세스하여 상기 텍스트의 분절을 결정하는 단계를 포함한다. 형태론적으로 유도된 단어는 상기 텍스트와 분절된 텍스트를 나타내는 출력에서 검출되며 형태론적으로 유도되는 단어를 형성하는 성분의 조합의 표시가 제공된다.

발명의 구성 및 작용

본 발명을 보다 상세히 설명하기 전에, 본 발명이 사용될 수 있는 예시적인 환경의 실시예를 설명한다. 도 1은 본 발명이 구현될 수 있는 적절한 컴퓨팅 시스템 환경(100)의 일 예를 나타낸다. 컴퓨팅 시스템 환경(100)은 적절한 컴퓨팅 환경의 단지 일 예이며, 본 발명의 사용범위 또는 기능에 대한 어떤 제한을 암시하려는 것이 아니다. 또한, 컴퓨팅 환경(100)은 예시적인 운영 환경(100)에서 설명되는 컴포넌트 중의 임의의 것 또는 그 조합에 대한 어떤 의존성이나 요건을 갖는 것을 해석되어서는 안 된다.

본 발명은 수많은 다른 범용 또는 특정 목적 컴퓨팅 시스템 환경 또는 구성에서 동작한다. 본 발명에서 사용될 수 있는 공지의 컴퓨팅 시스템, 환경 및/또는 구성의 예는, 개인용 컴퓨터, 서버 컴퓨터, 핸드헬드 또는 랩탑 장치, 멀티프로세서 시스템, 마이크로프로세서 기반 시스템, 셋탑 박스, 프로그래머블 소비자 전자제품, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터상기 시스템 또는 장치 중 임의의 것을 포함하는 분산 컴퓨팅 환경 등을 포함하지만 이에 한정되는 것은 아니다.

본 발명은 컴퓨터에 의해 실행되는 프로그램 모듈 등의 컴퓨터 실행가능 명령들의 일반적인 경우에 대하여 설명한다. 통상, 프로그램 모듈은 특정 작업을 수행하거나 특정 추상 데이터형을 구현하는 루틴, 프로그램, 객체, 컴포넌트, 데이터 구조 등을 포함한다. 당업자는 후술하는 임의의 컴퓨터 판독가능 매체의 형태로서 구현될 수 있는 컴퓨터 실행가능 명령들로서 여기서의 명세서 및/또는 도면을 구현할 수 있다.

본 발명은 또한 통신 네트워크를 통해 연결되는 원격 프로세싱 장치에 의해 작업이 수행되는 분산 컴퓨팅 환경에서 실시될 수 있다. 분산 컴퓨팅 환경에서, 프로그램 모듈은 메모리 스토리지 장치를 포함하는 로컬 및 원격 컴퓨터 스토리지 매체에 배치될 수 있다.

도 1을 참조하면, 본 발명을 구현하는 예시적인 시스템은 컴퓨터(110)의 형태의 범용 컴퓨팅 장치를 포함한다. 컴퓨터(110)의 컴포넌트는 처리부(120), 시스템 메모리(130), 및 상기 처리부(120)에 시스템 메모리 등의 여러 시스템 컴포넌트를 결합시키는 시스템 버스(121)를 포함한다. 시스템 버스(121)는 메모리 버스 또는 메모리 컨트롤러, 주변 버스, 및 다양한 버스 아키텍처 중 임의의 것을 사용하는 로컬 버스를 포함하는 여러 유형의 버스 아키텍처 중 임의의 것일 수 있다. 예를 들면, - 한정이 아님 -, 이러한 아키텍처는 산업 표준 아키텍처(ISA) 버스, 마이크로 채널 아키텍처(MCA) 버스, 개선된 ISA(EISA) 버스, 비디오 전자 표준 협회(VESA) 로컬 버스, 및 메자닌(Mezzanine) 버스로도 알려진 주변 컴포넌트 상호접속(PCI) 버스를 포함한다.

컴퓨터(110)는 통상 다양한 컴퓨터 판독가능 매체를 포함한다. 컴퓨터 판독가능 매체는 컴퓨터(110)에 의해 액세스될 수 있는 임의의 입수가능 매체일 수 있으며, 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 포함한다. 예를 들면, - 한정이 아님 -, 컴퓨터 판독가능 매체는 컴퓨터 스토리지 매체 및 통신 매체를 포함할 수 있다. 컴퓨터 스토리지 매체는 컴퓨터 판독가능 명령, 데이터 구조, 프로그램 모듈 또는 기타 데이터 등의 정보의 저장을 위한 임의의 방법 또는 기술로 구현되는 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 포함한다. 컴퓨터 스토리지 매체는, RAM, ROM, EEPROM, 플래시 메모리 또는 기타 메모리 기술, CD-ROM, 디지털 다기능 디스크(DVD), 또는 기타 광 디스크 스토리지, 자기 카세트, 자기 테이프, 자기 디스크 스토리지 또는 기타 자기 스토리지 장치 또는 원하는 정보를 저장하는데 사용될 수 있고 컴퓨터(110)에 의해 액세스될 수 있는 임의의 다른 매체를 포함하지만 이에 한정되는 것은 아니다. 통신 매체는 통상 반송파 또는 다른 전송 메커니즘 등의 변조된 데이터 신호로, 컴퓨터 판독가능 명령, 데이터 구조, 프로그램 모듈 또는 기타 데이터를 구현하며, 임의의 정보 전달 매체를 포함한다. "변조된 데이터 신호"라는 용어는 신호 내의 정보를 인코딩하는 방식으로 하나 이상의 특성이 설정 또는 변경되는 신호를 의미한다. 예를 들면, - 한정이 아님 -, 통신 매체는 유선 네트워크 또는 다이렉트 와이어드 접속 등의 유선 매체와, 어쿠스틱, RF, 적외선 및 기타 무선 매체 등의 무선 매체를 포함한다. 상기의 것의 임의의 조합은 컴퓨터 판독가능 매체의 범위 내에 또한 포함되어야 한다.

시스템 메모리(130)는 판독 전용 메모리(ROM; 131) 및 랜덤 액세스 메모리(RAM; 132) 등의 휘발성 및/또는 비휘발성 메모리의 형태의 컴퓨터 스토리지 매체를 포함한다. 기본 입출력 시스템(133; BIOS)은 시동 시와 같은 컴퓨터(110) 내의 성분들 간의 정보를 전송하는 것을 지원하는 기본 루틴을 포함하며, 통상 ROM(131)에 저장된다. RAM(132)은 즉시 액세스가능하고 및/또는 처리부(120) 상에 현재 처리되는 데이터 및/또는 프로그램 모듈을 통상 포함한다. 예를 들면, - 한정이 아님 -, 도 1은 운영 체계(134), 애플리케이션 프로그램(135), 기타 프로그램 모듈(136) 및 프로그램 데이터(1378)를 나타낸다.

컴퓨터(110)는 또한 다른 분리형/비분리형, 휘발성/비휘발성 컴퓨터 스토리지 매체를 포함할 수 있다. 단지 예를 들면, 도 1은 비분리형, 비휘발성 자기 매체로부터 판독하거나 이에 기입하는 하드 디스크 드라이브(141), 분리형, 비휘발성 자기 디스크(152)로부터 판독하거나 이에 기입하는 자기 디스크 드라이브(151), 및 CD-ROM이나 다른 광 매체 등과 같은 분리형, 비휘발성 광디스크(156)로부터 판독하거나 이에 기입하는 광 디스크 드라이브(155)를 나타낸다. 예시적인 운영 환경에서 사용될 수 있는 다른 분리형/비분리형, 휘발성/비휘발성 컴퓨터 스토리지 매체는 자기 테이프 카세트, 플래시 메모리 카드, 디지털 다기능 디스크, 디지털 비디오 테이프, 고체 상태 RAM, 고체 상태 ROM 등을 포함하지만 이에 한정되는 것은 아니다. 하드 디스크 드라이브(141)는 인터페이스(140) 등의 비분리형 메모리 인터페이스를 통해 시스템 버스(121)에 통상 접속되며, 자기 디스크 드라이브(151)와 광 디스크 드라이브(155)는 인터페이스(150)와 같은 분리형 메모리 인터페이스에 의해 시스템 버스(121)에 통상 접속된다.

상술되고 도 1에서 나타낸 드라이브 및 관련 컴퓨터 스토리지 매체는 컴퓨터(110)에 컴퓨터 판독가능 명령, 데이터 구조, 프로그램 모듈 및 기타 데이터를 제공한다. 도 1에서, 예를 들면, 하드 디스크 드라이브(141)는 운영 체계(144), 애플리케이션 프로그램(145), 기타 프로그램 모듈(146) 및 프로그램 데이터(147)를 저장하는 것으로 예시되어 있다. 이들 컴포넌트는 운영 체계(134), 애플리케이션 프로그램(135), 기타 프로그램 모듈(136) 및 프로그램 데이터(137)와 동일하거나 상이할 수 있다. 운영 체계(144), 애플리케이션 프로그램(145), 기타 프로그램 모듈(146) 및 프로그램 데이터(147)는 그들이 적어도 서로 다른 복사본임을 나타내기 위하여 여기서 서로 다른 번호가 부여된다.

사용자는 키보드(162), 마이크로폰(163), 및 마우스, 트랙볼 또는 터치 패드 등의 포인팅 장치(161)와 같은 입력 장치를 통해 컴퓨터(110)에 명령과 정보를 입력할 수 있다. 다른 입력 장치(미도시)는 조이스틱, 게임 패드, 위성 접시, 스캐너 등을 포함할 수 있다. 이들 및 다른 입력 장치는 시스템 버스에 결합되지만 병렬 포트, 게임 포트 또는 범용 직렬 버스(USB) 등의 다른 인터페이스와 버스 구조에 의해 접속될 수 있는 사용자 입력 인터페이스(160)를 통해 처리부(120)에 종종 접속된다. 모니터(191) 또는 다른 유형의 디스플레이 장치가 또한 비디오 인터페이스(190) 등의 인터페이스를 통해 시스템 버스(121)에 접속된다. 모니터에 더하여, 컴퓨터도 또한 스피커(197)와 프린터(196) 등의 다른 주변 출력 장치를 포함할 수 있으며, 이는 출력 주변 인터페이스(195)를 통해 접속될 수도 있다.

컴퓨터(110)는 원격 컴퓨터(180) 등의 하나 이상의 원격 컴퓨터에 대한 논리적 접속을 사용하여 네트워크화된 환경에서 동작할 수 있다. 원격 컴퓨터(180)는 개인용 컴퓨터, 핸드헬드 장치, 서버, 라우터, 네트워크 PC, 피어 장치 또는 기타 공통 네트워크 노드일 수 있으며, 통상, 컴퓨터(110)에 대하여 상술한 모든 또는 다수의 구성요소를 포함한다. 도 1에 도시된 논리적 접속은 근거리 네트워크(LAN; 171) 및 광역 네트워크(WAN; 173)를 포함하지만, 다른 네트워크를 포함할 수도 있다. 이러한 네트워킹 환경은 사무실, 범사내망 및 인터넷에서 흔히 볼 수 있다.

LAN 네트워킹 환경에서 사용되는 경우, 컴퓨터(110)는 네트워크 인터페이스 또는 어댑터(170)를 통해 LAN(171)에 접속된다. WAN 네트워킹 환경에서 사용되는 경우, 컴퓨터(110)는 인터넷 등의 WAN(173)을 통해 통신을 설정하는 모뎀(172) 또는 다른 수단을 통상 포함한다. 모뎀(172)은 내장형 또는 외장형일 수 있으며, 사용자 입력 인터페이스(160) 또는 다른 적절한 메커니즘을 통해 시스템 버스(121)에 접속될 수 있다. 네트워크화된 환경에서, 컴퓨터(110) 또는 그 일부에 대하여 도시된 프로그램 모듈은 원격 메모리 스토리지 장치에 저장될 수 있다. 예를 들면, -한정이 아님 -, 도 1은 원격 컴퓨터(180) 상에 상주하는 것으로서 원격 애플리케이션 프로그램(185)을 나타낸다. 도시된 네트워크 접속은 예시적이며, 컴퓨터들 사이의 통신을 설정하는 다른 수단이 사용될 수 있음이 이해될 것이다.

도 2는 언어 입력(202)을 수신하여 언어 출력(204)을 제공하는 언어 처리 시스템(200)을 일반적으로 나타낸다. 예를 들면, 언어 처리 시스템(200)은 언어 입력(202)으로서 비분절 텍스트를 수신하는 단어 분절 시스템 또는 모듈로서 구체화될 수 있다. 언어 처리 시스템(200)은 비분절 텍스트를 처리하고, 분절된 텍스트를 나타내는 출력(204) 및 분절된 텍스트에 대한 관련 정보를 제공한다.

처리 동안, 언어 처리 시스템(200)은 입력 텍스트(202)에 대한 분절을 결정하기 위해 언어 모델(206)을 액세스할 수 있다. 언어 모델(206)은 특정 유형의 표시 뿐만 아니라 다양한 유형의 단어를 정의하는 주석이 달린 언어자료로부터 구성될 수 있다. 당업자가 이해할 수 있는 바와 같이, 언어 처리 시스템(200)은, 예를 들어, 철자 검사, 문법 검사, 텍스트로부터 음성 합성, 음성 인식, 정보 검색 및 자연어 파싱 및 습득 수행 등의 다양한 상황에서 사용될 수 있다. 또한, 언어 모델(206)은 언어 처리 시스템(200)이 사용되는 특정 애플리케이션에 기초하여 개발될 수 있다.

분절을 제공하는 것에 더하여, 시스템(200)은 또한 각 분절 단어마다 단어형 표시를 제공한다. 일 실시예에서, 중국어 단어는 다음 4 가지 유형: (1) 주어진 어휘목록 내의 엔트리(이하, 어휘목록 단어 또는 LW로 지칭함), (2) 형태론적으로 유도된 단어(MDWs), (3) 날짜, 시간, 페센트, 돈 등의 팩토이드 및 (4) 인명(PNs), 지명(LNs) 및 기관명(ONs) 등의 개체명(NEs) 중 하나로서 정의된다. 다양한 서브타입이 또한 정의될 수 있다. 이들 단어형의 정의가 주어지면, 시스템(200)은 분절과 단어형을 나타내는 출력을 제공할 수 있다. 예를 들면, "친구가 12시 반에 점심식사를 위해 이준성 교수님 댁에 기쁘게 간다"라는 의미를 갖는, 아래 표 5의 비분절 문장을 고려하자.

표 5

朋友们十二点三十分高高兴兴到李俊生教授家吃饭

시스템(200)의 예시적인 출력이 아래 표 6에 도시되어 있다. 사각 꺽쇠괄호(square bracket)는 단어 경계(word boundary)를 나타내고, "+ "는 형태소(morpheme) 경계를 나타낸다. 태그는 브래킷 내에서 문장 내의 단어의 다양한 타입과 서브타입을 나타내도록 제공된다.

표 6

[朋友+们 MA_S] [十二点三十分 12:30 TIME] [高兴 MR_AABB]
[到] [李俊生 PN] [教授] [家] [吃饭]

분절을 제공하기 위해서, 언어 모델(206)은 입력 텍스트(202)에서 단어형을 검출한다. 어휘목록 단어의 경우, 단어가 어휘목록에 포함되면 단어 경계가 검출된다. 형태론적으로 유도된 단어의 경우, 형태론적 패턴이 검출되며, 예를 들어, 朋友+们 (친구+ 들 을 의미)는 명사 们 에 대한 복수 접미사 朋友 의 접사 첨가(affixation)에 의해 유도되고 (MA_S는 접미사 첨가(suffixation) 패턴을 가리키는 태그임), 高高兴兴 (기쁘게 라는 의미)는 高兴 (행복한)의 음절 중복(reduplication)이다 (MR_AABB는 AABB 음정 중복 패턴을 가리키는 태그임).

팩토이드의 경우, 그들의 타입과 정규화된 형태가 검출되며, 예를 들면, 12:30은 시간 표현 十二点三十分 의 정규화된 형태이다(TIME은 시간 표면을 가리키는 태그임). 개체명의 경우, 서브타입이 검출되며, 예를 들면, 李俊生 (Li Junsheng)은 인명이다(PN은 인명을 가리키는 태그임).

언어 모델(206)은 주석달린 언어자료(annotated corpus)로부터 생성될 수 있다. 도 3은 시스템(200)의 언어 모델(206)과 같이 단어 분절 시스템에서 언어 모델을 생성하는데 사용되는 주석달린 언어자료를 구현하는 방법(250)을 나타낸다. 단계 252에서, 단어 분절에 관련된 단어 및 규칙이 정의된다. 예를 들면, 중국어 단어 분절을 위한 어휘자료, 중국어 형태론적으로 유도된 단어에 대한 규칙 모음, 중국어 팩토이드의 가이드라인 및 개체명 및/또는 이들의 조합이 주석달린 언어자료를 구현하는데 정의될 수 있다. 단계 254에서, 다양한 텍스트뿐만 아니라 방대한 양의 텍스트를 포함하는 광범위한 언어자료가 제공된다. 광범위한 언어자료는 신문 및 잡지 등의 다양한 텍스트 소스로부터 선택될 수 있다. 다음으로, 단계 256에서, 단계 252에서 정의된 단어와 규칙에 일치하는 리스트가 광범위한 언어자료에서 추출되어 가능 단어(potential word) 리스트를 생성한다.

단계 258에서, 추출된 리스트는 원하는 경우 수동으로 점검되어 리스트 내의 임의의 잡음 또는 에러를 필터링할 수 있다. 그 후, 단계 260에서 리스트가 정의된 단어 및 규칙의 충분한 커버리지(coverage)를 갖는지를 결정한다. 일 실시예에서, 리스트는 폭넓은 영역과 스타일을 갖는 균형되고 독립적인 테스트 언어자료와 비교될 수 있다. 예를 들면, 영역과 스타일은, 예를 들어, 문화, 경제, 문학, 군사, 정치, 과학 및 기술, 사회, 스포츠, 컴퓨터 및 법률에 관련되는 텍스트를 포함할 수 있다. 다르게는, 특정 애플리케이션의 넓은 커버리지를 갖는 애플리케이션 특정 언어자료가 사용될 수 있다. 리스트가 충분한 커버리지를 갖는 것으로 판단되면, 언어자료는 단계 262에서 태깅된다. 언어자료의 태깅(tagging)은 후술하는 바와 같이 수행될 수 있다. 단계 264에서, 태깅된 언어자료가 검사될 수 있으며, 임의의 에러가 정정될 수 있다. 단계 266에서, 결과적인 언어자료는 트레이닝 또는 테스팅 언어자료로서 방대한 양의 텍스트를 태그하는데 시드(seed) 언어자료로서 사용된다. 그 결과, 도 4의 방법 280을 사용하여 평가될 수 있는 주석달린 언어자료가 구현된다.

도 4는 개선된 단어 분절을 제공하기 위해 언어 모델(206)을 생성하고 평가하는 방법(280)을 나타낸다. 단계 282에서, 그 프로세스가 도 3을 참조하여 상술된 주석달린 언어자료를 구현한다. 주석달린 언어자료가 주어지면, 단계 284에서 트레이닝 또는 테스팅 모델이 주석달린 언어자료에 기초하여 생성된다. 단계 286에서, 생성된 모델이 소정의 테스트 언어자료 또는 다른 모델과 비교하여 평가된다. 단계 286에서 수행된 평가가 주어지면, 언어 모델(206)의 효율성이 결정될 수 있다.

언어 모델을 평가하기 위해서, 이 모델을 사용한 단어 분절 시스템의 출력은 분절 시스템의 표준 출력의 역할을 수행하는 표준 주석 테스팅 언어자료와 비교될 수 있다. 신뢰성있는 평가를 얻기 위해서, 독립적이고 균형이 있으며, 적절한 크기의 기본 (주석이 없는) 테스트 언어자료가 선택될 수 있다. 독립적 테스트 언어자료는 언어 모델을 트레이닝하는데 사용되는 주석달린 언어자료와 상대적으로 적은 중첩을 가질 수 있다. 균형있는 언어자료는 폭넓은 영역, 스타일 및 시간을 갖는 문서를 포함한다. 충분히 많도록, 테스트 언어자료의 일 실시예는 대략 백만개의 중국어 문자를 포함한다. 테스트 언어자료를 구현한 후에, 이 언어자료는 수동으로 주석을 달아서, 주어진 테스트 언어자료에 대하여 중국어 단어 분절 시스템의 표준 출력으로서 사용된다. 테스트 언어자료는 후술하는 태깅 사양 또는 다른 태깅 사양을 사용하여 주석을 달 수 있다.

주석 테스트 언어자료가 주어지면, 정성 평가가 언어 모델의 성능을 평가하는데 사용될 수 있다. 표준 테스트 모음에서 단어 토큰의 총수가 "S", 테스트 모음에 적용되는 평가대상의 단어 분절 시스템의 출력의 단어 토큰의 총수는 "E", 표준 테스트 모음에서 단어 토크와 정확하게 일치하는 단어 토큰의 개수는 "M"이면, 언어 모델의 성능을 평가하도록 정성 수치는 계산된다. 아래 식 1 내지 3은 정확성, 리콜 및 F 점수에 대한 값을 나타낸다.

정확성= M/E (1)

리콜 = M/S (2)

F = 2 x 정확성 x 리콜/(정확성 + 리콜) (3)

더욱이, 평가는 상기 식 1 내지 3에 따라 여러 서브타입 상에 수행될 수 있다. 예를 들면, 인명(person name) 성능 평가가 수행될 수 있으며, S_{PN}은 표준 테스트 언어자료에서 인명 토큰의 총 수이다. E_{PN}은 평가될 단어 분절 시스템의 출력에서 인명 토큰의 총 수이고, M_{PN}은 표준 테스트 모음에서 인명과 정확하게 일치하는 출력에서의 인명 토큰의 개수이다. 그 결과, 성능 식은:

$정확성_{PN} = M_{PN}/E_{PN}$ (4)

$$리콜_{PN} = M_{PN}/S_{PN} \quad (5)$$

$$F_{PN} = 2 \times 정확성_{PN} \times 리콜_{PN}/(정확성_{PN} + 리콜_{PN}) \quad (6)$$

언어 모델의 성능을 평가할 때 다른 시스템 결과를 비교하는 것이 또한 유용하다. 예를 들면, (1) 인명, (2) 지명, (3) 기관명, (4) 중첩 불명료 문자열 및 (5) 커버링(covering) 불명료 문자열 등의 서로 다른 단어 분절 시스템의 출력의 여러 부위만을 비교하는 유용할 수 있다. 분절 시스템의 출력의 서브세트를 단지 평가함으로써, 에러가 분절에서 발생하는 경우에 보다 낳은 아이디어가 될 수 있다.

주석달린 언어자료를 구현하기 위해서, 태깅 사양은 상술한 중국어 단어형의 정의가 주어진 언어자료를 지속적으로 태깅하는데 사용된다. 어휘자료를 갖는 어휘자료 단어들은 추가 태깅없이 꺽쇠괄호에 의해 범위가 정해진다. 다른 유형은 후술하는 바와 같이 태깅된다.

도 5는 언어자료를 태깅하는 형태론적 카테고리에 대한 다이어그램을 나타낸다. 형태론적 카테고리는 접사 첨가(affixation), 음절 중복(reduplication), 분할(split), 병합(merge) 및 주요어 불변화사(head particle)를 포함한다. 각 형태론적 카테고리 또는 타입은 태깅 처리 동안 태깅될 수 있는 다양한 서브타입을 포함한다. 도 5에서의 포맷은 카테고리, 단어를 구성하는 성분 및 단어의 결과적인 품사를 나타낸다. 도 5의 다이어그램에서, "MP"는 형태론적 접두사, "MS"는 형태론적 접미사를 의미한다. "ML"은 분할, "MM"은 병합, 그리고 "MHP"는 형태론적 주요어 불변화사이다. 밑줄(_)과 (-) 사이의 성분은 형태론적으로 유도된 단어를 형성하는 성분의 조합이다. 음절 중복과 병합에서, 문자 A, B 및 C는 중국어 문자를 나타낸다.

도 5에서의 포맷은 형태론적 변화를 나타내고 다른 태깅 포맷이 변화를 나타내는데 사용될 수 있다. 접사 첨가는 접두사와 접미사의 서브카테고리를 포함하고, 여기서, 문자는 원래 문자로 표현되는 단어를 형태론적으로 변경하기 위해 다른 문자의 문자열에 추가된다. 접두사는 7개의 서브타입을 포함하고, 접미사는 13개의 서브타입을 포함한다. 음절 중복은 문자의 패턴을 구성하는 원래의 단어가 문자의 조합으로 구성된 다른 단어로 변경되는 경우에 발생하고 13개의 상이한 서브타입을 포함한다. 음절 중복은 동사를 나타내는 "V", 목적어를 나타내는 "0"를 포함하고, "1", "le" 및 "liaozhi"는 불변화사이다.

분할은 구문론 상으로 상이한 단어이지만, 의미론 상으로 단일 단어인 일련의 표현을 포함한다. 예를 들면, 문자열 ABC는 "이미 먹었다"라는 어구를 나타낼 수 있으며, 여기서, 2문자 단어 AC는 단어 "먹었다"를 나타내고 단어 "이미"를 나타내는 불변화사 문자 B로 분할된다. 분할은 두개의 서브타입을 포함한다. 하나의 서브타입은 동사와 목적어 사이에 문자 또는 문자들을 삽입하는 것을 포함하고,ㅡ 다른 서브타입은 어구 "qilai" 사이에 목적어를 삽입한다. 병합은 두개의 문자로 이루어진 하나의 단어와 두개의 문자로 이루어진 다른 단어가 조합되어 단일 단어를 형성하는 경우에 발생하고 3개의 서브타입을 포함한다. 주요어 불변화사는 동사 문자를 다른 문자들과 조합하여 단어를 형성하는 경우에 발생하고 형용사와 지시어(direction)를 그리고 동사와 지시어를 조합하는 두개의 서브타입을 포함한다.

개체명과 팩토이드에 대한 태깅 포맷은 아래 표 7에 제시되어 있다. 포맷-1은 사람이 빠르고 용이한 태깅을 행할 수 있게 하는 여러 타입과 서브타입에 대한 단순 태그를 포함한다. 예를 들면, 사람, 장소 및 기관에 대한 개체명은 단순히 P, L 및 O로 각각 태깅된다. 포맷-2는 제2 다국어 개체 작업 평가(MET-2)에 따라 표준화된 범용 마크업 언어(SGML)을 사용하는 태깅을 나타낸다. 원하는 경우, 포맷-1과 포맷-2 사이의 변환이 적절한 변환 프로그램을 통해 실현될 수 있다.

표 7

메인 카테고리	서브카테고리	포맷-1 태깅 모음	포맷-2 태깅 모음
사람	사람	P	사람
장소	장소	L	장소
기관	기관	O	기관
TIMEX	날짜	dat	DATE
	기간	dur	DURATION
	시간	tim	TIME
NUMEX	퍼센트	per	PERCENT
	돈	mon	MONEY
	회수	fre	FREQUENCY
	정수	int	INTEGER
	분수	fra	FRACTION
	소수	dec	DECIMAL
	서수	ord	ORDINAL
	비율	rat	RATE
MEASUREX	나이	age	AGE
	무게	wei	WEIGHT

	길이	len	LENGTH
	온도	tem	TEMPERATURE
	각도	ang	ANGLE
	영역	are	AREA
	용량	cap	CAPACITY
	속도	spe	SPEED
	기타 측정치	mea	MEASURE
주소	이메일	ema	EMAIL
	전화	pho	PHONE
	팩스	fax	FAX
	텔렉스	tel	TELEX
	WWW	www	WWW

표 7에서 태깅 포맷이 주어지면, 언어자료 내의 개체명과 팩토이드는 용이하게 태깅되어 주석달린 언어자료를 제공할 수 있다. 포맷-1과 포맷-2의 태깅 예는 아래에 제공된다.

포맷-1의 태깅:

예: on the morning of October 9th ---> on the[tim morning] of [data October 9th]

포맷-2의 태깅 포맷:

예: on the morning of October 9th ---> on the<TIMEX TYPE=TIME>morning </TIMEX> of <TIMEX TYPE=DATE> October 9th </TIMEX>

언어자료의 태깅 시에 항상성과 정확성을 보장하도록 일반적인 가이드라인을 제공하는 것이 유용하다. 다음 설명은 이들 가이드라인을 제공한다.

<u>일반 가이드라인</u>

(1) 개행을 위해 원문에 "엔터"를 하는 것이 회피되어야 한다.

(2) "-ms"로 표시된 태깅은 후술된다. 예는 [P-ms 邓小平]理论 "Deng Xiaoping theory"이다.

(3) 문자열을 멀티-태깅을 가지도록 허용된다. 주석자가 이러한 문자열에 대한 단일 태깅을 결정하기에 충분한 정보를 갖지 않으면, "/"이 다중 태깅을 위해 도입된다.

[L/O 西昌卫星发射中心]

(4) OPT: 주석자가 어떤 스트링이 태깅되어야 하는지 확신하지 못하는 경우, OPT 표시는 이 태깅이 검토 대상임을 나타내도록 도입된다.

[P/OPT 上帝]

<u>모든 개체명(사람, 장소, 기관)에 관련된 가이드 라인</u>

1. 적절한 명사는 객관적이고 특정 의미를 갖는 NEs이지만, 추상적이고 일반 의미를 갖는 NEs는 포함되지 않는다.

예: 표현, '老外Foreigner' '姑娘girl'은 적절한 명사가 아니다.

2. 합성되고 적절한 명사에 대하여, 임베디드 태깅은 허용되지 않는다. 즉, 최대 문자수를 갖는 분절 단어가 사용되는 경우에 최대 일치 접근법이 사용된다.

3. 인명, 지명 및 기관명에 임베디드되는 TIMES, NUMEX, MEASUREX 및 ADDRESS는 태깅되어야 하지 않는다.

[O 北京四中] --- 올바른 태그

[O北京[int 四]中] --- 틀린 태그

4. 개체 표현이 몇몇 문자열에서 영어와 중국어로 모두 포함하면서, 영어 문자열이 개체와 완전히 관련되는 경우에는 전체 표현이 개체명으로 태깅된다.

[O IBM中国公司]

[O American航空公司]

5. 소유격 구성에서, 소유자와 피소유 NE 서브문자열은 개별적으로 태깅되어야 한다. 중국어 철자 방식으로, 지시어(designator) "的"는 이러한 소유격 구성을 위한 기호이다.

[L 美国]的[L纽约]

[L 美国]的[P理查德本森]

주의: 문자열 "的"는 지시어로서 기능하지 않는 경우에는 개체의 일부로서 간주되어야 한다.

[O 美的电器集团]

6. 인용 표시는 개체명 내에 있고 개체명을 한정하지 않으면 태그 내에 포함된다. 중국어 텍스트에서, 제목 표시는 동일한 방식으로 처리된다.

[O“阿克布拉克”合资企业]

《[O 星岛日报]》的社论说

7. 분해 불가능한 합성 어구(complex phrase). 합성 표현이 전체로서 개체는 아니지만, 그 표현 내에 개체를 포함하고 있으면, 표현 내의 개체는 'P-ms', 'L-ms', 또는 'O-ms'로서 태깅되어야 한다.

주석자가 표현이 분해가능하지 여부를 확신하지 못하면, 그 표현은 분해가능으로 처리되고 그 안의 개체는 태깅되어야 한다. 예를 들어, [L_ms 香港]脚 "Hong Kong Foot"은 무좀과 동일한 의미이다. 이 표현은 전체로서 분해가 불가능하다. 가이드라인에 따르면, 단어 "Hong Kong"은 지명 'L-ms'로서 태깅될 수 있다. 예를 들어, [ord 第四十六] 届 [O太平洋亚洲旅行协会] 年会 "Forty-sixth Pacific Asia travel Association annual meeting"은 가이드라인에서 분해가능한 것으로 다루어진다:

'太平洋洲旅行协会 Pacific Asia travel Association'은 기관으로서 태깅되지만, '太平洋亚洲旅行协会年会 Pacific Asia travel Association annual meeting'은 기관이 아니다.

'인명 + 사상(또는: 이론, 법률, 이상)'의 표현에 있어서, 전체 표현은 'p-ms'로서 태깅되어야 한다.

[P_ms 马克思] 主义 "마르크스 이론"

[P_ms 毛泽东] 思想 "마오쩌둥 사상"

[P_ms 阿佛加罗] 定律 "아보가드로의 법칙"

8. '军' (...육군/...군대...)의 처리. 주요 구별은 영어의 'military'와 유사하게(즉, 비민간인)와 유사하게 '军'을 형용사로 해석하는 것과 "기관 지시자"로서 '军'을 해석하는 것이다. 후자의 해석을 위해서는, '军' 앞에 어느 병역의 '종류' 지시가('공군에서 空'공' 과 같은) 선행하는 지를 보는 것이다.

[L美]军飞机 "미국 공군"

[O斯里兰卡空军] "스리랑카 공군"

통상, 기관으로서 部队 "부대"로 끝나는 용어는 태깅하지 않는다. [L西非] 维和部队"서아프리카 평화 유지부대", 军事基地 "군사기지"는 장소로서 태깅되어야 하며, 기관으로 태깅되어서는 안된다. [L彼得森空军基地]"피터슨 공군 군사 기지".

9. 개체명(인명, 지명, 기관명)에 있어서, 멀티미디어(TV 및 라디오 쇼, 영화 및 책), 제품 또는 조약의 종류이면, "-ms" 태그로 태깅되어야 한다.

[P-ms 邓小平]一片的播出 "Deng Xiaoping" (CL-for-film)의 릴리스, 즉, "Deng Xiaoping"의 영화의 릴리스.

'邓小平Ding Xiao Ping'은 TV프로그램의 제목이기 때문이다. 가이드라인에 따르면, "Ding Xia Ping"은 'P-ms'로서 태깅되어야 한다.

[L_ms 广州]条约 《[L_ms 淮海]战役》这本书的出版

10. 가칭, 별칭, 개체의 약칭은 태깅되어야 한다.

[O ETS]
"[O深蓝]"
[O IBM]
[L沪]
[O 北约] <

개체명이 개체의 약칭으로 임베디드되면, 이는 태깅되지 않아야 한다. [O中共中央政治局],에서, '中'은 '중국'을 의미하며, 中은 마킹하지 않는다.

<u>사람에만 관련되는 가이드라인</u>

1. 사람의 직위

직위와 역할은 사람의 이름의 일부로 간주되지 않는다.

[P奥尔布赖特]国务卿 "올브라이트 국무부 장관"

[L英国]女王[P伊丽莎白] "영국 엘리자베스 여왕"

그러나, 세대 지시자인 "世", "代"는 인명의 일부로 간주된다.

[P 十四世达赖丹增加措] "제14대 달라이 텐진 기아쵸(tenzin gyatso)"

[L英国]女王[P 伊丽莎白二世] "영국 여왕 엘리자베스 2세"

사람의 직위가 성과 이름 사이에 있는 경우, 직위를 포함한다.

[P李主席登辉]先生 "이 주석 등휘 선생"

2. 가족명은 사람으로 태깅되어야 한다.

[P蒋]氏父子 "장씨, 부자"

[P西迪]兄弟 "서유 형제"

3. 동물명은 인명으로 태깅되어야 한다.

4. 성직자와 다른 종교 인물은 적절한 명칭이 인명으로 태깅되어야 한다.

[P 释迦穆尼]
[P 达赖]喇嘛

5. 소설 인물은 인명으로서 태깅되어야 한다.

6. 소설 둥물명과 사람이 아닌 캐릭터는 인명으로서 태깅되어야 한다.

7. 사람의 직위 또는 왕위가 특정인을 지칭하는 경우, 인명으로서 태깅된다.

[P 康熙] "강희, 즉 강희 황제"

[P 秦始皇] "진시황제"

[P 老子] "노자"

8. 여러가지 인명 태깅 대상이 아닌 것.

인명이 멀티미디어(TV 및 라디오 쇼, 영화 및 책), 제품 및 조약의 제목으로서 나타나는 경우, 이 이름들은 'p_ms'로서 태깅되어야 한다.

《[P_ms蒙娜丽莎]》 미술작품명(또는 책 제목)으로서의 "Mona Lisa"는 "P_ms"로 태깅되어야 한다.

다음 5가지 경우에서, 적절한 이름이 인명으로 태깅되지 않아야 한다: 인명을 따른 법률, 인명을 따른 법정 사건, 인명을 따른 날씨 형식, 질병/상장.

里氏六点二级 --- '里'에는 태그 없음

专家呼吁人们要注意沙氏杆菌 --- '沙'에는 태그 없음

[P_ms 诺贝尔]奖 --- '诺贝尔Nobel'을 'P_ms'로 태그

9. 중국어 이름의 일반 패턴

일반적으로, 성명은 두 성분: 성(FN)과 이름(GN)으로 구성된다.

#	성명 패턴	태깅 방식	예
1	성만 있는 경우(FN)	FN 태그	[P 李]
2	이름만 있는 경우(GN)	GN 태그	[P志东]
3	FN+ GN	전체 성명을 태그	[P王志东]
4	a. 성명(전체 성명 또는 성만 또는 이름만) + 직위 b. 직위+ 성명	성명만 태그, 즉 직위에는 태그안 함	[P李]教授 [P王志东]教授 [P志东]教授 [马]厂长 직위는, 대통령, 수상, 장관, 시장, 교수, 선생, 박사, 연구원, 주임 연구원, 의장, CEO 등
5	접두사+ 성명 성명+ 접미사	성명만 태그	大[P李] [P李]总
6	성명+ 성명	개별적으로 성명을 태그	[P李向东] [P李向阳]
7	외국인 성명	전체 성명을 태그	[P马拉多纳] [P比尔.盖茨]---- 문자 '.'가 인명 가운데 나타나면, 성명은 전체 개체로서 간주된다.

장소에만 관련되는 가이드라인

장소로서 태깅되는 문자열은, 바다, 대륙, 국가, 지방, 구, 시, 지역, 도로, 마음, 도시, 공항, 군사기지, 도로, 철로, 다리, 강, 바다, 운하, 해협, 만, 직선로, 모래 해변, 호수, 공원, 산, 평지, 초원, 광산, 박람회장 등, 소설 또는 신화 장소 및 에펠탑과 링컨상 등의 특정 구조물을 포함한다.

[L北京市] [L海淀区] [L知春路49号] "베이징 시, 하이디안 구, 지춘로 49호"

[L朝鲜] 南北对话 "코리아 남북 대화"에서 코리아에는 태깅하지만, 남/북에는 태깅하지 않으며, 阿[L以]冲突 "아랍과 이스라엘 충돌"은 이스라엘에는 태깅하지만 아랍에는 태깅하지 않으며, 이는 아랍이 특정 국가를 가리키는 것이 아니기 때문이다.

前[L 南]地区 "전 유고 지역"

震中位于[L 北纬三十六点二零度，东经九十点二九度]

"북위 36.0도 동경 95.9도에 위치한 진앙지".

1. 다른 지명에 임베디드된 지명에 대하여, 전체 엔티티가 태깅되어야 한다.

[L 美国空军基地] "미국 군사기지",는 …地区 "...지구/...영역"의 미국식 처리에 태깅하지 않는다. 地区이 특정 지구를 의미하면, 장소의 일부로서 태깅되어야 한다: '地区'이 일반적으로 특정 영역을 의미하면, 태깅되지 않아야 하고; 地区의 지점이 불명확하면, 태깅되지 않는다. [L 临沂地区]现更名为[L 临沂市] "린 이(Lin Yi) 지구는 지금 린 이 시로 개명하고 있다" 지명에 임베디드된 기관명에 있어서, 기관명은 태깅되지 않는다. [L 白宫玫瑰园] "백악관 장미원"은 백악관에 태깅하지 않는다.

2. 지명 지시자는 장소의 일부로서 태깅되어야 한다.

[L马里兰州] "메릴랜드 주"

[L 约旦河] "요르단 강"

장소명이 연속으로 열거된 복합 표현은 장소의 개별 인스턴스로서 태깅되어야 한다.
[L吉林省] [L 延边朝鲜族自治州] [L 图们市] "길림성 연변 조선족 자치주 두만시"

3. 다국적 지명 표현

[L 西非]国家领导人 "서아프리카 국자 지도자" [L 亚太] "아태 지역",은 하나의 개체로서 태깅되며, [L 西半球]国家 "서반구 국가", 发展中国家은 마크업하지 않는다.

국가내 영역은:

[L 华南] "남중국"

[L 西北五省区] "서북 5성"

使西南地区的客运 "서북지역의 관광객 서비스", 고정 기준 [L 华南]地区 "남중국 지구"이 없고, 여기서 남중국이 고정 기준이므로, "서북"에 마크업하지 않는다.

4. 지명 표현의 시간 변형. 역사적 시간 변형("구")은 태깅된 표현에 포함되지 않아야 한다.

前[L 南]地区 "구 유고 지구"

5. 지명 표현의 띄어쓰기 변형

[L 北爱尔兰] "북 아일랜드"

[L中西伯利亚] "중앙 시베리아"

[L 中] [L 南美] "중앙 및 남 아메리카" 이 표현은 두개의 지명 "중앙 아메리카"와 "남아메리카"를 포함하므로, 개별적으로 태깅되어야 한다.

6. 여러개의 지명 태깅가능하지 않는 것:

x-语 또는 x-文 의 형태 - 여기서, x는 장소 - 의 언어명인 지명에는 태그하지 않는다.

英语 "영어"에서 '英'에는 태그하지 않으며, 中文 "중국어"에서 '中'에는 태그하지 않는다.

x-话 의 형태 - 여기서, x는 장소 - 의 지명에는 태그한다. 用[L四川]话 "사천어 사용"은 四川의 장소에 태그한다.

7. 종족에서 族 또는 裔 으로 끝나는 명칭의 일부인 지명에는 태그하지 않는다.

目的是促进[L塞浦路斯]西族与土族的瓦解

"그 의도는 사이프러스 그리스 민족과 터키 민족 간의 평화와 이해를 증진하기 위한 것이었다"

이 표현에서, '华裔'、'汉族','华' 및 '汉'은 장소로서 태깅되지 않아야 한다. 그러나, 표현에서,

华人'、'华侨 '、'华商' 、'中医' 、'中草药' 、'中餐馆' ,'华' 및 '中'은 장소로서 태깅되어야 한다.

8. 장소의 일반 패턴

#	장소 패턴	태그 방식	예
1	지명만(LN)	LN 태그	[L山东]
2	LN + 지명 지시자	전체 표현을 태그	[L北京市] [L天安门广场]
3	명칭이 연속적으로 배치되는 복합 표현	개별 태그	[L山东省][L青岛市][L胜利广场]; [L北京]、[L天津] 、[L上海]
4	가칭 또는 별칭이 연속적으로 열거	개별 태그	[L鲁]、[L 冀]、[L京]; [L港][L澳][L台]地区 ; [L中][L俄]两国领导 人进行了会晤
5	LN 표현이 인명 또는 지명을 포함	인명 또는 지명에 태그 안함	[L李嘉诚广场] [L 南京路]
6	LN+ L 지시자가 전체로서 완전한 개념을 표현	최대 일치 접근법을 사용하여 표현을 태그	[L南非共和国] [L香港特别行政区]

<u>기관에만 관련되는 가이드라인</u>

기관으로 태깅되어야 하는 적절한 명칭은 증권거래소, 다국적 기관, 비즈니스, TV 또는 라디오 방송국, 정당, 종교 단체, 오케스트라 밴드, 또는 음악 그룹, 연합, "의회" 또는 "하원"과 같은 일반적이기 많은 정구 기구, 스포츠 팀 및 군대 (장소로 태깅되는 지명으로 나타재지 않은 경우) 뿐만 아니라 부속 기관을 포함한다.

회사 또는 기관 지시자는 기관명의 일부로 간주된다. 장소 태깅에 대한 기본 원리는 최대 일치 접근법을 사용하는 것이다.

前[O中国新华社香港分社]社长[P 许家屯]

"구 중국 신화사 항공 분사 사장 수 지아툰"

[O北京大学计算机系人工智能实验室] "북경대학 전산학부 인공지능 랩"

기관에 대한 일반 패턴

#	유형	태그	예
1	기관명 + 지시자	전체를 태그	[O海尔集团]
2	지명+ 기관명	전체를 태그	[O北京市电信局]
3	인명+ 기관명	전체를 태그	[O李嘉诚基金会]
4	가칭 또는 약칭	전체를 태그	[O北约]

1. 국가(국제) 입법부 및 부서 또는 수상은 기관으로 태깅되어야 한다.

当选[O 国会]议员

[O 内阁]改组将会在[dat八月]底前完成

在[O 总统府]分别约见了多位[O 国民党]中常委检察官

[P 刹瓦什]向[O 宪政法庭]提出动议

2. 기관명 바로 앞에 있는 지명의 처리. 일반적으로, 장소와 기관 간의 관계는 두개의 유형이 있다: 하나는 (法国航空航天局 "프랑스 항공 우주국" 등) 행렬식이고, 다른 하나는 (北京大学 "북경 대학교" 등) 지형 연결이다.

2.1. 지명으로 시작하는 기관 개체에 대해서는, 장소를 제거하면 특정되지 않는 장소가 되는 경우에는, 지명은 기관의 일부로 태깅되어야 한다.

[O北京大学] "북경 대학교"

[O深圳中学] "심천 중학교"

2.2. 상술한 기관 표현에 있어서, 이에 바로 앞서 하나의 지명(또는 하나 이상의 지명)이 있는 경우, 그 지명과 기관 표현은 별도로 태깅되어야 한다.

[L 中国] [O北京大学] "중국 북경 대학교"

[L 中国] [L广东] [O深圳中学] "중국 광동성 심천중학교"

2.3 비지명(non-location) 문자열로 시작하는 기관 개체(同济大学 "동제 대학교" 등)에 있어서, 이에 앞서 하나의 장소(또는 하나 이상의 장소)가 있으면, 이에 바로 앞서는 장소만이 기관의 일부로서 태깅되어야 한다.

[O上海同济大学] "상해 동제 대학교"

[L 中国] [O上海同济大学] "중국 상해 동제 대학교"

[O湖北省武钢三中] "호북성 무강 제3 중학교"

2.4 기관 개체가 둘 이상의 병렬 지명으로 시작하면, 이들 모든 장소들이 기관의 일부로서 태깅되어야 한다; 전체 기관 뒤에 다른 장소(들)가 있으면, 장소와 기관은 별도로 태깅되어야 한다.

[L 洛杉矶] [O亚太法律中心] "로스앤젤레스 아시아 태평양 법률센터"

[L 香港] [O中港贸易协会] "홍콩, 중국, 홍콩, 무역 협회"

2.5 몇몇 복잡한 경우에는, 기관명이 하나 또는 둘의 장소로 시작하는지 불명확하며, 태깅은 규칙 2.1 및 2.2에 따라 행해져야 한다.

예: 洛杉矶台北经济文化办事处 "로스앤젤레스 대만 경제문화 사무소", A: [L 洛杉矶] [O台北经济文化办事处] 또는 B: [O洛杉矶台北经济文化办事处] 를 태깅할 지, 이 경우는 기본값으로 태깅 A를 선택한다.

2.6 주석자가 기관이 장소로 시작하는지를 결정하기에 충분한 지식이 없는 경우.

예: "印度尼西亚莫巴蒂 努山打腊航空公司" 의 표현에서, 주석자는 莫巴蒂 努山打腊 이 장소명칭인지 확신하지 못한다. 그러나, 일단 이 문자열이 제거되면, 좌측 문자열은 특정 참조가 없음이 명확하다. 따라서, 2.1에 따라서, 이 표현은 [L 印度尼西亚] [O 莫巴蒂 努山打腊航空公司]과 같이 태깅되어야 한다.

2.7 지명 개체가 기관을 바로 뒤따르는 경우, 이들 가운데 어떤 변형 관계도 존재하지 않으면, 이들은 별도로 태깅되어야 한다.

促进了[L 中国] [O 东盟]的合作 "중국과 동남아시아 간의 협력을 증진함"

在 [L 日内瓦] [O 联合国] 人权会议上 "제네바 UN 인권 회의 상에서"

3. "…会" (미팅, 회의, 예술 페스티벌, 운동 경기)로 끝나는 어구는 이벤트를 나타내며, 기관으로 태깅되어서는 안된다. 그러나, 조직 구조 자체 -예를 들면, 지도부 등 --는 기관으로 태깅되어야 한다.

奥运会 "올림픽 스포트 경기"

[O 奥运会组委会] "올림픽 위원회"

어구 "…会"는 "의회" 또는 "하원"을 가리키는 경우, 이들은 기관으로 태깅되어야 한다. 의회(또는 하원)의 회의는 이벤트이기 때문에, 기관으로 태깅되지 않아야 함에 주의하자.

[O 全国政协] 八届五次会议将于
听取和审议 [O 全国政协八届五次会议常务委员会] 报告
[O 九届人大] 一次会议

4. 1인칭 대명사 "我","我们"이 기관 개체에 선행하는 변형자로서 기능하는 경우, 이 대명사는 기관의 일부로서 태깅되어서는 안된다. 我国 [O共产党] "내 나라 공산당" 我们 [O清华大学] "우리 칭화 대학교".

5. 대사관과 영사관

대사관, 영사관 및 다른 외교 사절의 명칭은 그들이 나타내는 나라와 그들의 장소가 마크업에 포함될 수 있는 경우에만 기관으로 표시되어야 한다.

后来调任 [O美国驻洪都拉斯大使馆] "온두라스 주재 미국 대사관으로 전달됨"

대사관 기술자(descriptor)가 그것이 나타내는 나라/지역과 연결되면, 나라/지역은 기관의 일부로서 태깅되어야 한다.

前往 [L香港] 的 [O洪都拉斯领事馆] "홍콩에 있는 온두라스 대사관으로 가라" 대사관 기술자가 지형 위치와 연결되어 있으면, 장소로서 임의의 위치를 별도로 표시하고, 대사관은 기관으로 태깅하지 않는다.

[L美国] 在通过驻 [L金沙萨] 大使馆和其他正常渠道 "미국의 킨샤사 주재 대사관과 다른 정상 채널을 통과"

6. 제조물 및 제품

제조물 및 제품이 명명된 경우, 제조물은 기관으로 태깅되어야 하지만 제품은 태깅되어야 하지 않는다. 제품은 제조품(예를 들어, 차량) 뿐만 아니라 계산치(예를 들어, 증권지수) 및 미디어 제품(예를 들어, 텔레비젼 쇼)를 포함하도록 넓게 정의되어야 한다.

[O 道琼] 工业平均指数 "다우 존스 산업 평균 지수".

7. 뉴스원(신문, 라디오 및 TV 방송국, 및 뉴스 저널)은 기관으로 태깅한다. 출판사와 출판물은 기관으로 태깅되어야 한다. TV 방송국은 TV 쇼와 상이하며, 후자는 태깅하지 않는다.

[O人民日报] 海外版第三版 "인민" 일보 해외판 제3판"

这是 [O 中央台] 报道的 "이는 중앙 보도국이다"

8. 기관유사 비태깅 대상

"정부"와 같은 일반적인 개체명은 태깅되지 않아야 한다.

[L中国] 政府 "중국 정부"

[L 新疆自治区] 政府 "신장 자치구 정부" [O中国公安部] 门 "중국공안부(들)".

中央 "중앙"이라는 용어는 그 자체로서는 기관으로 표시하지 않는다. 그러나, 党中央 은 "당 중앙"은 기관으로 표시한다.

在中央的领导下 "중앙 영도 하에서"

以[P江泽民]同志为核心的[O 党中央]周围 "장쩌밍 동지를 핵으로 하는 중앙 당". 交易会' "박람회"를 기관으로 태그하지 않는다.

[L 中国][L 天津]出口商品交易会 "중국 천진 출구 상품 박람회"

9. 여러 고유 명사에 대한 태그

[L 人民大会堂] "인민대회당"

[O 白宫] "백악관"

[O 克里姆林宫]表示 "크레믈린에 따르면"

TIMEX 태그 방식

TIME 유형은 "초, 분 또는 시" 등의 전체 하루보다 짧은 시간 단위로서 정의된다. DATA 서브타입은 "날, 주, 달, 분기, 년, 세기 등"의 전체 하루 또는 그 이상의 시간 단위이다. DURATION 서브타입은 기간을 캡쳐한다.

1. 날짜

문자열 前/头/下+ + 기간의 형태에서, 기간은 태깅되지 않도록 DAT에서 임베디드되지 않기 때문에 전체 어구는 dat_MET로서 태깅된다.

[dat_MET 前3天] "처음 3일"

[dat 秋季]报告 "가을 보고서"

[dat第四季度] "4분기"

[dat 十五世纪] "15세기"

[dat春节] "봄 축제"

문자열 "(上/中/下)旬 한 달의 상순/중순/하순은 태깅되어야 하고, [dat五月上旬]은 "5월의 하순"과 같은 '주변' 또는 '근방' 등의 표현을 변형하는 단어 또는 어구이다.

大约[dat五月四日] "5월 4일 근방"

2. 시간

[tim 凌辰三四点钟] "오전 서너시"

[tim 北京时间5时59分] "베이징 시각 5시 59분"

[tim_MET上午]、[tim_MET中午]、[tim_MET下午]、[tim_MET晚上] "오전, 정오, 오후, 저녁"

"大约 약/대략"의 처리

[tim 晚上大约七点] 到达 "저녁 약 7시에 도달"

이 어구에서, 문자열 '약'은 두 시간에 의해 한정되며 분해가능하지 않으므로, 태깅되어야 한다.

[dat 九月十三日] 大约 [tim七点] 到达北京 "베이징에 9월 13일 7시 정각에 도달"

이 어구에서, 문자열 大约 은 날짜와 시간에 의해 한정되므로, 분해가능하다.

3. 기간

[dur 10天] "10일"

在水门丑闻 [dur 四分之一世纪] 时发表的评论 "워터게이트 사건 이후 사반분기의 논의"

포함 여부가 차이가 거의 없기 때문에, 문자열 "整整"은 기간 태그에서 포함되어야 하지 않는다.

整整[dur 十五年] "정확히 15년"

[dur 九点]整到达北京站 "베이지 역에 정확히 9시 정각에 도달"

十年九旱 "십년에 9년은 가뭄, 즉, 자주 가뭄을 겪음", 모두 가상 번호(virtual number)의 경우이기 때문에, '9'와 '10'에는 마크하지 않는다.

4. 태그 비대상

"방금 전, 최근, 협상 이후, 순간" 등의 절대 시간 척도를 갖는 시간 표현은 태깅되지 않아야 한다. 축제 표현이 절대 시간을 갖지 않는 경우에는 태깅되지 않는다.

[L 印度]国际电影节 "인도 국제 영화 페스티벌"

[L 中国]旅游年 "1997년은 중국 방문의 해"

[L美国]的独立日 "미국 독립 기념일", 이벤트에 근접으로 인한 독립기념일에 대해 마크업하지 않음.

春联 "봄 2행 연구(spring couplet)"에서 春 "봄"을 태그하지 않는다.

5. 특별 케이스:

두 시간 표현이 상이한 서브타입이면, 그들은 별도로 태깅되어야 한다. 두 표현은 분해가능하지 않으면, 함께 태깅되어야 한다.

[dat 2月12日][tim 上午8点] "2월 12일 오전 8시 정각"

[dat星期一][tim 8点] "월요일 8시 정각"

장소 개체가 시간 표현 내에 임베디드되면, '_MET' 표시는 MET-2 가이드라인을 지칭하도록 도입된다. "ER99"는 다른 사양에 따라 태깅하는데 사용될 수 있다.

[tim北京时间1997年2月9号19点28分]

"작년", "어제", "오늘 아침"과 같은 표현은 MET-2에 따라 태깅되어야 하며, 그 차이를 주석자에게 알리기 위해 여분의 표시를 그에 따라 사용한다.

[dat_MET 去年[dat_ER99 上半年]]

[dat_MET 今年[dat_ER99 夏天]]

[dat_MET 今年[dat_ER99 三月一日]]

[dat_MET 今年[dat_ER99 4月17日]] [tim_MET 下午]

[dat_MET 去年[dat_ER99春夏之交]]

[tim_MET 昨天[tim_ER99 夜里]]

[dat_MET 今天][tim_MET 晚上]

[tim_MET 今早[tim_ER99 六点]]

[tim 早上六点]

[dat_MET 当日][tim_MET 下午]

[dat_MET 当日][tim 下午16时30分]

每日[tim_MET [tim_ER99 上午11时]至[tim_ER99 深夜3时]]

[tim_MET 晨]练、[tim_MET 晚]宴

"今早 오늘 아침"의 표현에 있어서, ER-99는 상대 시간 개체로서 이를 처리하고, 태깅되어야 하지 않지만, MET-2에서는 태깅되어야 한다.

[dur_ER99 [dat_MET [dat_ER99 11月24] 至 [dat_ER9927日]]]

[dat_MET [dat_ER99 11月24] 至 [dat_ER9927日]]

[tim_MET 昨夜]

迄[tim_MET今]

[tim_MET今]后

"数年 수년"의 표현에 있어서, ER-99은 고정 기간으로서 이를 처리하고, 태깅되어야 하지만, "多年 여러 해"는 비고정(non-fixed) 기간이고 태깅되지 않아야 한다.

一年 "일년"의 표현은 기간으로 태깅되어야 한다.

新的[一年] 即将开始

入伍 [dur 一年]多 的时间里

硬是在地下室干了 [dur 一年] 的公司

一年创产值效益……

一年便多收入……

聘金为一年 [mon 900万美元] 的价码

"每年 매년" / "年 년, 해마다"라는 표현은 不标注 年产值…… 每年创产值效益…… 每年收入…….

NUMEX 태그 방식

1. 퍼센트

[per 百分之三十九] "39퍼센트"

大约 [per 5%] "약 5퍼센트"

[per 九成] "90퍼센트"

2. 돈

[mon 四万五千块钱] "4400 위안"

[mon 四万五千人民币] "4400 RMB"

[mon 人民币四万五千元] "RMB 4400 위안"

● 동일 액의 돈이 상이한 통화의 철자를 갖는 경우에는, 별도로 태깅되어야 한다. 돈에서 임베디드된 장소는 태깅되지 않아야 한다.

[mon 43.6亿美元] "436억 USD"

● 문자열 "约 약"은 절대 개념을 갖지 않으므로, 태깅되지 않아야 한다.

约 [mon 十万元] "약 10만 위안"

多于 [mon $90,000] "$90000 이상"

● 문자열 "几 여러"는 특정 수로 변경될 수 있으며 절대 액을 표현하므로, 태깅되어야 한다.

[mon 几十万元] "수십만 위안"

● 문자열 "余 이상"은 일반적으로 태깅되지 않아야 하며; 다음 경우, 전체 표현이 분해가능하지 않으므로 태깅된다.

[mon 二十七万余元] "27만 이상의 위안"

● 이 가이드라인에서, 통화에 임베디드된 지명에 있어서, 약어로 기재되면, 태깅되지 않으며, 그렇지 않으면 '-ms'로서 태깅되어야 한다.

[mon 2000新元] "2000 SID"

[mon 2000 [L_ms 新加坡] 元] "2000 싱가포크 달러 위안".

3. 회수/정수/분수/소수/서수

[fre 26次]
[fre 十多次]
[fre 多次]
[fra 3/4]
[fra 四分之三]
[fra 百万分之八]
[fra 百万分之三百六十四]
[fra 半]
[fra 4倍半]
[dec 3.14]
[dec 三点一四]
[ord 第二] 故乡
[ord 1174号]文件
[ord 6路]汽车
[ord 第一]天
[ord 第二]年
[int 20 名]杰出教师
[int 亿万] 人民
[int 几千万盆]

정수/분수/소수가 변형자로서 단위를 가지면, 단위는 태깅되어야 한다.

[int 几家] 工厂 "여러 가구('jia' factories)" 一家 [int 5口] 人 "하나의 가족에 5명의 가족('kou' persons)" [int 58倍] "58배".

4. 특별한 경우

● 탭 번호는 태깅되지 않는다.

一靠政策调动农民的积极性；
二靠科技；
三靠投入
1．自卑的羞耻感。
2．依赖的恐惧感。
3．温饱即安的安全感。
(1)加强爱国主义的宣传教育。
(2)加强正确的理想、信念、人生观、价值观的宣传教育。
(3)加强马克思主义的唯物辩证法的宣传教育。

● 일부 관용구에서 숫자, "一会儿 잠시", "一起 함께", "一流 일류", "唯一 유일한" 등은 태깅되지 않아야 한다.

● 인명, 지명 또는 기관명에 임베디드된 숫자는 태깅되지 않아야 한다.

[O 一中] "제1 중학교"

<[L三明市] "삼명시"

任队长的 [O 1205钻井队]

● 문자열 "一"이 관사 'a'로서 기능하는 경우, 태깅되지 않는다. "一倍 한 배 이상"은 태깅되어야 한다. 서수의 일부로서 "一"은 태깅되어야 한다.

一座城市 "도시(a city)"

最大的企业之一 "최대 회사들 중 하나"

[ord 一等]奖 "1등 상"

我的收入是它的[int 一倍] "내 수입은 그보다 1배 이상이다."

MEASUREX 태그 방식

MEASUREX는, 나이, 무게, 길이, 온도, 각, 면적, 용량, 속도 및 비율을 포함한다.

[age 34岁]

[age 六十寿辰]

[age 花甲]老人

产量达到 [wei 数千万吨]

开掘到 [len 一米六七] 深度时

高 [len 五米] 宽 [len 一百米]

积温高([tem 2800度])

钝角就是大于 [ang 90度] 的角

农田 [are 20万亩]

运输量为 [cap 34个立方]

一[cap两箩]谷子

最高速度 [spe 360米每秒]

[wei 二十万吨] 级以上

[tem 零下5] 到 [tem 6摄氏度]

주의: 물리학과 화학에서 무게 및 측정의 서로 다른 단위에 대하여, 이들은 "mea"로 태깅되어야 한다.

[mea 5.5瓦特] "5.5 와트"

[mea 1.5 牛顿] "1.5 뉴튼"

ADDRESSX 태그 방식

ADDRESX는 이메일, 전화, 팩스, 텔렉스, WWW를 포함한다.

[ema exp@email.com.cn]

Tel: [pho 86-10-66665555]

电话: [pho 86-10-66665555]

FAX: [fax 86-10-66665555]

TELEX: [tel 86-10-66665555]

[www http:---- www.hotmail.com]

전화 또는 팩스 번호에 있어서, "tel, 电话 등의 지시자만이 있는 경우에는 태깅되어야 한다.

본 발명은 특정 실시예를 참조하여 설명하였지만, 당업자는 본 발명의 취지 및 범위를 벗어남이 없이 형태 및 세부사항에서 변경이 행해질 수 있음을 인식할 것이다.

발명의 효과

상술한 본 발명에 따르면, 복수의 문자열과 복수의 문자열에 관련되는 복수의 형태론적 태그를 포함하는 언어자료를 구현하고, 상기 텍스트와 분절된 텍스트를 나타내는 출력에서 형태론적으로 유도된 단어를 검출하고, 성분 조합의 표시를 제공함으로써, 우수한 단어 분절을 제공할 수 있다.

(57) 청구의 범위

청구항 1.

언어 모델을 트레이닝하기 위해 컴퓨터 판독가능 매체에 저장되는 언어자료(corpus)에 있어서,

복수의 문자(character); 및

상기 복수의 문자의 복수의 문자열(sequence of character)에 관련되고, 관련 문자열의 형태론적 타입 및 형태론적 서브타입을 형성하는 성분(part)의 조합(combination)을 나타내는 복수의 형태론적 태그(morphological tag)를 포함하는 언어자료.

청구항 2.

제1항에 있어서,

상기 형태론적 타입은, 접사 첨가(affixation), 음절 중복(reduplication), 분할(split), 병합(merge) 및 주요어 불변화사(head particle)인 언어자료.

청구항 3.

제1항에 있어서,

상기 형태론적 타입은 접사 첨가이고, 상기 성분의 조합은 단어와 적어도 하나의 접두사와 접미사를 포함하는 언어자료.

청구항 4.

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 성분의 조합은 상기 단어에 대한 품사(part of speech)를 나타내는 언어자료.

청구항 5.

제1항에 있어서,

상기 형태론적 타입은 음절 중복이고, 상기 성분의 조합은 문자의 패턴을 포함하는 언어자료.

청구항 6.

제1항에 있어서,

상기 형태론적 타입은 병합이고, 상기 성분의 조합은 문자의 패턴을 포함하는 언어자료.

청구항 7.

제1항 내지 제6항 중 어느 한 항에 있어서,

문자열이 팩토이드(factoid)인지의 여부를 제공하는 복수의 팩토이드 태그를 더 포함하는 언어자료.

청구항 8.

제1항 내지 제7항 중 어느 한 항에 있어서,

문자열이 개체명(named entity)인지의 여부를 제공하는 복수의 개체명 태그를 더 포함하는 언어자료.

청구항 9.

제1항 내지 제8항 중 어느 한 항에 있어서,

문자열이 어휘목록(lexicon) 내에 포함되는지 여부를 더 포함하는 언어자료.

청구항 10.

단어 분절(word segmentation)을 수행하기 위한 명령들을 갖는 컴퓨터 판독가능 매체에 있어서, 상기 명령들은,

비분절 텍스트(unsegmented text)의 입력을 수신하는 단계;

언어 모델에 액세스하여 상기 텍스트의 분절을 결정하는 단계;

상기 텍스트에서 형태론적으로 유도된 단어(morphologically derived word)를 검출하는 단계; 및

분절 텍스트(segmented text)의 출력과 상기 형태론적으로 유도된 단어를 형성하는 성분 조합의 표시를 제공하는 단계를 포함하는 컴퓨터 판독가능 매체.

청구항 11.

제10항에 있어서,

상기 명령들은 상기 형태론적으로 유도된 단어가 접사 첨가, 음절 중복, 분할, 병합 및 주요어 불변화사 중 하나임을 나타내는 단계를 더 포함하는 컴퓨터 판독가능 매체.

청구항 12.

제10항 또는 제11항에 있어서,

상기 명령들은 상기 텍스트 내의 어휘목록을 검출하는 단계를 더 포함하는 컴퓨터 판독가능 매체.

청구항 13.

제10항 내지 제12항 중 어느 한 항에 있어서,

상기 명령들은 상기 텍스트 내의 팩토이드를 검출하는 단계를 더 포함하는 컴퓨터 판독가능 매체.

청구항 14.

제10항 내지 제13항 중 어느 한 항에 있어서,

상기 명령들은 상기 텍스트 내의 개체명을 검출하는 단계를 더 포함하는 컴퓨터 판독가능 매체.

청구항 15.

제10항 내지 제14항 중 어느 한 항에 있어서,

상기 출력 제공 단계는, 상기 성분 조합에서 품사를 나타내는 단계를 더 포함하는 컴퓨터 판독가능 매체.

청구항 16.

제10항 내지 제15항 중 어느 한 항에 있어서,

상기 출력 제공 단계는, 상기 성분 조합을 형성하는 문자 패턴을 나타내는 단계를 더 포함하는 컴퓨터 판독가능 매체.

청구항 17.

언어 모델을 트레이닝하기 위한 언어자료를 개발하는 방법에 있어서,

정의된 단어와 규칙에 일치하는 언어자료로부터 가능한 단어 리스트를 추출하는 단계;

상기 리스트가 충분한 수의 정의된 단어와 규칙을 포함하는지 결정하는 단계;

상기 언어자료에 주석을 달아서 단어형의 표시를 제공하는 단계; 및

관련 문자열의 형태론적 타입 및 형태론적 서브타입를 형성하는 성분 조합을 나타내는 상기 어휘자료내의 형태론적 태그를 제공하는 단계를 포함하는 언어자료 개발 방법.

청구항 18.

제17항에 있어서,

상기 주석 단계는, 상기 단어가 어휘자료, 형태론적으로 유도된 단어, 팩토이드 및 개체명인지 여부를 제공하는 단계를 더 포함하는 언어자료 개발 방법.

청구항 19.

제17항 또는 제18항에 있어서,

상기 형태론적 타입은 접사 첨가, 음절 중복, 분할, 병합 및 주요어 불변화사 중 하나인 언어자료 개발 방법.

청구항 20.

제17항 내지 제19항 중 어느 한 항에 있어서,

상기 형태론적 태그 제공 단계는, 상기 성분 조합에 대한 품사를 나타내는 단계를 더 포함하는 언어자료 개발 방법.

청구항 21.

제17항 내지 제20항 중 어느 한 항에 있어서,

상기 형태론적 태그 제공 단계는 상기 성분 조합에 대한 문자 패턴을 나타내는 단계를 더 포함하는 언어자료 개발 방법.

청구항 22.

제17항 내지 제21항 중 어느 한 항에 있어서,

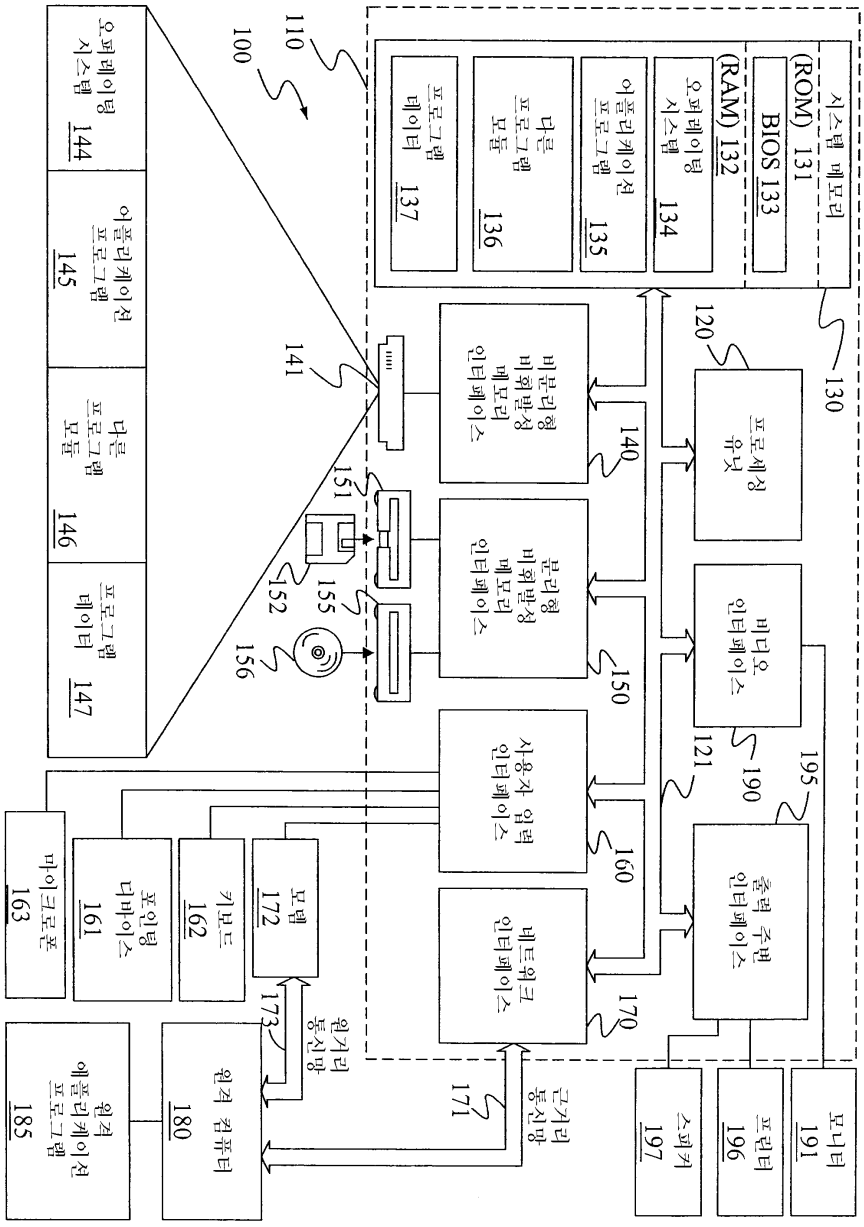
상기 언어자료 내에 형태론적 태그를 제공하는 단계 후에, 상기 언어자료를 사용하여 보다 많은 텍스트 양에 주석을 다는 언어자료 개발 방법.

청구항 23.

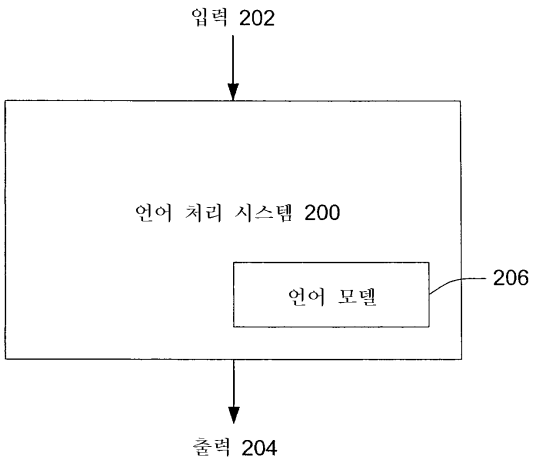
제17항 내지 제22항 중 어느 한 항을 실행하기에 적합한 컴퓨터 시스템.

도면

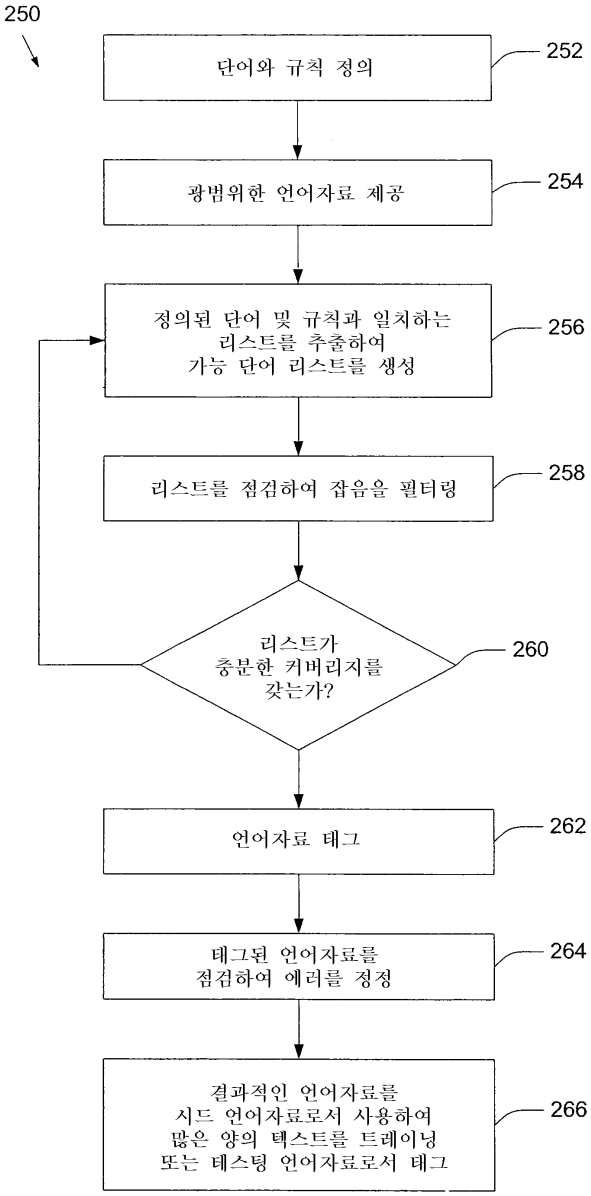
도면1


100
110
시스템 메모리
(ROM) 131
BIOS 133
(RAM) 132
오퍼레이팅 시스템 134
어플리케이션 프로그램 135
다른 프로그램 모듈 136
프로그램 데이터 137
130
120
프로세싱 유닛
121
비디오 인터페이스
190
195
출력 주변 인터페이스
비분리형 비휘발성 메모리 인터페이스
140
분리형 비휘발성 메모리 인터페이스
150
사용자 입력 인터페이스
160
네트워크 인터페이스
170
141
151
152
155
156
오퍼레이팅 시스템 144
어플리케이션 프로그램 145
다른 프로그램 모듈 146
프로그램 데이터 147
마이크로폰 163
포인팅 디바이스 161
키보드 162
모뎀 172
173
원거리 통신망
원격 컴퓨터 180
원격 애플리케이션 프로그램 185
171
근거리 통신망
스피커 197
프린터 196
모니터 191

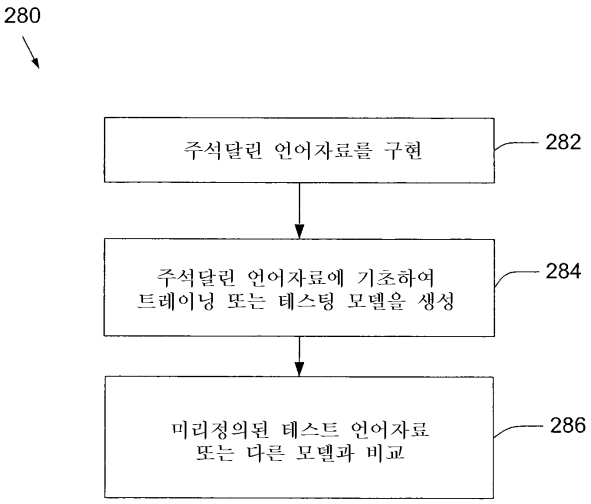
도면2


입력 202
언어 처리 시스템 200
언어 모델
206
출력 204

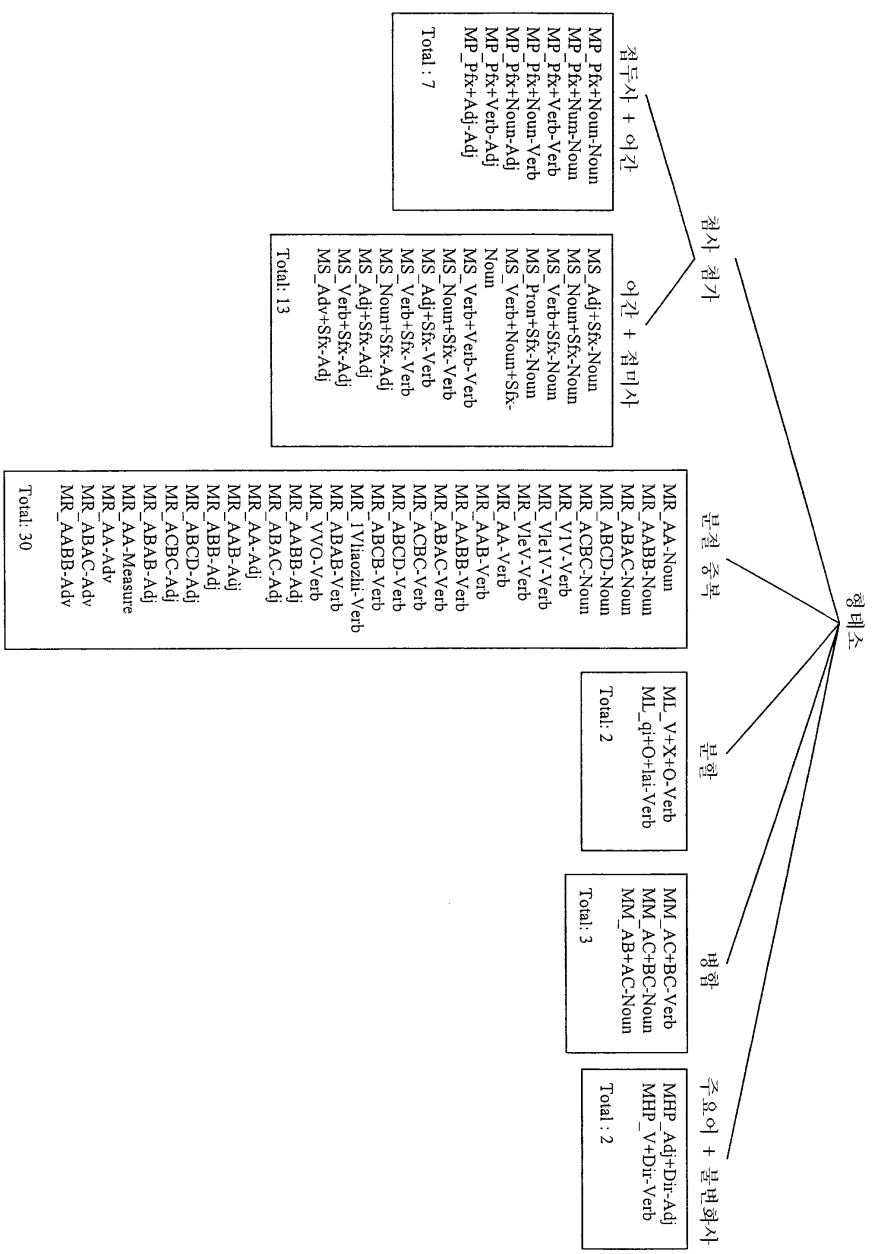
도면3


250
단어와 규칙 정의
252
광범위한 언어자료 제공
254
정의된 단어 및 규칙과 일치하는
리스트를 추출하여
가능 단어 리스트를 생성
256
리스트를 점검하여 잡음을 필터링
258
리스트가
충분한 커버리지를
갖는가?
260
언어자료 태그
262
태그된 언어자료를
점검하여 에러를 정정
264
결과적인 언어자료를
시드 언어자료로서 사용하여
많은 양의 텍스트를 트레이닝
또는 테스팅 언어자료로서 태그
266

도면4


280
주석달린 언어자료를 구현
282
주석달린 언어자료에 기초하여
트레이닝 또는 테스팅 모델을 생성
284
미리정의된 테스트 언어자료
또는 다른 모델과 비교
286

도면5


형태소
첨사 첨가
접두사 + 어간
MP_Pfx+Noun-Noun
MP_Pfx+Num-Noun
MP_Pfx+Verb-Verb
MP_Pfx+Noun-Verb
MP_Pfx+Noun-Adj
MP_Pfx+Verb-Adj
MP_Pfx+Adj-Adj
Total : 7
어간 + 접미사
MS_Adj+Sfx-Noun
MS_Noun+Sfx-Noun
MS_Verb+Sfx-Noun
MS_Pron+Sfx-Noun
MS_Verb+Noun+Sfx-Noun
MS_Verb+Verb-Verb
MS_Noun+Sfx-Verb
MS_Adj+Sfx-Verb
MS_Verb+Sfx-Verb
MS_Noun+Sfx-Adj
MS_Adj+Sfx-Adj
MS_Verb+Sfx-Adj
MS_Adv+Sfx-Adj
Total: 13
분절 중복
MR_AA-Noun
MR_AABB-Noun
MR_ABAC-Noun
MR_ABCD-Noun
MR_ACBC-Noun
MR_V1V-Verb
MR_V1eV-Verb
MR_V1eV-Verb
MR_AA-Verb
MR_AAB-Verb
MR_AABB-Verb
MR_ABAC-Verb
MR_ACBC-Verb
MR_ABCD-Verb
MR_ABCB-Verb
MR_1Vliaozhi-Verb
MR_ABAB-Verb
MR_VVO-Verb
MR_AABB-Adj
MR_ABAC-Adj
MR_AA-Adj
MR_AAB-Aoj
MR_ABB-Adj
MR_ABCD-Adj
MR_ACBC-Adj
MR_ABAB-Adj
MR_AA-Measure
MR_AA-Adv
MR_ABAC-Adv
MR_AABB-Adv
Total: 30
분할
ML_V+X+O-Verb
ML_qi+O+lai-Verb
Total: 2
병합
MM_AC+BC-Verb
MM_AC+BC-Noun
MM_AB+AC-Noun
Total: 3
주요어 + 불변화사
MHP_Adj+Dir-Adj
MHP_V+Dir-Verb
Total : 2