(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년12월06일
(11) 등록번호 10-1683324
(24) 등록일자 2016년11월30일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) *G06F 17/27* (2006.01)
G06F 17/28 (2006.01)
(21) 출원번호 10-2011-7027693
(22) 출원일자(국제) 2010년05월14일
심사청구일자 2015년04월16일
(85) 번역문제출일자 2011년11월21일
(65) 공개번호 10-2012-0026063
(43) 공개일자 2012년03월16일
(86) 국제출원번호 PCT/US2010/035033
(87) 국제공개번호 WO 2010/135204
국제공개일자 2010년11월25일
(30) 우선권주장
12/470,492 2009년05월22일 미국(US)
(56) 선행기술조사문헌
KR1020040013097 A*
US20030204400 A1*
US20040102957 A1
US20020198701 A1
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
마이크로소프트 테크놀로지 라이센싱, 엘엘씨
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
돌란 윌리엄 비
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
브록켓 크리스토퍼 제이
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션
(뒷면에 계속)
(74) 대리인
제일특허법인

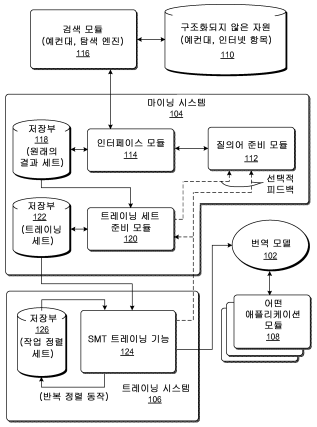
전체 청구항 수 : 총 15 항 심사관 : 경연정

(54) 발명의 명칭 **구조화되지 않은 자원으로부터의 문구 쌍의 마이닝**

(57) 요 약

마이닝 시스템은 질의어를 적용하여 구조화되지 않은 자원으로부터 결과 항목을 검색한다. 구조화되지 않은 자원은 네트워크 액세스 가능한 자원 항목의 저장소에 대응할 수 있다. 검색되는 결과 항목은 자원 항목과 관련된 텍스트 세그먼트(예컨대, 문장 단편)에 대응할 수 있다. 마이닝 시스템은 결과 항목을 필터링하여 결과 항목의 각각의 쌍을 확립함으로써 구조화된 트레이닝 세트를 생성한다. 트레이닝 시스템은 트레이닝 세트를 이용하여 통계적 번역 모델을 생성할 수 있다. 번역 모델은 단일 언어로 의미론적 관련 문구 사이에서 번역하기 위해 하나의 언어를 사용하는 콘텍스트에서 이용될 수 있다. 번역 모델은 또한 2개의 각각의 언어로 표현되는 문구 사이에서 번역하기 위해 2개의 언어를 사용하는 콘텍스트에서 이용될 수 있다. 번역 모델의 여러 적용이 또한 설명된다.

대 표 도



(72) 발명자

캐스틸로 줄리오 제이

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

반더벤드 루크레티아 에이치

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이 엘씨에이 - 인터내셔널 페이턴츠 마이크로소프트 코포레이션

명 세 서

청구범위

청구항 1

처리 장치에 의해 수행되는 방법으로서,

자연어로 표현된 자연어 질의어 용어를 포함하는 자연어 질의어를 구성하는 단계와,

상기 자연어 질의어를 웹 검색 엔진에 제시하는 단계 - 상기 웹 검색 엔진은 웹 문서들의 인덱스를 유지하기 위한 웹 크롤링(crawling) 동작을 수행하고 상기 인덱스를 사용하여 상기 자연어 질의어의 상기 자연어 질의어 용어와 매칭되는 매칭 웹 문서를 식별하도록 구성됨 - 와,

상기 웹 검색 엔진으로부터 웹 검색 결과 세트를 수신하는 단계 - 상기 웹 검색 결과 세트는 상기 자연어 질의어의 상기 자연어 질의어 용어와 매칭되는 상기 매칭 웹 문서로부터 상기 웹 검색 엔진에 의해 식별된 웹 검색 결과 항목을 제공하고, 상기 웹 검색 결과 항목도 상기 자연어로 표현됨 - 와,

트레이닝 세트를 생성하기 위해 상기 웹 검색 결과 세트에 대한 처리를 수행하는 단계 - 상기 트레이닝 세트는 상기 웹 검색 결과 세트 내의 상기 웹 검색 결과 항목의 쌍을 식별함 -

를 포함하되,

상기 트레이닝 세트 내의 각각의 쌍은

상기 자연어로 표현되는 제1 복수의 단어들을 포함하는 상기 웹 검색 엔진으로부터 수신된 제1 웹 검색 결과 항목과,

상기 자연어로 표현되는 제2 복수의 단어들을 포함하는 상기 웹 검색 엔진으로부터 수신된 제2 웹 검색 결과 항목

을 포함하고,

상기 트레이닝 세트는 전기적 트레이닝 시스템이 통계적 번역 모델을 학습할 수 있는 토대를 제공하는

방법.

청구항 2

제1항에 있어서,

각각의 웹 검색 결과 항목은 연관 웹 문서의 개요 또는 요약을 포함하는

방법.

청구항 3

제1항에 있어서,

상기 자연어 질의어를 구성하는 단계는, 상기 웹 검색 엔진으로 사용자에 의해 이전에 전송된 실제 질의어 세트를 추출하는 단계를 포함하는

방법.

청구항 4

제1항에 있어서,

상기 처리를 수행하는 단계는, 적어도 하나의 고려 사항에 기초하여 상기 트레이닝 세트로부터 소정의 웹 검색 결과 항목을 배제시키는 단계를 포함하는

방법.

청구항 5

제4항에 있어서,

상기 배제시키는 단계는, 상기 웹 검색 결과 항목과 연관된 순위 스코어에 기초하여 쌍별 매칭을 위한 후보를 상기 웹 검색 결과 항목으로부터 식별하는 단계를 포함하는

방법.

청구항 6

제4항에 있어서,

상기 트레이닝 세트로부터 상기 소정의 웹 검색 결과 항목을 배제시키는 단계는,

상기 웹 검색 결과 세트의 복수의 웹 검색 결과 항목 내에서 발견된 단어의 공통성에 기초하여 각각의 웹 검색 결과 세트에 대한 어휘 특징(lexical signature)을 결정하는 단계와,

상기 소정의 웹 검색 결과 항목이 상기 어휘 특징과 미리 정해진 양만큼 상이하다는 것을 판정하는 단계

를 포함하는

방법.

청구항 7

제1항에 있어서,

상기 제1 웹 검색 결과 항목과 상기 제2 웹 검색 결과 항목은 상기 제1 웹 검색 결과 항목의 상기 제2 웹 검색 결과 항목에 대한 유사도를 반영하는 유사도 스코어에 기초하여 상기 트레이닝 세트 내의 상기 각각의 쌍으로서 지정되는

방법.

청구항 8

제1항에 있어서,

상기 웹 검색 결과 세트에 대해 k-최근접 이웃 클러스터링 기법(k-nearest neighbor clustering technique)을 수행하여 웹 검색 결과 항목의 복수의 상이한 클러스터들을 식별하는 단계와,

각각의 클러스터 내의 각각의 클러스터화된 결과 항목이 쌍을 이루게 하고 상기 트레이닝 세트 내에 상기 쌍을 이룬 클러스터화된 결과 항목을 포함시키는 단계

를 더 포함하는 방법.

청구항 9

제1항에 있어서,

상기 자연어로 된 제1 문구를 구문변경(paraphrase)하여 상기 자연어로 된 제2 문구를 획득하기 위해, 상기 트

레이닝 세트를 사용하여 상기 통계적 번역 모델을 트레이닝시키는 단계

를 더 포함하는 방법.

청구항 10

제9항에 있어서,

종료 조건이 충족될 때까지 상기 트레이닝 세트 내의 문구의 임시적 쌍별 정렬(tentative pairwise alignments)을 획득하기 위해 상기 통계적 번역 모델을 반복적으로 적용하는 단계

를 더 포함하는 방법.

청구항 11

제10항에 있어서,

상기 종료 조건이 충족되면, 상기 통계적 번역 모델을 정의하는 통계적 매개 변수를 생성하는 단계

를 더 포함하는 방법.

청구항 12

제1항에 있어서,

웹-기반 백과사전적 기준 소스로부터 상기 자연어 질의어 용어를 추출하는 단계

를 더 포함하는 방법.

청구항 13

제1항에 있어서,

상기 트레이닝 세트에 기초하여 상기 통계적 번역 모델을 생성하고 상기 통계적 번역 모델을 적용하는 단계

를 더 포함하되,

상기 적용하는 단계는,

상기 통계적 번역 모델을 이용하여 탐색 질의어를 확장하는 단계와,

상기 통계적 번역 모델을 이용하여 문서 인덱스 결정을 가능하게 하는 단계와,

상기 통계적 번역 모델을 이용하여 텍스트 콘텐츠를 수정하는 단계와,

상기 통계적 번역 모델을 이용하여, 새로운 트리거 키워드를 생성함으로써 트리거 키워드를 포함하는 광고 정보를 확장시키는 단계

를 포함하는

방법.

청구항 14

시스템으로서,

처리 장치와,

명령어를 저장하는 컴퓨터 판독가능 매체

를 포함하되,

상기 명령어는 상기 처리 장치에 의해 실행될 경우 상기 처리 장치로 하여금,

복수의 자연어 질의어 용어를 포함하는 자연어 질의어를 구성하게 하고,

네트워크를 통해 웹 검색 엔진으로 상기 자연어 질의어를 전송하게 하고 - 상기 웹 검색 엔진은 웹 문서들의 인덱스를 유지하기 위한 웹 크롤링(crawling) 동작을 수행하고 상기 인덱스를 사용하여 상기 자연어 질의어와 매칭되는 매칭 웹 문서를 식별하도록 구성됨 -,

상기 웹 검색 엔진으로부터 웹 검색 결과 세트를 수신하게 하고 - 상기 웹 검색 결과 세트는 상기 자연어 질의어와 매칭되는 상기 웹 검색 엔진에 의해 식별된 상기 매칭 웹 문서의 개요 또는 요약을 포함하는 매칭 웹 문서 발췌를 제공함 -,

상기 매칭 웹 문서 발췌를 처리하여 트레이닝 세트를 생성하게 하되 - 상기 트레이닝 세트는 상기 매칭 웹 문서 발췌의 쌍을 식별하고, 매칭 웹 문서 발췌의 각각의 쌍은, 상기 웹 검색 엔진에 의해 식별된 제1 매칭 웹 문서의 제1 개요 또는 제1 요약을 포함하는 제1 매칭 웹 문서 발췌와, 상기 웹 검색 엔진에 의해 식별된 제2 매칭 웹 문서의 제2 개요 또는 제2 요약을 포함하는 제2 매칭 웹 문서 발췌를 포함함 -,

상기 트레이닝 세트는 전기적 트레이닝 시스템이 통계적 번역 모델을 학습할 수 있는 토대를 제공하는

시스템.

청구항 15

실행가능한 명령어를 포함하는 하나 이상의 컴퓨터-판독가능 저장 장치로서,

상기 명령어는 하나 이상의 처리 장치에 의해 실행될 경우 상기 하나 이상의 처리 장치로 하여금,

웹 검색 엔진으로 자연어 질의어를 제시하는 동작 - 상기 웹 검색 엔진은 웹 문서들의 인덱스를 유지하고 상기 인덱스를 사용하여 상기 자연어 질의어와 매칭되는 매칭 웹 문서를 포함하는 웹 결과 세트를 식별하도록 구성됨 - 과,

상기 검색 엔진으로부터 상기 웹 결과 세트를 수신하는 동작 - 상기 웹 결과 세트는 상기 검색 엔진에 의해 식별된 상기 매칭 웹 문서를 제공함 - 과,

상기 웹 결과 세트에 대한 처리를 수행하여 트레이닝 세트를 생성하는 동작 - 상기 트레이닝 세트는 상기 매칭 웹 문서의 쌍을 식별함 - 과,

상기 웹 검색 엔진에 제시되는 각각의 자연어 질의어에 의해 검색되는 제1 매칭 웹 문서와 제2 매칭 웹 문서를 포함하는 웹 결과 항목의 적어도 각각의 쌍에 대하여, 상기 제1 매칭 웹 문서 내의 제1 문구를 상기 제2 매칭 웹 문서 내의 제2 문구와 정렬시키는 동작 -상기 제1 문구와 상기 제2 문구 중 어느 것도 상기 제1 매칭 웹 문서와 상기 제2 매칭 웹 문서를 검색하기 위한 상기 웹 검색 엔진에 제시된 상기 각각의 자연어 질의어로부터의 자연어 질의어 용어를 포함하지 않음 -

을 수행하게 하는

컴퓨터-판독가능 저장 장치.

발명의 설명

배 경 기 술

[0001] 최근에 통계적 기계 번역 기술에 대한 관심이 상당히 있다. 이러한 기술은 먼저 트레이닝 세트(training set)를 확립하여 동작한다. 전통적으로, 트레이닝 세트는 제 1 언어에서의 텍스트의 본문(body) 및 제 2 언어에서의 텍스트의 대응하는 본문과 같은 텍스트의 병렬 코퍼스(parallel corpus)를 제공한다. 트레이닝 모듈은 텍스트의 제 1 본문이 대체적으로 텍스트의 제 2 본문에 맵핑하는 방식을 결정하기 위해 통계적 기술을 이용한다. 이러한 분석은 번역 모델을 생성시킨다. 디코딩 단계에서, 번역 모델은 제 1 언어에서의 텍스트의 사례(instances)를

제 2 언어에서의 텍스트의 대응하는 사례에 맵핑하는데 이용될 수 있다.

[0002] 통계적 번역 모델의 유효성은 종종 번역 모델을 생성하는데 이용되는 트레이닝 세트의 강건성(robustness)에 의존한다. 그러나, 고 품질의 트레이닝 세트를 제공하는 것이 어려운 작업이다. 부분적으로, 이것은 트레이닝 모듈이 전형적으로 상당량의 트레이닝 데이터를 필요로 하기 때문이며, 게다가 이와 같은 정보를 공급하기 위한 사전 확립된 병렬 코퍼스 타입의 자원의 부족 문제가 있다. 전통적인 경우에, 트레이닝 세트는 예컨대 인간 번역기의 사용을 통해 병렬 텍스트를 수동으로 생성함으로써 획득될 수 있다. 그러나, 이들 텍스트의 수동 생성은 엄청나게 시간이 소비되는 작업이다.

[0003] 더욱 자동화된 방식으로 병렬 텍스트를 식별하는 많은 기술이 존재한다. 예컨대, 웹 사이트가 다수의 서로 다른 언어로 동일한 정보를 전달하는 경우를 고려하며, 이 정보의 각 버전은 별도의 네트워크 주소(예컨대, 별도의 URL)와 관련된다. 한 기술에서, 검색 모듈은, 예컨대 URL 내의 특성 정보에 기초하여 이들 병렬 문서를 식별하려는 시도에서 검색 인덱스를 검사할 수 있다. 그러나, 이러한 기술은 비교적 제한된 수의 병렬 텍스트로의 액세스를 제공할 수 있다. 더욱이, 이러한 접근법은 많은 경우에 사실이 아닐 수 있는 가정에 의존할 수 있다.

[0004] 상기 예들은 2개의 서로 다른 자연 언어 사이로 텍스트를 변환하는 모델의 콘텍스트에서 맞추어졌다. 하나의 언어를 사용하는(monolingual) 모델이 또한 제안되었다. 이와 같은 모델은 입력 텍스트와 동일한 언어로 출력 텍스트를 생성하도록 입력 텍스트를 바꿔쓰기를 시도한다. 일 응용에서, 예컨대, 이러한 타입의 모델은, 예컨대 탐색 질의어를 표현하는 부가적인 방식을 식별하여 사용자의 탐색 질의어(search query)를 수정하는데 이용될 수 있다.

[0005] 하나의 언어를 사용하는 모델은 상술한 바와 동일한 결점을 갖는다. 실제로, 그것은 특히 동일한 언어 내에서 기존의 병렬 코퍼스를 찾기가 힘들어질 수 있다. 즉, 2개의 언어를 사용하는(bilingual) 콘텍스트에서는, 서로 다른 판독기의 모국어를 수용하도록 서로 다른 언어의 병렬 텍스트를 생성하기 위한 기존의 필요성이 존재한다. 동일한 언어의 텍스트의 병렬 버전을 생성하기 위한 필요성은 더욱 제한된다.

[0006] 그럼에도 불구하고, 이와 같은 하나의 언어를 사용하는 정보는 소량으로 존재한다. 예컨대, 통상의 시소러스(thesaurus)는 유사한 의미를 가진 동일한 언어의 단어에 관한 정보를 제공한다. 다른 경우에, 일부 책은 서로 다른 번역기에 의해 동일한 언어로 번역되었다. 서로 다른 번역은 하나의 언어를 사용하는 병렬 코퍼스 역할을 할 수 있다. 그러나, 이러한 타입의 병렬 정보는 너무 전문화되어 보다 일반적인 콘텍스트에 효율적으로 이용될 수 없다. 더욱이, 상술한 바와 같이, 이러한 타입의 정보의 비교적 소량만이 존재한다.

[0007] 또한, 동일한 주제에 관한 하나의 언어를 사용하는 문서의 본문을 자동으로 식별하여, 병렬 문장의 존재에 대한 이들 문서를 마이닝하기 위한 시도가 행해졌다. 그러나, 일부 경우에, 이들 접근법은 이들의 효율성 및 일반성을 제한할 수 있는 콘텍스트 특정 가정(context specific assumption)에 의존하였다. 이들 곤란 이외에, 텍스트는 다양한 방식으로 바꾸어질 수 있으며; 따라서, 하나의 언어를 사용하는 콘텍스트에서 병렬을 식별하는 것은 잠재적으로 2개의 언어를 사용하는 콘텍스트 내의 관련된 텍스트를 식별하는 것보다 더 복잡한 작업이다.

발명의 내용

과제의 해결 수단

[0008] 구조화되지 않은 자원으로부터 구조화된 트레이닝 세트를 추려내는 마이닝 시스템이 여기에서 설명된다. 즉, 구조화되지 않은 자원에는 잠재적으로 반복 콘텐츠 및 교대 형 콘텐츠(alternation-type content)가 풍부할 수 있다. 반복 콘텐츠는 구조화되지 않은 자원이 텍스트의 동일한 사례의 많은 반복을 포함한다는 것을 의미한다. 교대 형 콘텐츠는 구조화되지 않은 자원이 형식면에서 상이하지만, 유사한 의미론적(semantic) 콘텐츠를 표현하는 텍스트의 많은 사례를 포함한다는 것을 의미한다. 마이닝 시스템은 구조화되지 않은 자원의 이들 특성을 노출시켜 추출하며, 이 프로세스를 통해, 번역 모델을 트레이닝할 시에 이용하기 위해 구조화되지 않은 원시(raw) 콘텐츠를 구조화된 콘텐츠로 변환한다. 하나의 경우에, 구조화되지 않은 자원은 네트워크 액세스 가능한 자원 항목(예컨대, 인터넷 액세스 가능한 자원 항목)의 저장소에 대응할 수 있다.

[0009] 일 예시적인 구현에 따르면, 마이닝 시스템은 질의어를 검색 모듈에 제시하여 동작한다. 검색 모듈은 질의어를 이용하여 구조화되지 않은 자원 내에서 탐색을 행하여 결과 항목을 제공한다. 결과 항목은 구조화되지 않은 자원 내에 제공되는 관련된 자원 항목을 요약하는 텍스트 세그먼트에 대응할 수 있다. 마이닝 시스템은 결과 항목

을 필터링하여 결과 항목의 각각의 쌍을 식별함으로써 구조화된 트레이닝 세트를 생성한다. 트레이닝 시스템은 트레이닝 세트를 이용하여 통계적 번역 모델을 생성할 수 있다.

[0010] 일 예시적인 양태에 따르면, 마이닝 시스템은 동일한 주제를 처리하는 자원 항목의 그룹을 사전 식별하지 않고 질의어의 제시에만 기초하여 결과 항목을 식별할 수 있다. 환언하면, 마이닝 시스템은 대체로 자원 항목(예컨대, 문서)의 주제에 관한 불가지론 접근법(agnostic approach)을 취할 수 있으며, 마이닝 시스템은 하위 문서 조각 레벨(sub-document snippet level)에서 구조화되지 않은 자원 내의 구조를 노출시킨다.

[0011] 다른 예시적인 양태에 따르면, 트레이닝 세트는 문장 단편(fragment)에 대응하는 항목을 포함할 수 있다. 환언하면, 트레이닝 시스템은 (트레이닝 시스템이 또한 문장 전체를 포함하는 트레이닝 세트를 성공적으로 처리할 수 있을지라도) 문장 레벨 병렬 처리의 식별 및 개발(identification and exploitation of sentence-level parallelism)에 의존하지 않는다.

[0012] 다른 예시적인 양태에 따르면, 번역 모델은 단일 언어로 입력 문구를 출력 문구로 변환하도록 하나의 언어를 사용하는 콘텍스트에서 이용될 수 있으며, 여기서, 입력 문구 및 출력 문구는 유사한 의미론적 콘텐츠를 갖지만, 서로 다른 표현 형식을 갖는다. 환언하면, 번역 모델은 입력 문구의 바꿔쓴 버전(paraphrased version)을 제공하는데 이용될 수 있다. 번역 모델은 또한 제 1 언어의 입력 문구를 제 2 언어의 출력 문구로 번역하도록 2개의 언어를 사용하는 콘텍스트에서 이용될 수 있다.

[0013] 다른 예시적인 양태에 따르면, 번역 모델의 여러 응용이 설명된다.

[0014] 상기 접근법은 여러 타입의 시스템, 구성 요소, 방법, 컴퓨터 판독 가능한 매체, 데이터 구조, 제조물 등에 명시될 수 있다.

[0015] 이러한 요약은 단순한 형식으로 개념의 선택을 도입하기 위해 제공되고; 이들 개념은 상세한 설명에서 아래에 더 설명된다. 이러한 요약은 청구된 주 문제의 기본 특징 또는 필수 특징을 식별하기 위해 의도되지 않고, 청구된 주 문제의 범주를 제한하는데 이용되는 것으로도 의도되지 않는다.

도면의 간단한 설명

[0016] 도 1은 통계적 기계 번역 모델을 생성하여 적용하는 예시적인 시스템을 도시한 것이다.

도 2는 네트워크 관련 환경 내에서 도 1의 시스템의 구현을 도시한 것이다.

도 3은 하나의 결과 세트 내의 일련의 결과 항목의 일례를 도시한 것이다. 도 1의 시스템은 검색 모듈로의 질의어의 제시에 응답하여 결과 세트를 복귀시킨다.

도 4는 도 1의 시스템이 결과 세트 내에 결과 항목의 쌍을 어떻게 확립할 수 있는지를 증명하는 일례를 도시한 것이다.

도 5는 도 1의 시스템이 서로 다른 결과 세트에 대해 수행되는 분석에 기초하여 트레이닝 세트를 어떻게 생성할 수 있는지를 증명하는 일례를 도시한 것이다.

도 6은 도 1의 시스템의 동작의 개요를 제시하는 예시적인 절차를 도시한 것이다.

도 7은 도 6의 절차 내에서 트레이닝 세트를 확립하는 예시적인 절차를 도시한 것이다.

도 8은 도 1의 시스템을 이용하여 생성된 번역 모델을 적용하는 예시적인 절차를 도시한 것이다.

도 9는 상술한 도면에 도시된 특징의 어떤 양태를 구현하는데 이용될 수 있는 예시적인 처리 기능을 도시한 것이다.

개시 및 도면에서 동일한 번호는 동일한 구성 요소 및 특징을 참조하는데 이용된다. 100 계열의 번호는 원래 도 1에서 발견되는 특징을 나타내고, 200 계열의 번호는 원래 도 2에서 발견되는 특징을 나타내며, 300 계열의 번호는 원래 도 3에서 발견되는 특징을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 개시는 통계적 번역 모델을 확립하는데 이용될 수 있는 트레이닝 세트를 생성하기 위한 기능을 기술한다. 본 개시는 또한 통계적 번역 모델을 생성하여 적용하기 위한 기능을 설명한다.

[0018] 본 개시물은 다음과 같이 조직화된다. 섹션 A는 상술한 바와 같이 요약된 기능을 수행하는 예시적인 시스템을

기술한다. 섹션 B는 섹션 A의 시스템의 동작을 설명하는 예시적인 방법을 기술한다. 섹션 C는 섹션 A 및 B에 기술된 특징의 어떤 양태를 구현하는데 이용될 수 있는 예시적인 처리 기능을 기술한다.

[0019] 예비적인 문제로서, 도면의 일부는 하나 이상의 구조적 구성 요소와 관련한 개념을 기술하고, 기능, 모듈, 특징, 요소 등으로 다양하게 지칭된다. 도면에 도시된 여러 구성 요소는 예컨대 소프트웨어, 하드웨어(예컨대, 이산 논리 구성 요소 등), 펌웨어 등, 또는 이들 구현의 어떤 조합에 의해 어떤 방식으로 구현될 수 있다. 하나의 경우에, 도면에서 별개의 유닛으로의 여러 구성 요소의 예시된 분리는 실제 구현에서 대응하는 별개의 구성 요소의 사용을 반영할 수 있다. 대안적으로 또는 부가적으로, 도면에 예시된 어떤 단일 구성 요소는 다수의 실제 구성 요소에 의해 구현될 수 있다. 대안적으로 또는 부가적으로, 도면에서 어떤 2 이상의 별도의 구성 요소의 묘사는 단일의 실제 구성 요소에 의해 수행되는 서로 다른 기능을 반영할 수 있다. 다음에 논의되는 도 9는 도면에 도시된 기능의 일 예시적인 구현에 관한 부가적인 상세 사항을 제공한다.

[0020] 다른 도면은 흐름도의 형식의 개념을 기술한다. 이러한 형식에서, 어떤 동작은 어떤 순서로 수행되는 별개의 블록을 구성하는 것으로 기술된다. 이와 같은 구현은 예시적이고 비제한적이다. 여기에 기술된 어떤 블록은 서로 그룹화되고, 단일 동작으로 수행될 수 있으며, 어떤 블록은 다수의 구성 요소의 블록으로 분해될 수 있고, 어떤 블록은 여기에 예시된 것과 상이하고 (블록을 수행하는 병렬 방식을 포함하는) 순서로 수행될 수 있다. 흐름도에 도시된 블록은 소프트웨어, 하드웨어(예컨대, 이산 논리 구성 요소 등), 펌웨어, 수동 처리 등, 또는 이들 구현의 어떤 조합에 의해 구현될 수 있다.

[0021] 용어에 관해, 문구 "하도록 구성된(configured to)"는 어떤 종류의 기능이 식별된 동작을 수행하도록 구성될 수 있는 어떤 방식을 포함한다. 이 기능은 예컨대, 소프트웨어, 하드웨어(예컨대, 이산 논리 구성 요소 등), 펌웨어 등, 또는 이의 어떤 조합을 이용하여 동작을 수행하도록 구성될 수 있다.

[0022] 용어 "논리"는 태스크를 수행하기 위한 어떤 기능을 포함한다. 예컨대, 흐름도에 예시된 각 동작은 이 동작을 수행하기 위한 논리에 대응한다. 동작은 예컨대, 소프트웨어, 하드웨어(예컨대, 이산 논리 구성 요소 등), 펌웨어 등, 또는 이의 어떤 조합을 이용하여 수행될 수 있다.

[0023] A. 예시적인 시스템

[0024] 도 1은 번역 모델(102)을 생성하여 적용하는 예시적인 시스템(100)을 도시한 것이다. 번역 모델(102)은 입력 문구를 출력 문구에 맵핑하기 위한 통계적 기계 번역(SMT) 모델에 대응하며, 여기서 "문구(phrase)"는 어떤 하나 이상의 텍스트 문자열(text strings)을 나타낸다. 번역 모델(102)은 규칙 기반 접근법보다는 통계적 기술을 이용하여 이러한 동작을 수행한다. 그러나, 다른 구현에서, 번역 모델(102)은 규칙 기반 접근법의 하나 이상의 특징을 포함함으로써 통계적 분석을 보충할 수 있다.

[0025] 하나의 경우에, 번역 모델(102)은 하나의 언어를 사용하는 콘텍스트에서 동작한다. 여기서, 번역 모델(102)은 입력 문구와 동일한 언어로 표현되는 출력 문구를 생성한다. 환언하면, 출력 문구는 입력 문구의 바꿔쓴 버전으로 고려될 수 있다. 다른 경우에, 번역 모델(102)은 2개의 언어를 사용하는(또는 다수의 언어를 사용하는) 콘텍스트에서 동작한다. 여기서, 번역 모델(102)은 입력 문구와 비교해 다른 언어로 출력 문구를 생성한다. 또 다른 경우에, 번역 모델(102)은 음역(transliteration) 콘텍스트에서 동작한다. 여기서, 번역 모델은 입력 문구와 동일한 언어로 출력 문구를 생성하지만, 출력 문구는 입력 문구와 비교해 다른 기록 형식으로 표현된다. 번역 모델(102)은 또 다른 번역 시나리오에 적용될 수 있다. 이와 같은 모든 콘텍스트에서, 단어 "번역(translation)"은 광범위하게 해석될 수 있으며, 한 상태에서 다른 상태까지의 텍스트 정보의 모든 종류의 대화를 나타낸다.

[0026] 시스템(100)은 3개의 주요 구성 요소인 마이닝 시스템(104), 트레이닝 시스템(106), 및 애플리케이션 모듈(108)을 포함한다. 개요로서, 마이닝 시스템(104)은 번역 모델(102)을 트레이닝할 시에 이용하기 위한 트레이닝 세트를 생성한다. 트레이닝 시스템(106)은 트레이닝 세트에 기초하여 번역 모델(102)을 유도하도록 반복 접근법을 적용한다. 애플리케이션 모듈(108)은 특정 사용 관련 시나리오에서 입력 문구를 출력 문구에 맵핑하도록 번역 모델(102)을 적용한다.

[0027] 하나의 경우에, 단일 엔티티 또는 다수의 엔티티의 어떤 조합에 의해 관리되는 바와 같이, 단일 시스템은 도 1에 도시된 모든 구성 요소를 구현할 수 있다. 다른 경우에는, 다시, 단일 엔티티 또는 다수의 엔티티의 어떤 조합에 의해 관리되는 바와 같이, 어떤 2 이상의 별도의 시스템이 도 1에 도시된 어떤 2 이상의 구성 요소를 구현할 수 있다. 어느 하나의 경우에, 도 1에 도시된 모든 구성 요소는 단일 사이트에 위치될 수 있거나, 다수의 각각의 사이트를 통해 분산될 수 있다. 다음의 설명은 도 1에 도시된 모든 구성 요소에 관한 부가적인 상세 사항을 제공한다.

[0028] 마이닝 시스템(104)부터 시작하면, 이러한 구성 요소는 구조화되지 않은 자원(110)으로부터 결과 항목을 검색하여 동작한다. 구조화되지 않은 자원(110)은 자원 항목의 어떤 국부화 또는 분산된 소스를 나타낸다. 자원 항목은 결과적으로 텍스트 정보의 어떤 유닛에 대응할 수 있다. 예컨대, 구조화되지 않은 자원(110)은 인터넷과 같은 원거리 통신망에 의해 제공되는 자원 항목의 분산된 저장소를 나타낸다. 여기서, 자원 항목은 어떤 타입의 네트워크 액세스 가능한 페이지 및/또는 관련된 문서에 대응할 수 있다.

[0029] 구조화되지 않은 자원(110)은 사전에 병렬 코퍼스의 방식으로 배열되지 않기 때문에 구조화되지 않은 것으로 고려된다. 환언하면, 구조화되지 않은 자원(110)은 어떤 오버아칭(overarching) 기법에 따라 자원 항목을 서로 관련시키지 않는다. 그럼에도 불구하고, 구조화되지 않은 자원(110)에는 잠재적으로 반복 콘텐츠 및 교대 형 콘텐츠가 풍부할 수 있다. 반복 콘텐츠는 구조화되지 않은 자원(110)이 텍스트의 동일한 사례의 많은 반복을 포함한다는 것을 의미한다. 교대 형 콘텐츠는 구조화되지 않은 자원(110)이 형식면에서 상이하지만, 유사한 의미론적 콘텐츠를 표현하는 텍스트의 많은 사례를 포함한다는 것을 의미한다. 이것은 트레이닝 세트를 구성할 시에 이용하기 위해 마이닝될 수 있는 구조화되지 않은 자원(110)의 기본 특징이 있다는 것을 의미한다.

[0030] 마이닝 시스템(104)의 하나의 목적은 구조화되지 않은 자원(110)의 상술한 특성을 노출시켜, 이 프로세스를 통해, 번역 모델(102)을 트레이닝할 시에 이용하기 위해 구조화되지 않은 원시 콘텐츠를 구조화된 콘텐츠로 변환하는 것이다. 마이닝 시스템(104)은 부분적으로 질의어 준비 모듈(112) 및 인터페이스 모듈(114)을 검색 모듈(116)과 함께 이용하여 이러한 목적을 달성한다. 질의어 준비 모듈(112)은 질의어의 그룹을 공식화한다. 각 질의어는 타겟 대상(target subject)으로 향한 하나 이상의 질의어(query terms)를 포함할 수 있다. 인터페이스 모듈(114)은 질의어를 검색 모듈(116)에 제시한다. 검색 모듈(116)은 질의어를 이용하여 구조화되지 않은 자원(110) 내에서 탐색을 수행한다. 이러한 탐색에 응답하여, 검색 모듈(116)은 서로 다른 각각의 질의어에 대한 다수의 결과 세트를 복귀시킨다. 각 결과 세트는 결과적으로 하나 이상의 결과 항목을 포함한다. 결과 항목은 구조화되지 않은 자원(110) 내의 각각의 자원 항목을 식별한다.

[0031] 하나의 경우에, 마이닝 시스템(104) 및 검색 모듈(116)은 동일한 시스템에 의해 구현되고, 동일한 엔티티 또는 서로 다른 각각의 엔티티에 의해 관리된다. 다른 경우에는, 마이닝 시스템(104) 및 검색 모듈(116)은 2개의 각각의 시스템에 의해 구현되고, 다시, 동일한 엔티티 또는 서로 다른 각각의 엔티티에 의해 관리된다. 예컨대, 일 구현에서, 검색 모듈(116)은 워싱턴 레드몬드 소재의 마이크로소프트사에 의해 제공되는 Live Search 엔진과 같지만, 이에 제한되지 않는 탐색 엔진을 나타낸다. 사용자는 탐색 엔진(예컨대, API 등)에 의해 제공되는 인터페이스와 같은 어떤 메카니즘을 통해 탐색 엔진으로의 액세스를 제공할 수 있다. 탐색 엔진은 어떤 탐색 전략 및 순위 전략(ranking strategy)을 이용하여 제시된 질의어에 응답하여 결과 세트를 식별하여 공식화할 수 있다.

[0032] 하나의 경우에, 결과 세트 내의 결과 항목은 각각의 텍스트 세그먼트에 대응한다. 서로 다른 탐색 엔진은 질의어의 제시에 응답하여 텍스트 세그먼트를 공식화할 시에 서로 다른 전략을 이용할 수 있다. 많은 경우에, 텍스트 세그먼트는 제시된 질의어 뿐만 아니라 자원 항목의 관련성을 전달하는 자원 항목의 대표 부분(예컨대, 발췌 부분)을 제공한다. 설명을 위해, 텍스트 세그먼트는 이들의 관련된 완전한 자원 항목의 간략한 초록 또는 요약으로 고려될 수 있다. 특히, 하나의 경우에, 텍스트 세그먼트는 기본적 전체 자원 항목으로부터 취해지는 하나 이상의 문장에 대응할 수 있다. 한 시나리오에서, 인터페이스 모듈(114) 및 검색 모듈(116)은 문장 단편을 포함하는 자원 항목을 공식화할 수 있다. 다른 시나리오에서, 인터페이스 모듈(114) 및 검색 모듈(116)은 전체 문장(또는 전체 단락 등과 같은 텍스트의 보다 큰 단위)을 포함하는 자원 항목을 공식화할 수 있다. 인터페이스 모듈(114)은 결과 세트를 저장부(118) 내에 저장한다.

[0033] 트레이닝 세트 준비 모듈(120)(간결함을 위한 "준비 모듈")은 결과 세트 내의 원시 데이터를 처리하여 트레이닝 세트를 생성한다. 이러한 동작은 개별적으로 또는 함께 수행될 수 있는 2개의 구성 요소의 동작, 즉 필터링 및 매칭(matching)을 포함한다. 필터링 동작에 관해, 준비 모듈(120)은 하나 이상의 제한된 고려 사항(constraining consideration)에 기초하여 결과 항목의 원래의 세트를 필터링한다. 이러한 처리의 목적은 상대(pairwise) 매칭을 위한 적절한 후보자(candidates)인 결과 항목의 서브세트를 식별하여, 결과 세트로부터 "노이즈"를 제거하기 위한 것이다. 필터링 동작은 필터링된 결과 세트를 생성한다. 매칭 동작에 관해, 준비 모듈(120)은 필터링된 결과 세트 상에서 쌍별 매칭을 수행한다. 쌍별 매칭은 결과 세트 내의 결과 항목의 쌍을 식별한다. 준비 모듈(120)은 상기 동작에 의해 생성되는 트레이닝 세트를 저장부(122) 내에 저장한다. 준비 모듈(120)의 동작에 관한 부가적인 상세 사항은 이러한 설명의 이후 시점에서 제공될 것이다.

[0034] 트레이닝 시스템(106)은 저장부(122) 내의 트레이닝 세트를 이용하여 번역 모델(102)을 트레이닝한다. 이를 위

해, 트레이닝 시스템(106)은 구문형 SMT 기능과 같은 어떤 타입의 통계적 기계 번역(SMT) 기능(124)을 포함할 수 있다. SMT 기능(124)은 트레이닝 세트 내의 패턴을 식별하도록 통계적 기술을 이용하여 동작한다. SMT 기능(124)은 이들 패턴을 이용하여 트레이닝 세트 내의 문구의 상관 관계를 식별한다.

[0035] 특히, SMT 기능(124)은 그의 트레이닝 동작을 반복적 방식으로 수행한다. 각 단계에서, SMT 기능(124)은 트레이닝 세트 내의 문구의 쌍대 정렬에 관한 임시적 가정(tentative assumption)에 도달하도록 하는 통계적 분석을 수행한다. SMT 기능(124)은 이들 임시적 가정을 이용하여 그의 통계적 분석을 반복하여, 갱신된 임시적 가정에 도달하도록 한다. SMT 기능(124)은 종료 조건이 만족되는 것으로 생각될 때까지 이러한 반복 동작을 반복한다. 저장부(126)는 SMT 기능(124)에 의해 수행되는 처리의 과정을 통해 임시 정렬 정보의 작업(working) 세트(예컨대, 번역 테이블 등의 형식으로) 유지할 수 있다. 처리의 종료에서, SMT 기능(124)은 번역 모델(102)을 규정하는 통계적 매개 변수를 생성한다. SMT 기능(124)에 관한 부가적인 상세 사항은 이러한 설명의 이후 시점에서 제공될 것이다.

[0036] 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 입력 문구를 의미론적 관련 출력 문구로 변환한다. 상술한 바와 같이, 입력 문구 및 출력 문구는 동일한 언어 또는 서로 다른 각각의 언어로 표현될 수 있다. 애플리케이션 모듈(108)은 여러 애플리케이션 시나리오와 관련하여 이러한 변환을 수행할 수 있다. 애플리케이션 모듈(108) 및 애플리케이션 시나리오에 관한 부가적인 상세 사항은 이러한 설명의 이후 시점에서 제공될 것이다.

[0037] 도 2는 도 1의 시스템(100)의 일 대표적 구현을 도시한 것이다. 이 경우에, 컴퓨팅 기능(202)은 마이닝 시스템(104) 및 트레이닝 시스템(106)을 구현하는데 이용될 수 있다. 단일 엔티티 또는 다수의 엔티티의 어떤 조합에 의해 유지되는 바와 같이, 컴퓨팅 기능(202)은 단일 사이트에서 유지되거나 다수의 사이트를 통해 분산되는 어떤 처리 기능을 나타낼 수 있다. 한 대표적인 경우에, 컴퓨팅 기능(202)은 개인용 데스크탑 컴퓨팅 장치, 서버형 컴퓨팅 장치 등과 같은 어떤 타입의 컴퓨팅 장치에 대응한다.

[0038] 하나의 경우에, 구조화되지 않은 자원(110)은 네트워크 환경(204)에 의해 제공되는 자원 항목의 분산 저장소에 의해 구현될 수 있다. 네트워크 환경(204)은 어떤 타입의 근거리 통신망 또는 원거리 통신망에 대응할 수 있다. 예컨대, 제한 없이, 네트워크 환경(204)은 인터넷에 대응할 수 있다. 이와 같은 환경은 예컨대 네트워크 액세스 가능한 페이지 및 링크된 콘텐츠 항목에 대응하는 잠재적으로 상당수의 자원 항목으로의 액세스를 제공한다. 검색 모듈(116)은 예컨대 네트워크 크롤링(crawling) 기능을 이용하여 통상적인 방식으로 네트워크 환경(204)에서 이용 가능한 자원 항목의 인덱스를 유지할 수 있다.

[0039] 도 3은 질의어(304)의 제시에 응답하여 검색 모듈(116)에 의해 복귀될 수 있는 가상 결과 세트(302)의 부분의 일례를 도시한 것이다. 이러한 일례는 도 1의 마이닝 시스템(104)의 일부 개념적 토대를 설명하기 위한 수단(vehicle) 역할을 한다.

[0040] 질의어(304)인 "shingles zoster(대상 포진)"은 잘 알려진 질병에 대한 것이다. 질의어는 상당량의 관계없는 정보를 배제하도록 충분히 초점을 맞춘 타겟 대상 문제를 정확하게 나타내도록 선택된다. 이 예에서, "shingles"은 질병의 일반적인 이름을 나타내는 반면에, (예컨대, herpes zoster에서와 같은) "zoster"은 질병의 더욱 공식적인 이름을 나타낸다. 따라서, 질의어 용어의 이런 조합은 단어 "shingles"의 관계없는 그리고 의도되지 않은 의미에 속하는 결과 항목의 검색을 감소시킬 수 있다.

[0041] 결과 세트(302)는 R1-RN으로 표시되는 일련의 결과 항목을 포함하며; 도 3은 이들 결과 항목의 작은 샘플을 도시한다. 각 결과 항목은 대응하는 자원 항목으로부터 추출된 텍스트 세그먼트를 포함한다. 이 경우에, 텍스트 세그먼트는 문장 단편을 포함한다. 그러나, 인터페이스 모듈(114) 및 검색 모듈(116)은 또한 전체 문장(또는 전체 단락 등)을 포함하는 자원 항목을 제공하도록 구성될 수 있다.

[0042] shingles의 질병은 현저한 특성을 갖는다. 예컨대, shingles은 수두(chicken pox)를 유발시키는 동일한 바이러스(herpes zoster)의 재활성화에 의해 유발되는 질병이다. 다시 각성될 시에, 바이러스는 신체의 신경을 따라 이동하여, 외관에 불그스름한 고통스러운 발진(painful rash)에 이르고, 물집이 작은 클러스터를 특징으로 한다. 이러한 질병은 종종 면역 체계가 손상될 발생하여, 물리적 외상, 다른 질병, 스트레스 등에 의해 유발될 수 있다. 이러한 질병은 종종 중장년층(elderly) 등을 괴롭힌다.

[0043] 서로 다른 결과 항목은 질병의 현저한 특성에 초점을 맞추는 콘텐츠를 포함하는 것으로 예상될 수 있다. 결과적으로, 결과 항목은 어떤 잡아내기 쉬운 문구(telltale phrases)를 반복하는 것으로 예상될 수 있다. 예컨대, 사례(306)에 의해 나타낸 바와 같이, 수개의 결과 항목은 다양하게 표현된 바와 같이 고통스러운 발진의 발생을 언급한다. 사례(308)에 의해 나타낸 바와 같이, 수개의 결과 항목은 다양하게 표현된 바와 같이 질병이 약화된

면역 체계와 관련됨을 언급한다. 사례(310)에 의해 나타낸 바와 같이, 수개의 결과 항목은 다양하게 표현된 바와 같이 질병이 신체 내의 신경을 따라 이동하는 바이러스를 생성함을 언급한다. 이들 예는 단지 예시적이다. 다른 결과 항목은 타겟 대상과 크게 관계없을 수 있다. 예컨대, 결과 항목(312)은 건축 자재와 관련하여 용어"shingles"에 이용하여, 주제와 밀접한 관계가 없다. 그러나, 이러한 관계없는 결과 항목(12)은 다른 결과 항목과 공유되는 문구를 포함할 수 있다.

[0044] 여러 통찰들이 결과 세트(302)에 명시된 패턴으로부터 수집될 수 있다. 이들 통찰의 일부는 좁게는 타겟 대상, 즉 shingles(대상 포진)의 질병에 속한다. 예컨대, 마이닝 시스템(104)은 결과 세트(302)를 이용하여 "shingles" 및 "herpes zoster"가 동의어임을 추론할 수 있다. 다른 통찰력은 일반적으로 의료 분야에 속한다. 예컨대, 마이닝 시스템(104)은 문구 "painful rash(고통스러운 발진)"이 문구 "a rash that is painful(고통스러운 발진)"로 의미있게 대체될 수 있음을 추론할 수 있다. 더욱이, 마이닝 시스템(104)은 문구 "impaired(훼손된)"가 면역 체계(immuned system)(및 잠재적으로 다른 주제)를 논의할 때 "weakened(약화된)" 또는 "compromised(손상된)"으로 의미있게 대체될 수 있음을 추론할 수 있다. 다른 통찰력은 글로벌 또는 도메인 독립적인 범위를 가질 수 있다. 예컨대, 마이닝 시스템(104)은 문구 "moves along(따라서 움직이다)"가 "travels over(위를 여행하다)" 또는 "moves over(위를 움직이다)"으로 의미있게 대체될 수 있고, 문구 "elderly(장년층)"가 "old people(노인들)" 또는 "old folks(노인들) 또는 "senior citizens(노인 시민들)" 등으로 대체될 수 있음을 추론할 수 있다. 이들 등가성은 결과 세트(302) 내에서 의료 콘텍스트에 제시되지만, 이들은 다른 콘텍스트에 적용할 수 있다. 예컨대, 일하러 가는 여행을 도로 "travels over(위를 여행하다)" 또는 도로 "moves over(위를 움직이다)"라고 기술될 수 있다.

[0045] 도 3은 또한 트레이닝 시스템(106)이 문구 중에서 의미있는 유사도를 식별할 수 있는 한 메카니즘을 예시하는데에 유용하다. 예컨대, 결과 항목은 "rash(발진)", "elderly(장년층)", "nerves(신경)", "immune system(면역체계)" 등과 많은 동일한 단어를 반복한다. 빈번히 나타나는 이들 단어는 의미론적 관련 문구의 존재에 대한 텍스트 세그먼트를 조사하는 앵커 포인트(anchor points) 역할을 할 수 있다. 예컨대, 일반적으로 생성하는 문구 "면역 체계"와 관련된 앵커 포인트에 초점을 맞춤으로써, 트레이닝 시스템(106)은 "impaired", "weakened" 및 "compromised"이 의미론적으로 교환 가능한 단어에 대응할 수 있다는 결론을 도출할 수 있다. 트레이닝 시스템(106)은 점진적인 방식으로 이러한 조사에 접근할 수 있다. 즉, 그것은 문구의 정렬에 관한 임시적 가정을 도출할 수 있다. 이들 가정에 기초하여, 그것은 새로운 임시적 가정을 도출하도록 조사를 반복할 수 있다. 어떤 시점에서, 임시적 가정에 의해, 트레이닝 시스템(106)이 결과 항목의 관련성으로 부가적인 통찰력을 도출시킬 수 있으며; 대안적으로, 이러한 가정은 스텝 백(step back)을 나타내어, 추가적 분석을 흐리게 할 수 있다(이 경우에, 가정은 수정될 수 있다). 이러한 프로세스를 통해, 트레이닝 시스템(106)은 결과 세트 내의 문구의 관련성에 관한 가정의 안정 세트에 도달하기를 시도한다.

[0046] 일반적으로, 예는 또한 마이닝 시스템(104)이 동일한 주제를 처리하는 자원 항목의 그룹(예컨대, 기본 문서)을 사전 식별하지 않고 질의어의 제시에만 기초하여 결과 항목을 식별할 수 있다. 환언하면, 마이닝 시스템(104)은 대체로 자원 항목의 주제에 관한 불가지론 접근법을 취할 수 있다. 도 3의 예에서, 자원 항목의 대부분은 사실상 동일한 주제(질병 shingles)에 속할 것 같다. 그러나, (1) 이러한 유사도는 문서의 메타 레벨 분석보다는 질의어만에 기초하여 드러나며, (2) 자원 항목이 동일한 주제에 속한다는 요건이 없다.

[0047] 도 4로 진행하면, 이러한 도면은 (도 1의) 준비 모듈(120)이 결과 세트(R_A) 내에 결과 항목(R_{A1}-R_{AN})의 초기 쌍 이루기(initial pairing)를 확립하는데 이용될 수 있는 방식을 도시한 것이다. 여기서, 준비 모듈(120)은 각 결과 항목과 (결과 항목의 자기 동일적(self-identical) 쌍 이루기를 제외한) 결과 세트 내의 모든 다른 결과 항목 사이의 링크를 확립할 수 있다. 예컨대, 제 1 쌍은 결과 항목(R_{A1})을 결과 항목(R_{A2})과 연결한다. 제 2 쌍은 결과 항목(R_{A1})을 결과 항목(R_{A3})과 연결한다. 사실상, 준비 모듈(120)은 하나 이상의 필터링 고려 사항에 기반하는 결과 항목 사이의 관련(associations)을 제한할 수 있다. 섹션 B은 준비 모듈(120)이 결과 항목의 쌍별 매칭을 제한할 수 있는 방식에 관한 부가적인 정보를 제공할 것이다.

[0048] 반복하기 위해, 상기 방식으로 쌍을 이루게 되는 결과 항목은 문장 단편을 포함하는 각각의 자원 항목의 어떤 부분에 대응할 수 있다. 이것은 마이닝 시스템(104)이 병렬 문장을 식별하는 표현 태스크 없이 트레이닝 세트를 확립할 수 있다는 것을 의미한다. 환언하면, 트레이닝 시스템(106)은 문장 레벨 병렬 처리의 개발에 의존하지 않는다. 그러나, 트레이닝 시스템(106)은 또한 결과 항목이 전체 문장(또는 텍스트의 보다 큰 단위)을 포함하는 트레이닝 세트를 성공적으로 처리할 수 있다.

[0049] 도 5는 서로 다른 결과 세트로부터의 쌍대 매핑이 저장부(122) 내에 트레이닝 세트를 형성하도록 조합될 수 있는 방식을 예시한 것이다. 즉, 질의어(Q_A)는 결과 세트(R_A)에 이르고 나서, 쌍별 매칭된 결과 세트(TS_A)에 이른다. 질의어(Q_B)는 결과 세트(R_B)에 이르고 나서, 쌍별 매칭된 결과 세트(TS_B)에 이른다. 준비 모듈(120)은 이들 서로 다른 쌍별 매칭된 결과 세트를 조합하고 연쇄시켜 트레이닝 세트를 생성한다. 대체로, 트레이닝 세트는 추가적 조사를 위한 결과 항목 사이의 임시 정렬의 초기 세트를 확립한다. 트레이닝 시스템(106)은 실제 관련된 텍스트 세그먼트를 나타내는 정렬의 서브세트를 식별하도록 반복적 방식으로 트레이닝 세트에서 동작한다. 궁극적으로, 트레이닝 시스템(106)은 정렬 내에 나타나는 의미론적 관련 문구를 식별하고자 한다.

[0050] 이러한 섹션 내의 최종 포인트로서, 도 1에서, 시스템(100)의 서로 다른 구성 요소의 사이에 점선이 도시되어 있음에 주목한다. 이것은 어떤 구성 요소에 의해 도달되는 결론이 다른 구성 요소의 동작을 수정하는데 이용될 수 있음을 그래프로 나타낸다. 예컨대, SMT 기능(124)은 준비 모듈(120)이 결과 항목의 초기 필터링 및 쌍 이루기를 수행하는 방식과 관계가 있다는 어떤 결론에 도달할 수 있다. 준비 모듈(120)은 이러한 피드백을 수신하여, 이에 응답하여 이의 필터링 또는 매칭 행동을 수정할 수 있다. 다른 경우에, SMT 기능(124) 또는 준비 모듈(120)은, 예컨대, 반복 콘텐츠 및 교대 형 콘텐츠가 풍부한 결과 세트를 추출하는 질의어 공식화 전략의 능력과의 관계로서 어떤 질의어 공식화 전략의 유효성에 관한 결론에 도달할 수 있다. 질의어 준비 모듈(112)은 이러한 피드백을 수신하여, 이에 응답하여 이의 행동을 수정할 수 있다. 특히, 하나의 경우에, SMT 기능(124) 또는 준비 모듈(120)은 질의어의 다른 라운드(round) 내에 포함하는데 유용할 수 있는 주요 용어 또는 주요 문구를 발견하여 분석을 위한 부가적인 결과 세트에 이를 수 있다. 피드백을 위한 또 다른 기회는 시스템(100) 내에 존재할 수 있다.

[0051] B. 예시적인 프로세스

[0052] 도 6-도 8은 도 1의 시스템(100)의 동작의 한 방식을 설명하는 절차(600, 700, 800)를 도시한 것이다. 시스템(100)의 동작의 기본 원칙이 이미 섹션 A에서 소개되었으므로, 어떤 동작은 이 섹션에서 요약 형식으로 처리될 것이다.

[0053] 도 6에서 개시하면, 이 도면은 마이닝 시스템(104) 및 트레이닝 시스템(106)의 동작의 개요를 나타내는 절차(600)를 도시한 것이다. 특히, 동작의 제 1 문구는 마이닝 시스템(104)에 의해 수행되는 마이닝 동작(602)을 나타내지만, 동작의 제 2 문구는 트레이닝 시스템(106)에 의해 수행되는 트레이닝 동작(604)을 나타낸다.

[0054] 블록(606)에서, 마이닝 시스템(104)은 질의어의 세트를 구성하여 프로세스(600)를 개시한다. 마이닝 시스템(104)은 서로 다른 전략을 이용하여 이러한 태스크를 수행할 수 있다. 하나의 경우에, 마이닝 시스템(104)은, 예컨대, 질의어 로그 등으로부터 획득되는 바와 같이 이전에 사용자에 의해 탐색 엔진에 제시된 실제 질의어의 세트를 추출할 수 있다. 다른 경우에, 마이닝 시스템(104)은 어떤 기준 소스 또는 기준 소스의 조합에 기초하여 "인공적(artificial)" 질의어를 작성할 수 있다. 예컨대, 마이닝 시스템(104)은 위키피디어(Wikipedia) 등과 같은 백과사전적(encyclopedic) 기준 소스의 분류 인덱스 또는 시소러스 등으로부터 질의어 용어를 추출할 수 있다. 단지 일례를 인용하기 위해, 마이닝 시스템(104)은 기준 소스를 이용하여 서로 다른 질병 이름을 포함하는 질의어의 수집을 생성할 수 있다. 마이닝 시스템(104)은 질명 이름에 하나 이상의 다른 용어를 보충하여, 복귀되는 결과 세트에 초점을 맞추는데 도움을 줄 수 있다. 예컨대, 마이닝 시스템(104)은 "shingles AND zoster"에서와 같이 정식 의료 상당 어구(formal medical equivalent)와 각각의 공통 질병 이름을 결합할 수 있다. 또는, 마이닝 시스템(104)은 "shingles AND prevention" 등과 같은 질병 이름에 약간 직교하는(orthogonal) 다른 질의어 용어와 각 질병 이름을 결합할 수 있다.

[0055] 더욱 광범위하게 고려하면, 블록(606)에서의 질의어 선택은 서로 다른 오버아칭 목표(overarching objectives)에 의해 조절될 수 있다. 하나의 경우에, 마이닝 시스템(104)은 특정 도메인에 초점을 맞추는 질의어를 준비하기를 시도할 수 있다. 이러한 전략은 특정 도메인으로 약간 웨이트(weight)되는 문구를 표면화시킬 시에 효과적일 수 있다. 다른 경우에, 마이닝 시스템(104)은 도메인의 광범한 범위를 조사하는 질의어를 준비하기를 시도할 수 있다. 이러한 전략은 사실상 더욱 도메인에 독립적인 문구를 표면화시킬 시에 효과적일 수 있다. 하여튼, 마이닝 시스템(104)은 상술한 바와 같이 반복 콘텐츠 및 교대 형 콘텐츠의 양방이 풍부한 결과 항목을 획득하고자 한다. 더욱이, 질의어는 자원 항목 중 유사한 주제의 어떤 타입의 사전 분석 보다는 구조화되지 않은 자원으로부터 병렬 처리(parallelism)를 추출하는 1차 수단으로 된다.

[0056] 최종으로, 마이닝 시스템(104)은 질의어의 선택의 유효성을 드러내는 피드백을 수신할 수 있다. 이러한 피드백에 기초하여, 마이닝 시스템(104)은 질의어를 작성하는 방법을 관리하는 규칙을 수정할 수 있다. 게다가, 피드

백은 질의어를 공식화하는데 이용될 수 있는 특정 키워드 또는 주요 문구를 식별할 수 있다.

[0057] 블록(608)에서, 마이닝 시스템(104)은 질의어를 검색 모듈(116)에 제시한다. 그 후, 검색 모듈(116)은 질의어를 이용하여 구조화되지 않은 자원(110) 내에서 탐색 동작을 수행한다.

[0058] 블록(610)에서, 마이닝 시스템(104)은 검색 모듈(116)로부터 다시 결과 세트를 수신한다. 결과 세트는 결과 항목의 각각의 그룹을 포함한다. 각 결과 항목은 구조화되지 않은 자원(110) 내의 대응하는 자원 항목에서 추출된 텍스트 세그먼트에 대응할 수 있다.

[0059] 블록(612)에서, 마이닝 시스템(104)은 결과 세트의 초기 처리를 수행하여 트레이닝 세트를 생성한다. 상술한 바와 같이, 이러한 동작은 2개의 구성 요소를 포함할 수 있다. 필터링 구성 요소에서, 마이닝 시스템(104)은 의미론적 관련 문구를 식별할 시에 유용할 것 같지 않은 정보를 제거하거나 무시하도록 결과 세트를 제한한다. 매칭 구성 요소에서, 마이닝 시스템(104)은 예컨대 세트별 기준으로(on a set-by-set basis) 결과 항목의 쌍을 식별한다. 도 4는 예시적인 결과 세트와 관련하여 이러한 동작을 그래프로 예시한 것이다. 도 7은 블록(612)에서 수행되는 동작에 관한 부가적인 상세 사항을 제공한다.

[0060] 블록(614)에서, 트레이닝 시스템(106)은 번역 모델(102)을 도출하기 위해 트레이닝 세트에서 동작하는 통계적 기술을 이용한다. 어떤 통계적 기계 번역 접근법은 어떤 타입의 문구 지향 접근법과 같이 이러한 동작을 수행하는데 이용될 수 있다. 일반적으로, 번역 모델(102)은 $P(y|x)$ 로서 나타낼 수 있고, 이는 출력 문구 y가 주어진 입력 문구 x를 나타내는 확률을 규정한다. Bayes 규칙을 이용하여, 이것은 $P(y|x) = P(x|y)P(y)/P(x)$ 로 표현될 수 있다. 트레이닝 시스템(106)은 트레이닝 세트의 조사에 기초하여 이러한 식에 의해 정해지는 확률을 밝히기 위해 동작하며, 학습의 목적은 $P(x|y)P(y)$ 을 최대화하는 경향이 있는 입력 문구 x로부터 맵핑한다. 상술한 바와 같이, 조사는 사실상 반복된다. 각 동작 단계에서, 트레이닝 시스템(106)은 트레이닝 세트 내의 문구(및 대체로 텍스트 세그먼트)의 정렬에 관한 임시적 결론에 도달할 수 있다. 문구 지향 SMT 접근법에서, 임시적 결론은 번역 테이블 등을 이용하여 표현될 수 있다.

[0061] 블록(616)에서, 트레이닝 시스템(616)은 종료 조건이 도달되었는지를 판단하여, 만족스러운 정렬 결과가 달성되었음을 나타낸다. 어떤 메트릭(metric)은 잘 알려진 BLEU(Bilingual Evaluation Understudy) 스코어와 같은 이러한 판단을 하는데 이용될 수 있다.

[0062] 블록(618)에서, 만족스러운 결과가 아직 달성되지 않았다면, 트레이닝 시스템(106)은 트레이닝 시에 이용되는 어떤 가정을 수정한다. 이것은 결과 항목 내의 문구가 서로 어떻게 관계되는지(및 대체로 텍스트 세그먼트가 서로 어떻게 관계되는지)에 관한 일반적인 작업 가설을 수정하는 효과를 갖는다.

[0063] 종료 조건이 충족되었을 때, 트레이닝 시스템(106)은 트레이닝 세트 내의 의미론적 관련 문구 사이의 매핑을 식별할 것이다. 이들 매핑을 규정하는 매개 변수는 번역 모델(102)을 확립한다. 이와 같은 번역 모델(102)의 사용의 기초가 되는 추정은 텍스트의 새롭게 생성된 사례가 트레이닝 세트 내에서 발견된 패턴을 닮을 것이다.

[0064] 도 6의 절차는 서로 다른 방식으로 변경될 수 있다. 예컨대, 대안적 구현에서, 블록(614)에서의 트레이닝 동작은 번역 모델(102)을 도출하기 위해 만족스러운 분석 및 규칙 기반 분석의 조합을 이용할 수 있다. 다른 수정에서, 블록(614)에서의 트레이닝 동작은 트레이닝 태스크를 다수의 서브태스크로 나누어, 사실상 다수의 번역 모델을 생성할 수 있다. 그리고 나서, 트레이닝 동작은 다수의 번역 모델을 단일 번역 모델(102)에 병합할 수 있다. 다른 수정에서, 블록(614)에서의 트레이닝 동작은 시소러스 등으로부터 획득되는 정보와 같은 기준 소스를 이용하여 초기화되거나 "프라임(primed)"될 수 있다. 또 다른 수정이 가능하다.

[0065] 도 7은 도 6의 블록(612)에서 마이닝 시스템(104)에 의해 수행되는 필터링 및 매칭 처리에 관한 부가적인 상세 사항을 제공하는 절차(700)를 도시한다.

[0066] 블록(702)에서, 마이닝 시스템(104)은 하나 이상의 고려 사항에 기초하여 원래의 결과 세트를 필터링한다. 이러한 동작은 쌍별 매칭을 위한 가장 적절한 후보자인 결과 항목의 서브세트를 식별하는 효과를 갖는다. 이러한 동작은 (예컨대, 관련성이 낮은 것으로 평가되는 결과 항목을 제거하거나 무시함으로써) 트레이닝 세트의 복잡성 및 트레이닝 세트의 노이즈의 양을 감소시키는데 도움을 준다.

[0067] 하나의 경우에, 마이닝 시스템(104)은 결과 항목과 관련된 순위 스코어에 기초하여 쌍별 매칭을 위한 적절한 후보자로서 결과 항목을 식별할 수 있다. 부정적으로 언급되는 바와 같이(stated in the negative), 마이닝 시스

템(104)은 소정의 관련성 임계값 이하의 순위 스코어를 가진 결과 항목을 제거할 수 있다.

[0068] 대안적으로 또는 부가적으로, 마이닝 시스템(104)은 (예컨대, 결과 세트에 나타나는 단어의 공통성에 기초하여) 결과 세트 내에서 발견된 전형적인 텍스트 특징을 표현하는 각각의 결과 세트에 대한 어휘 특징(lexical signature)을 생성시킬 수 있다. 그 후, 마이닝 시스템(104)은 결과 세트와 관련된 어휘 특징과 각 결과 항목을 비교할 수 있다. 마이닝 시스템(104)은 이러한 비교에 기초하여 쌍별 매칭을 위한 적절한 후보자로서 결과 항목을 식별할 수 있다. 부정적으로 명시되는 바와 같이, 마이닝 시스템(104)은 소정량만큼 어휘 특징과 상이한 결과 항목을 제거할 수 있다. 덜 공식적으로 명시되는 바와 같이, 마이닝 시스템(104)은 각각의 결과 세트 내에서 "돌출(stand out)"되는 결과 항목을 제거할 수 있다.

[0069] 대안적으로 또는 부가적으로, 마이닝 시스템(104)은 각 결과 항목이 결과 세트 내에서 각 다른 결과 항목에 대해 얼마나 유사한지를 식별하는 유사도 스코어를 생성할 수 있다. 마이닝 시스템(104)은 코사인 유사도 메트릭과 같이 이러한 판단을 하는 어떤 유사도 메트릭에 의존할 수 있지만, 이에 제한되지 않는다. 마이닝 시스템(104)은 이들 유사도 스코어에 기초하여 쌍별 매칭을 위한 적절한 후보자로서 결과 항목을 식별할 수 있다. 부정적으로 명시되는 바와 같이, 마이닝 시스템(104)은 매칭을 위한 양호한 후보자가 아닌 결과 항목의 쌍을 식별할 수 있는데, 그 이유는 이들이 유사도 스코어에 의해 나타나는 바와 같이 소정량 이상만큼 서로 상이하기 때문이다.

[0070] 대안적으로 또는 부가적으로, 마이닝 시스템(104)은, 예컨대, k-최근접 이웃 클러스터링 기법(k-nearest neighbor clustering technique) 또는 어떤 다른 클러스터링 기법을 이용하여 유사한 결과 항목의 그룹을 결정하도록 결과 세트 내의 결과 항목에 대한 클러스터 분석을 수행할 수 있다. 그 후, 마이닝 시스템(104)은 서로 다른 클러스터에 걸친 후보자가 아닌 쌍별 매칭을 위한 적절한 후보자로서 각 클러스터 내의 결과 항목을 식별할 수 있다.

[0071] 마이닝 시스템(104)은 구조화되지 않은 자원(110)으로부터 수집되는 결과 항목을 필터링하거나 "클린 업(clean up)"하는 또 다른 동작을 수행할 수 있다. 블록(702)은 필터링된 결과 세트를 생성시킨다.

[0072] 블록(704)에서, 마이닝 시스템(104)은 필터링된 결과 세트 내의 쌍을 식별한다. 상술한 바와 같이, 도 4는 이러한 동작이 예시적 결과 세트의 콘텍스트 내에서 어떻게 수행될 수 있는지를 도시한다.

[0073] 블록(706)에서, 마이닝 시스템(104)은 트레이닝 세트를 제공하도록 (개개의 결과 세트와 관련된) 블록(704)의 결과를 조합할 수 있다. 상술한 바와 같이, 도 5는 이러한 동작이 어떻게 수행될 수 있는지를 도시한다.

[0074] 블록(704)이 설명을 용이하게 하기 위해 블록(702)에서 분리하여 도시되지만, 블록(702 및 704)은 통합된 동작으로 수행될 수 있다. 더욱이, 블록(702 및 704)의 필터링 및 매칭 동작은 다수의 동작 단계를 통해 분산될 수 있다. 예컨대, 마이닝 시스템(104)은 블록(706)에 따른 결과 항목에 대한 추가적 필터링을 수행할 수 있다. 더욱이, 트레이닝 시스템(106)은 (도 6의 블록(614-618)으로 나타낸 바와 같이) 반복 처리 중에 결과 항목에 대한 추가적 필터링을 수행할 수 있다.

[0075] 다른 변형으로서, 블록(704)은 개개의 결과 세트 내의 결과 항목의 쌍의 확립과 관련하여 설명되었다. 그러나, 다른 모드에서, 마이닝 시스템(104)은 서로 다른 결과 세트에 걸쳐 후보자 쌍을 확립할 수 있다.

[0076] 도 8은 번역 모델(102)의 예시적 애플리케이션을 설명하는 절차(800)를 도시한다.

[0077] 블록(802)에서, 애플리케이션 모듈(108)은 입력 문구를 수신한다.

[0078] 블록(804)에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 입력 문구를 출력 문구로 변환한다.

[0079] 블록(806)에서, 애플리케이션 모듈(108)은 출력 문구에 기초하여 출력 결과를 생성한다. 서로 다른 애플리케이션 모듈은 서로 다른 각각의 이득을 달성하도록 서로 다른 각각의 출력 결과를 제공할 수 있다.

[0080] 한 시나리오에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 질의어 수정 동작을 수행할 수 있다. 여기서, 애플리케이션 모듈(108)은 입력 문구를 탐색 질의어로 간주한다. 애플리케이션 모듈(108)은 출력 문구를 이용하여 탐색 질의어를 대신하거나 보충할 수 있다. 예컨대, 입력 문구가 "shingles"이면, 애플리케이션 모듈(108)은 출력 문구 "zoster"를 이용하여 "shingles AND zoster"의 보충된 질의어를 생성할 수 있다. 그리고 나서, 애플리케이션 모듈(108)은 확장 질의어를 탐색 엔진에 제공할 수 있다.

[0081] 다른 시나리오에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 인덱스 분류 결정을 행할 수 있다. 여기서, 애플리케이션 모듈(108)은 분류될 문서에서 어떤 텍스트 콘텐츠를 추출하여, 그 텍스트 콘텐츠를 입력

문구로 간주할 수 있다. 애플리케이션 모듈(108)은 출력 문구를 이용하여 문서의 주제에 관한 부가적인 통찰력을 모을 수 있으며, 이는 결과적으로 문서의 적절한 분류를 제공하는데 이용될 수 있다.

[0082] 다른 시나리오에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 어떤 타입의 텍스트 수정 동작을 수행할 수 있다. 여기서, 애플리케이션 모듈(108)은 입력 문구를 텍스트 수정을 위한 후보자로 간주할 수 있다. 애플리케이션 모듈(108)은 출력 문구를 이용하여 입력 문구가 수정될 수 있는 방식을 제시할 수 있다. 예컨대, 입력 문구는 오히려 장황한 텍스트 "rash that is painful(고통스러운 발진)"에 대응하는 것으로 추정한다. 애플리케이션 모듈(108)은 이러한 입력 문구가 더욱 간결한 "painful rash(고통스러운 발진)"로 대체될 수 있음을 제시할 수 있다. 이러한 제시를 행할 시에, 애플리케이션 모듈(108)은 원래의 문구에서 모든 문법 및/또는 철자 오류를 고칠 수 있다(출력 문구는 문법 및/또는 철자 오류를 포함하지 않는 것으로 가정한다). 하나의 경우에, 애플리케이션 모듈(108)은 입력 문구를 수정할 수 있는 방법에 관한 다수의 선택을 사용자에게 제공할 수 있으며, 사용자가 서로 다른 수정의 적합성을 측정하도록 하는 어떤 타입의 정보와 결합될 수 있다. 예컨대, 애플리케이션 모듈(108)은 당신의 아이디어가 (대표예만을 인용하도록) 저자의 80%가 사용하는 말씨의 방법을 표시하여 특정 수정에 주석을 달수 있다. 대안적으로, 애플리케이션 모듈(108)은 하나 이상의 고려 사항에 기초하여 자동으로 수정할 수 있다.

[0083] 다른 텍스트 수정 케이스에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 텍스트 절단(text truncation) 동작을 수행할 수 있다. 예컨대, 애플리케이션 모듈(108)은 이동 전화 장치 등과 같은 작은 스크린 보기(small-screened viewing) 장치 상의 프레젠테이션(presentation)을 위한 원본 텍스트를 수신할 수 있다. 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여, 입력 문구로 간주되는 텍스트를 텍스트의 단축된 버전으로 변환할 수 있다. 다른 케이스에서, 애플리케이션 모듈(108)은 이러한 접근법을 이용하여, 트위터형 통신 메카니즘과 같이 메시지에 사이즈 제한을 가하는 어떤 메시지 전송 메카니즘과 양립할 수 있도록 원래의 문구를 짧게 할 수 있다.

[0084] 다른 텍스트 수정 케이스에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 문서 또는 문구를 요약할 수 있다. 예컨대, 애플리케이션 모듈(108)은 이러한 접근법을 이용하여 원래의 초록의 길이를 감소시킬 수 있다. 다른 케이스에서, 애플리케이션 모듈(108)은 이러한 접근법을 이용하여 텍스트의 더욱 긴 메시지에 기반하는 타이틀을 제시할 수 있다. 대안적으로, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 문서 또는 문구를 확장할 수 있다.

[0085] 다른 시나리오에서, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 광고 정보의 확장을 수행할 수 있다. 여기서, 예컨대, 광고주는 광고 콘텐츠(예컨대, 웹 페이지 또는 다른 네트워크 액세스 가능한 콘텐츠)와 관련되는 초기 트리거 키워드를 선택할 수 있다. 최종 사용자가 이들 트리거 키워드에 들어가거나, 사용자가 이들 트리거 키워드와 관련되는 콘텐츠를 소비하면, 광고 메카니즘은 트리거 키워드와 관련되는 광고 콘텐츠로 사용자를 지향시킬 수 있다. 여기서, 애플리케이션 모듈(108)은 트리거 키워드의 초기 세트를 번역 모델(102)을 이용하여 확장될 입력 문구로 간주할 수 있다. 대안적으로 또는 부가적으로, 애플리케이션 모듈(108)은 광고 콘텐츠 자체를 입력 문구로 간주할 수 있다. 그 후, 애플리케이션 모듈(108)은 번역 모델(102)을 이용하여 광고 콘텐츠와 관계되는 텍스트를 제시할 수 있다. 광고주는 제시된 텍스트에 기초하여 하나 이상의 트리거 키워드를 제공할 수 있다.

[0086] 상술한 애플리케이션은 대표적이고, 철저한 것이 아니다. 다른 애플리케이션이 가능하다.

[0087] 상기 논의에서, 출력 문구는 입력 문구와 동일한 언어로 표현된다는 가정을 한다. 이 경우에, 출력 문구는 입력 문구의 바꿔쓰기(paraphrasing)로 고려될 수 있다. 다른 경우에, 마이닝 시스템(104) 및 트레이닝 시스템(106)은 제 1 언어의 문구를 다른 언어(또는 다수의 다른 언어)의 대응하는 문구로 변환하는 번역 모델(102)을 생성하는데 이용될 수 있다.

[0088] 2개의 언어 또는 다수의 언어를 사용하는 콘텍스트에서 동작하기 위해, 마이닝 시스템(104)은 2개의 언어 또는 다수의 언어를 사용하는 정보에 대해 상술한 바와 같은 기본 동작을 수행할 수 있다. 하나의 경우에, 마이닝 시스템(104)은 네트워크 환경 내의 병렬 질의어를 제시하여 2개의 언어를 사용하는 결과 세트를 확립할 수 있다. 즉, 마이닝 시스템(104)은 제 1 언어로 표현되는 질의어의 한 세트 및 제 2 언어로 표현되는 질의어의 다른 세트를 제시할 수 있다. 예컨대, 마이닝 시스템(104)은 문구 "rash zoster"를 제시하여 영어 결과 세트를 생성할 수 있고, 문구 "zoster erupcion de piel"를 제시하여 영어 결과 세트의 스페인어 대응(Spanish counterpart)을 생성할 수 있다. 그 후, 마이닝 시스템(104)은 영어 결과 항목을 스페인어 결과 항목에 링크하는 쌍을 확립할 수 있다. 이러한 매칭 동작의 목적은 트레이닝 시스템(106)이 영어 및 스페인어의 의미론적 관련 문구 사이

의 링크를 식별하도록 하는 트레이닝 세트를 제공하기 위한 것이다.

[0089] 다른 경우에, 마이닝 시스템(104)은 질의어 "shingles rash erupcion de piel"의 경우에서와 같이 영어 및 스페인어 주요 용어의 양방을 조합하는 질의어를 제시할 수 있다. 이러한 접근법에서, 검색 모듈(116)은 영어로 표현되는 결과 항목 및 스페인어로 표현되는 결과 항목을 조합하는 결과 세트를 제공하는 것으로 예상될 수 있다. 그 후, 마이닝 시스템(104)은 결과 항목이 영어 또는 스페인어로 표현되는지를 구별하지 않고 이러한 혼합된 결과 세트 내의 서로 다른 결과 항목 사이에 링크를 확립할 수 있다. 트레이닝 시스템(106)은 혼합된 트레이닝 세트 내의 기본 패턴에 기초하여 단일 번역 모델(102)을 생성할 수 있다. 사용 중에, 번역 모델(102)은 하나의 언어를 사용하는 모드로 적용될 수 있으며, 여기서 그것은 입력 문구와 동일한 언어로 출력 문구를 생성시키도록 제한된다. 또는 번역 모델(102)은 2개의 언어를 사용하는 모드로 동작할 수 있으며, 여기서 그것은 입력 문구와 비교해 서로 다른 언어로 출력 문구를 생성시키도록 제한된다. 또는 번역 모델(102)은 두 언어의 결과를 제시하는 구속받지 않는 모드로 동작할 수 있다.

[0090] C. 대표적 처리 기능

[0091] 도 9는 상술한 기능의 어떤 양태를 구현하기 위해 이용될 수 있는 예시적인 전기적 데이터 처리 기능을 설명한 것이다. 도 1 및 2를 참조하면, 예컨대, 도 9에 도시된 처리 기능(900)의 타입은 시스템(100) 또는 컴퓨팅 기능(202) 등의 어떤 양태를 구현하기 위해 이용될 수 있다. 하나의 경우에, 처리 기능(900)은 하나 이상의 처리 장치를 포함하는 컴퓨팅 장치의 어떤 타입에 대응할 수 있다.

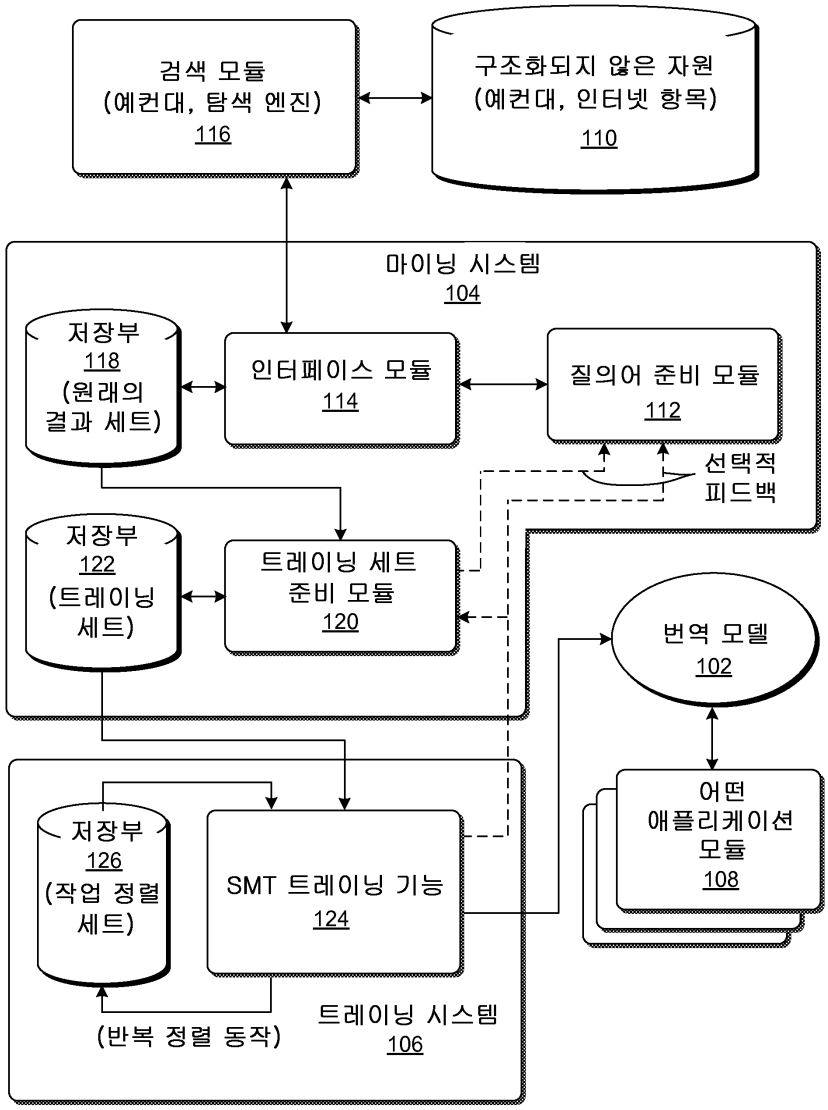
[0092] 처리 기능(900)은 하나 이상의 처리 장치(906) 뿐만 아니라 RAM(902) 및 ROM(904)와 같은 휘방성 및 비휘발성 메모리를 포함할 수 있다. 처리 기능(900)은 또한 선택적으로 하드 디스크 모듈, 광 디스크 모듈 등과 같은 여러 매체 장치(908)를 포함한다. 처리 기능(900)은 처리 장치(906)가 메모리(예컨대, RAM(902), ROM(904)등)에 의해 유지되는 명령어를 실행할 때에 상기 식별된 여러 동작을 수행할 수 있다. 일반적으로, 명령어 및 다른 정보는 정적 메모리 저장 장치, 자기 저장 장치, 광 저장 장치등을 포함하지만, 이에 제한되지 않는 어떤 컴퓨터 판독 가능한 매체(910) 상에 저장될 수 있다. 용어 컴퓨터 판독 가능한 매체는 또한 다수의 저장 장치를 포함한다. 용어 컴퓨터 판독 가능한 매체는 또한, 예컨대, 전선, 케이블, 무선 전송 등을 통해 제 1 위치에서 제 2 위치로 전송되는 신호를 포함한다.

[0093] 처리 기능(900)은 또한 (입력 모듈(914)을 통해) 사용자로부터 여러 입력을 수신하여, (출력 모듈을 통해) 여러 출력을 사용자에 제공하는 입력/출력 모듈(912)을 포함한다. 한 특정 출력 메카니즘은 프레젠테이션 모듈(916) 및 관련된 그래픽 사용자 인터페이스(GUI)(918)를 포함할 수 있다. 처리 기능(900)은 또한 하나 이상의 통신 관로(922)를 통해 데이터를 다른 장치와 교환하기 위한 하나 이상의 네트워크 인터페이스(920)를 포함할 수 있다. 하나 이상의 통신 버스(924)는 상술한 구성 요소를 서로 통신 가능하게 결합한다.

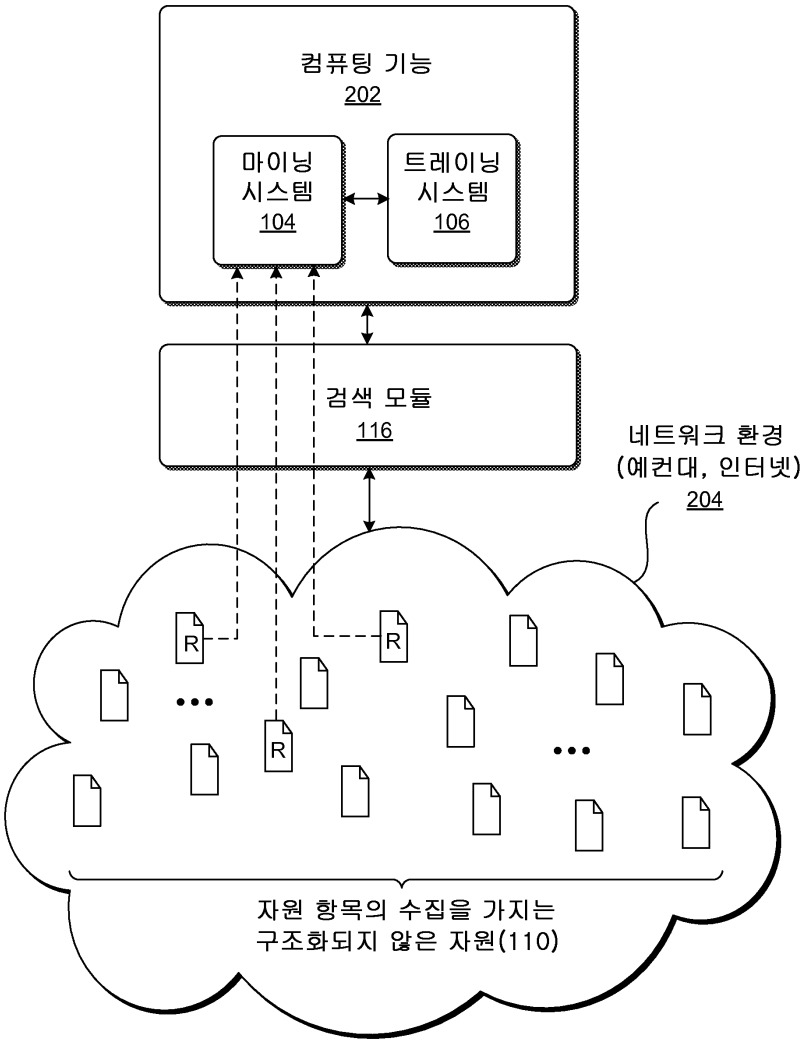
[0094] 내용이 구조적 특징 및/또는 방법 논리적 동작에 특정된 언어로 설명되었지만, 첨부된 청구범위에 정의된 내용은 반드시 상술된 특징 또는 동작으로 제한되지 않는 것으로 이해되어야 한다. 오히려, 상술된 특정한 특징 및 동작은 청구범위를 구현하는 예시적인 형식으로서 개시되었다.

도면

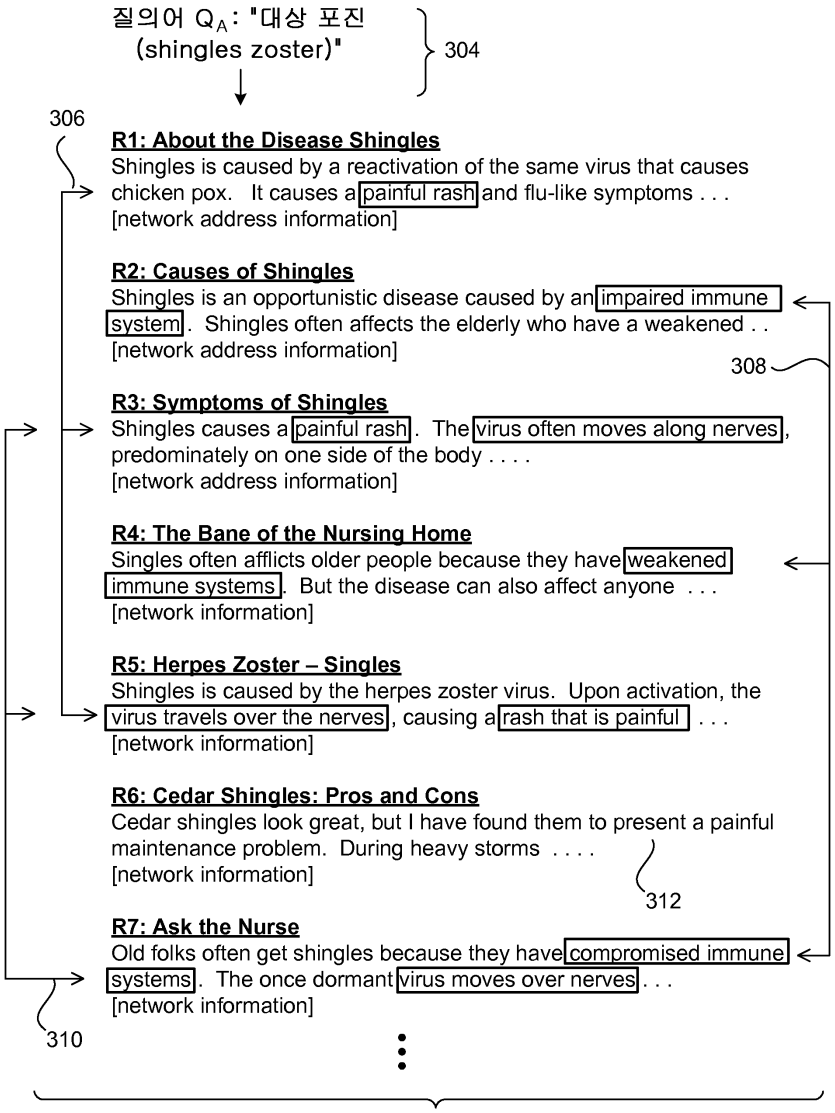
도면1


검색 모듈
(예컨대, 탐색 엔진)
116
구조화되지 않은 자원
(예컨대, 인터넷 항목)
110
마이닝 시스템
104
저장부
118
(원래의
결과 세트)
인터페이스 모듈
114
질의어 준비 모듈
112
선택적
피드백
저장부
122
(트레이닝
세트)
트레이닝 세트
준비 모듈
120
번역 모델
102
어떤
애플리케이션
모듈
108
저장부
126
(작업 정렬
세트)
SMT 트레이닝 기능
124
(반복 정렬 동작)
트레이닝 시스템
106

도면2

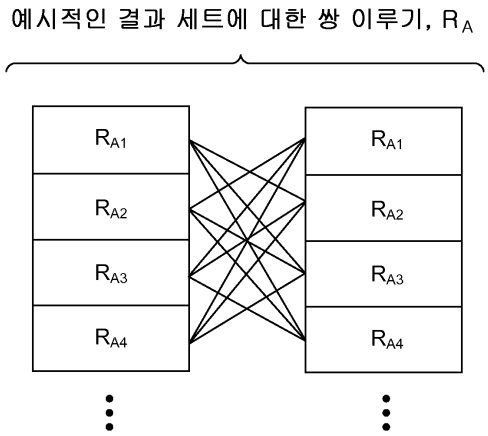

컴퓨팅 기능
202
마이닝 시스템
104
트레이닝 시스템
106
검색 모듈
116
네트워크 환경
(예컨대, 인터넷)
204
R
R
R
자원 항목의 수집을 가지는
구조화되지 않은 자원(110)

도면3

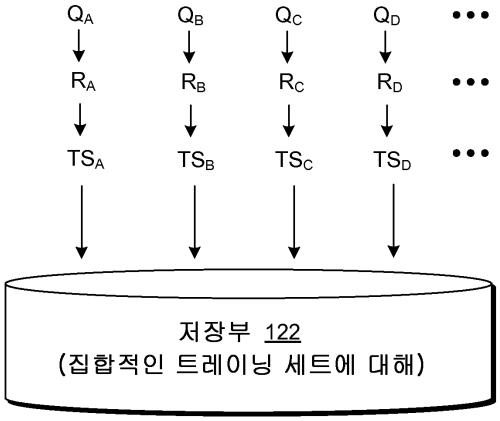


예시적인 결과 세트 302

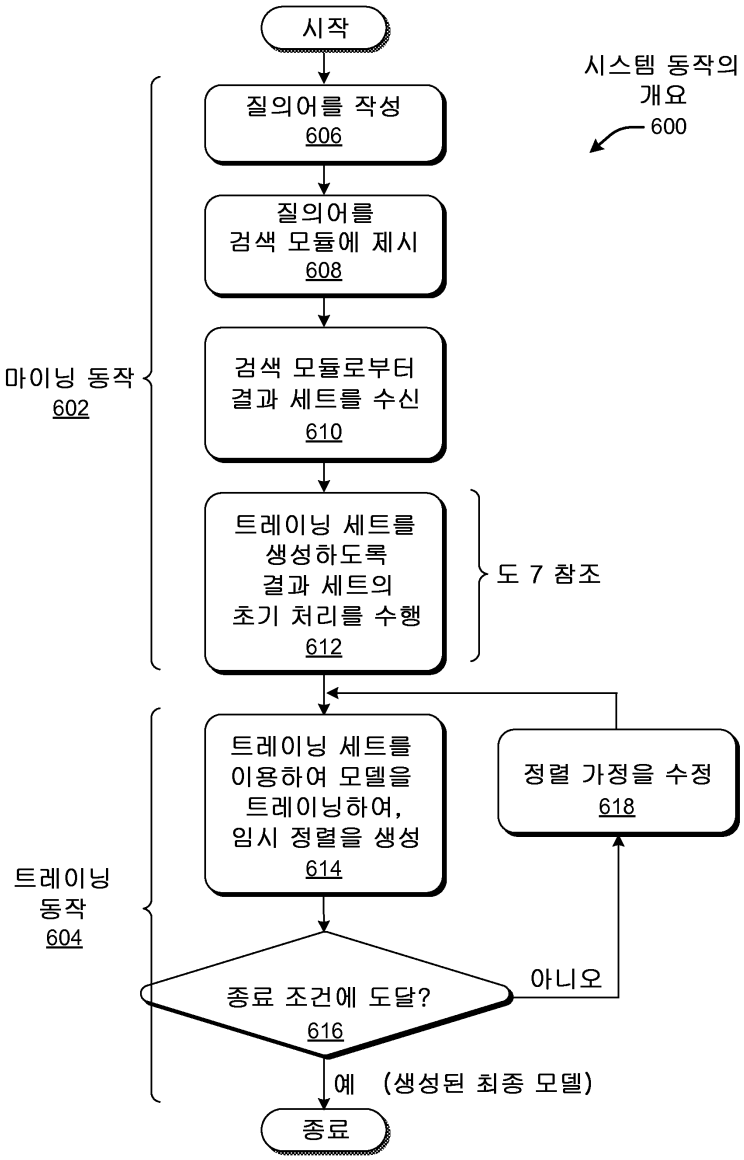
도면4



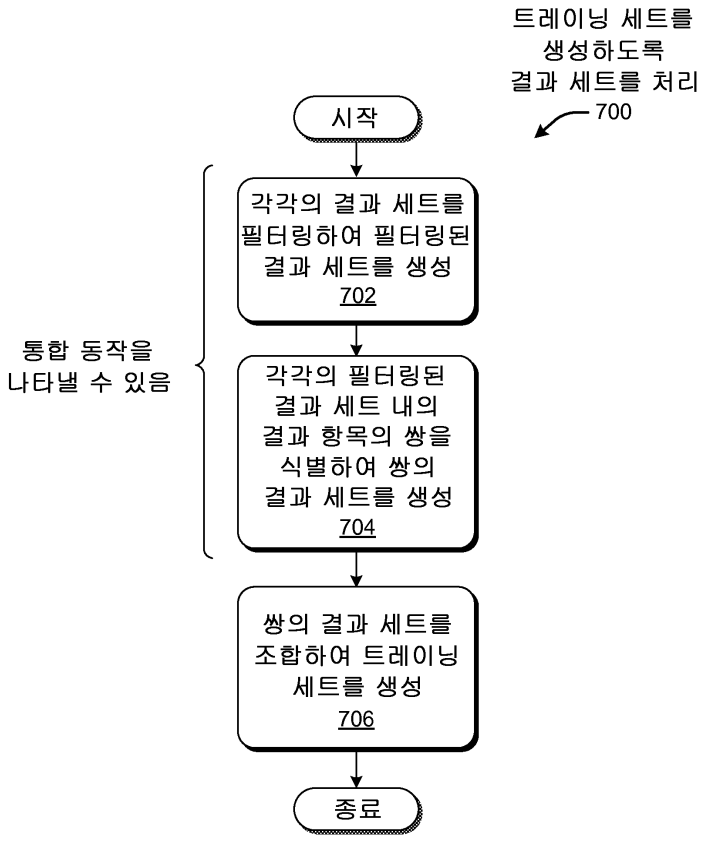
도면5


Q_A Q_B Q_C Q_D •••
R_A R_B R_C R_D •••
TS_A TS_B TS_C TS_D •••
저장부 122
(집합적인 트레이닝 세트에 대해)

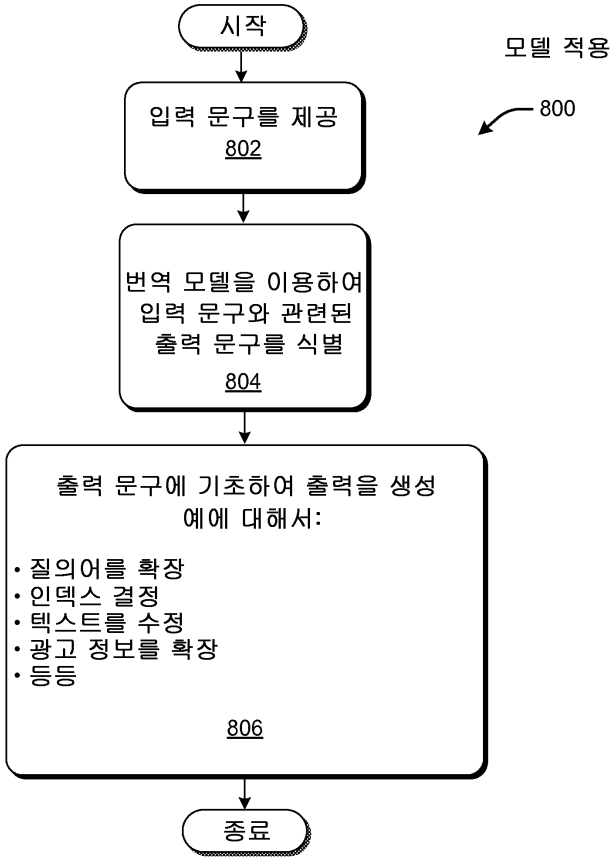
도면6


시작
시스템 동작의 개요
600
질의어를 작성
606
질의어를
검색 모듈에 제시
608
검색 모듈로부터
결과 세트를 수신
610
트레이닝 세트를
생성하도록
결과 세트의
초기 처리를 수행
612
도 7 참조
마이닝 동작
602
트레이닝 세트를
이용하여 모델을
트레이닝하여,
임시 정렬을 생성
614
정렬 가정을 수정
618
트레이닝
동작
604
종료 조건에 도달?
616
아니오
예 (생성된 최종 모델)
종료

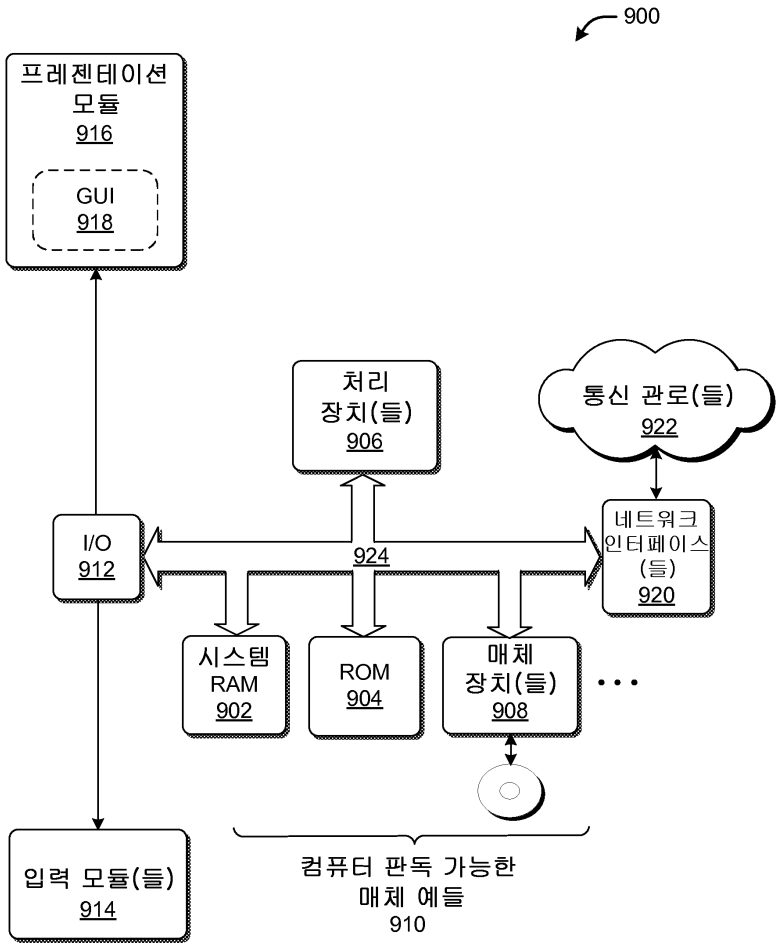
도면7


트레이닝 세트를
생성하도록
결과 세트를 처리
700
시작
각각의 결과 세트를
필터링하여 필터링된
결과 세트를 생성
702
통합 동작을
나타낼 수 있음
각각의 필터링된
결과 세트 내의
결과 항목의 쌍을
식별하여 쌍의
결과 세트를 생성
704
쌍의 결과 세트를
조합하여 트레이닝
세트를 생성
706
종료

도면8


시작
모델 적용
800
입력 문구를 제공
802
번역 모델을 이용하여
입력 문구와 관련된
출력 문구를 식별
804
출력 문구에 기초하여 출력을 생성
예에 대해서:
• 질의어를 확장
• 인덱스 결정
• 텍스트를 수정
• 광고 정보를 확장
• 등등
806
종료

도면9


900
프레젠테이션 모듈
916
GUI
918
처리 장치(들)
906
통신 관로(들)
922
I/O
912
924
네트워크 인터페이스(들)
920
시스템 RAM
902
ROM
904
매체 장치(들)
908
입력 모듈(들)
914
컴퓨터 판독 가능한 매체 예들
910