(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0042829
(43) 공개일자 2012년05월03일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01)
(21) 출원번호 10-2012-7000254
(22) 출원일자(국제) 2010년05월19일
심사청구일자 없음
(85) 번역문제출일자 2012년01월04일
(86) 국제출원번호 PCT/US2010/035413
(87) 국제공개번호 WO 2010/141219
국제공개일자 2010년12월09일
(30) 우선권주장
12/479,522 2009년06월05일 미국(US)

(71) 출원인
구글 인코포레이티드
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
(72) 발명자
사이트스 리차드 엘.
미국 캘리포니아주 94025 멘로 파크 캄포 벨로 레인 145
(74) 대리인
특허법인태평양

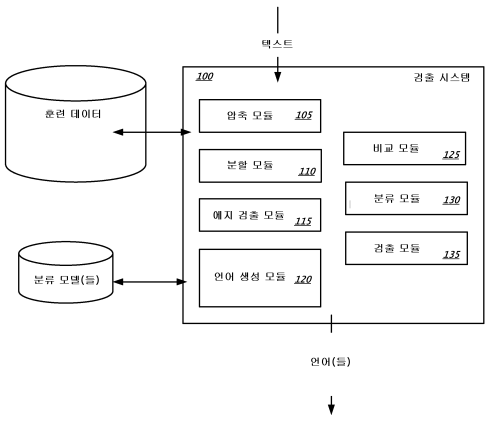
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **쓰기 체계 및 언어 검출**

(57) 요 약

쓰기 체계들과 언어들을 검출하는 방법들, 시스템들, 컴퓨터 프로그램 제품들을 포함하는 장치들이 개시된다. 일 구현예에서는, 방법이 제공된다. 그 방법은 텍스트를 수신하는 단계; 상기 텍스트의 제1 세그먼트를 검출하는 단계?상기 제1 세그먼트의 상당한 양이 제1 언어를 표현함?; 상기 텍스트의 제2 세그먼트를 검출하는 단계?상기 제2 세그먼트의 상당한 양이 제2 언어를 표현함?; 상기 텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별하는 단계; 및 상기 점수들의 변화량에 기초하여 상기 텍스트에서 상기 제1 언어에서 상기 제2 언어로의 전환(transition)을 식별하는 에지(edge)를 검출하는 단계를 포함한다.

대 표 도 - 도1



특허청구의 범위

청구항 1

컴퓨터 구현 방법으로서,

텍스트를 수신하는 단계;

상기 텍스트의 제1 세그먼트를 검출하는 단계?상기 제1 세그먼트의 상당한 양이 제1 언어를 표현함?;

상기 텍스트의 제2 세그먼트를 검출하는 단계?상기 제2 세그먼트의 상당한 양이 제2 언어를 표현함?;

상기 텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별하는 단계; 및

상기 점수들의 변화량에 기초하여 상기 텍스트에서 상기 제1 언어에서 상기 제2 언어로의 전환(transition)을 식별하는 에지(edge)를 검출하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 2

청구항 1에 있어서, 점수는 상기 n-그램이 특정 언어를 표현할 가능성을 표현하는 것을 특징으로 하는 방법.

청구항 3

청구항 1에 있어서, 점수는 상기 n-그램이 상기 제1 언어를 표현할 가능성과 상기 n-그램이 상기 제2 언어를 표현할 가능성 간의 차이를 표현하는 것을 특징으로 하는 방법.

청구항 4

청구항 3에 있어서, 에지를 검출하는 단계는

연속하는 n-그램의 제1 그룹에 대한 점수들의 제1 평균을 계산하는 단계?여기서, 연속하는 n-그램은 제1 왼쪽 콘텍스트와 제1 오른쪽 콘텍스트를 포함하는 제3 n-그램과, 제2 왼쪽 콘텍스트와 제2 오른쪽 콘텍스트를 포함하는 제4 n-그램로서 정의되고, 상기 제2 왼쪽 콘텍스트는 상기 제1 오른쪽 콘텍스트이고, 연속하는 n-그램의 상기 제1 그룹은 마지막 n-그램을 포함하는, 특정 개수의 연속하는 n-그램을 포함하는 것으로 정의됨?;

연속하는 n-그램의 제2 그룹에 대한 점수들의 제2 평균을 계산하는 단계?연속하는 n-그램의 상기 제2 그룹은 개시 n-그램을 포함하는, 동일한 개수의 연속적인 n-그램을 포함하는 것으로서 정의되고, 상기 마지막 n-그램은 상기 개시 n-그램과 인접함?; 및

상기 제1 평균과 상기 제2 평균 간의 차이에 기초하여 에지를 식별하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 5

컴퓨터 구현 방법으로서,

텍스트를 수신하는 단계;

상기 텍스트의 제1 부분에서 표현된 쓰기 체계(writing system)를 식별하는 단계?상기 쓰기 체계는 하나 이상의 제1 언어를 표현함?; 및

상기 텍스트의 상기 제1 부분에 표현된 하나 이상의 제1 언어에서만 특정 언어를 검출하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 6

청구항 5에 있어서, 상기 쓰기 체계를 식별하는 단계는 상기 텍스트의 제1 부분에 있는 인코딩(encoding)으로 된 문자들을 상응하는 쓰기 체계에 매핑하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 7

청구항 6에 있어서, 상기 인코딩은 유니코드(Unicode) 인 것을 특징으로 하는 방법.

청구항 8

컴퓨터 구현 방법으로서,

문서를 수신하는 단계;

상기 문서의 제1 부분을 식별하는 단계?여기서, 제1 부분에 있는 텍스트의 상당한 양이 제1 쓰기 체계로 된 텍스트를 표현함?;

상기 문서의 제1 부분에서 하나 이상의 세그먼트를 식별하는 단계?여기서, 하나 이상의 세그먼트 각각에 있는 텍스트의 상당한 양이 상기 제1 쓰기 체계의 언어로 표현됨?; 및

상기 하나 이상의 세그먼트에서 텍스트의 상당한 양에 의해 표현되는 상기 제1 쓰기 체계의 특정 언어를 검출하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 9

청구항 8에 있어서,

상기 문서의 제2 부분을 식별하는 단계?여기서, 상기 제2 부분에 있는 텍스트의 상당한 양이 제2 쓰기 체계로 된 텍스트를 표현함?; 및

상기 문서의 상기 제2 부분에 있는 하나 이상의 세그먼트를 식별하는 단계?하나 이상의 세그먼트의 각 부분에 있는 텍스트의 상당한 양이 상기 제2 쓰기 체계의 언어로 표현됨?를 더 포함하는 것을 특징으로 하는 방법.

청구항 10

청구항 8에 있어서, 상기 제1 쓰기 체계는 중국어, 일본어, 및 한국어를 표현하기 위해 사용되는 쓰기 체계들을 포함하는 병합된 쓰기 체계(merged writing system)인 것을 특징으로 하는 방법.

청구항 11

유형가능 프로그램 캐리어로 부호화되어 데이터 프로세싱 장치로 하여금 동작들을 수행할 수 있게 하는 컴퓨터 프로그램 제품으로서, 상기 동작들은

텍스트를 수신하는 단계;

상기 텍스트의 제1 세그먼트를 검출하는 단계?상기 제1 세그먼트의 상당한 양이 제1 언어를 표현함?;

상기 텍스트의 제2 세그먼트를 검출하는 단계?상기 제2 세그먼트의 상당한 양이 제2 언어를 표현함?;

상기 텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별하는 단계; 및

상기 점수들의 변화량에 기초하여 상기 텍스트에서 상기 제1 언어에서 상기 제2 언어로의 전환을 식별하는 에지를 검출하는 단계를 포함하는 것을 특징으로 하는 컴퓨터 프로그램 제품.

청구항 12

유형가능 프로그램 캐리어로 부호화되어 데이터 프로세싱 장치로 하여금 동작들을 수행할 수 있게 하는 컴퓨터 프로그램 제품으로서, 상기 동작들은

텍스트를 수신하는 단계;

상기 텍스트의 제1 부분에서 표현된 쓰기 체계를 식별하는 단계?상기 쓰기 체계는 하나 이상의 제1 언어를 표현함?; 및

상기 텍스트의 상기 제1 부분에 표현된 하나 이상의 제1 언어에서만 특정 언어를 검출하는 단계를 포함하는 것을 특징으로 하는 컴퓨터 프로그램 제품.

청구항 13

유형가능 프로그램 캐리어로 부호화되어 데이터 프로세싱 장치로 하여금 동작들을 수행할 수 있게 하는 컴퓨터

프로그램 제품으로서, 상기 동작들은

문서를 수신하는 단계;

상기 문서의 제1 부분을 식별하는 단계?여기서, 제1 부분에 있는 텍스트의 상당한 양이 제1 쓰기 체계로 된 텍스트를 표현함?;

상기 문서의 제1 부분에서 하나 이상의 세그먼트를 식별하는 단계?여기서, 하나 이상의 세그먼트 각각에 있는 텍스트의 상당한 양이 상기 제1 쓰기 체계의 언어로 표현됨?; 및

상기 하나 이상의 세그먼트에서 텍스트의 상당한 양에 의해 표현되는 상기 제1 쓰기 체계의 특정 언어를 검출하는 단계를 포함하는 것을 특징으로 하는 컴퓨터 프로그램 제품.

청구항 14

시스템으로서,

프로그램 제품을 포함하는 기계 판독가능 저장 디바이스; 및

상기 프로그램 제품들을 실행시켜 동작들을 수행할 수 있는 하나 이상의 컴퓨터를 포함하고, 상기 동작들은

텍스트를 수신하는 단계;

상기 텍스트의 제1 세그먼트를 검출하는 단계?상기 제1 세그먼트의 상당한 양이 제1 언어를 표현함?;

상기 텍스트의 제2 세그먼트를 검출하는 단계?상기 제2 세그먼트의 상당한 양이 제2 언어를 표현함?;

상기 텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별하는 단계; 및

상기 점수들의 변화량에 기초하여 상기 텍스트에서 상기 제1 언어에서 상기 제2 언어로의 전환을 식별하는 에지를 검출하는 단계를 포함하는 것을 특징으로 하는 시스템.

청구항 15

시스템으로서,

프로그램 제품을 포함하는 기계 판독가능 저장 디바이스; 및

상기 프로그램 제품들을 실행시켜 동작들을 수행할 수 있는 하나 이상의 컴퓨터를 포함하고, 상기 동작들은

텍스트를 수신하는 단계;

상기 텍스트의 제1 부분에서 표현된 쓰기 체계를 식별하는 단계?상기 쓰기 체계는 하나 이상의 제1 언어를 표현함?; 및

상기 텍스트의 상기 제1 부분에 표현된 하나 이상의 제1 언어에서만 특정 언어를 검출하는 단계를 포함하는 것을 특징으로 하는 시스템.

청구항 16

시스템으로서,

프로그램 제품을 포함하는 기계 판독가능 저장 디바이스; 및

상기 프로그램 제품들을 실행시켜 동작들을 수행할 수 있는 하나 이상의 컴퓨터를 포함하고, 상기 동작들은

문서를 수신하는 단계;

상기 문서의 제1 부분을 식별하는 단계?여기서, 제1 부분에 있는 텍스트의 상당한 양이 제1 쓰기 체계로 된 텍스트를 표현함?;

상기 문서의 제1 부분에서 하나 이상의 세그먼트를 식별하는 단계?여기서, 하나 이상의 세그먼트 각각에 있는 텍스트의 상당한 양이 상기 제1 쓰기 체계의 언어로 표현됨?; 및

상기 하나 이상의 세그먼트에서 텍스트의 상당한 양에 의해 표현되는 상기 제1 쓰기 체계의 특정 언어를 검출하는 단계를 포함하는 것을 특징으로 하는 시스템.

명 세 서

기 술 분 야

[0001] 본 명세서는 쓰기 체계들(writing system)과 언어들을 검출하는 것에 관한 것이다.

배 경 기 술

[0002] 쓰기 체계는 언어의 소리들을 표현하기 위해 기호들(예컨대, 문자들 또는 문자소(grapheme))을 사용한다. 쓰기 체계 내에 있는 기호들의 집합은 스크립트로서 불릴 수 있다. 예를 들어, 하나 이상의 로마 스크립트로 된 로마자의 집합을 포함하는 라틴 쓰기 체계가 영어를 표현하기 위해 사용될 수 있다. 특정 쓰기 체계가 둘 이상의 언어를 표현하기 위해 사용될 수 있다. 예를 들어, 라틴 쓰기 체계는 프랑스어를 표현하기 위해서도 사용될 수 있다.

[0003] 이에 더하여, 주어진 언어가 둘 이상의 쓰기 체계로 표현될 수도 있다. 예를 들어, 중국어는 제1 쓰기 체계(예컨대, 병음(Pinyin), 즉 로마자화된 중국어)로 표현될 수 있다. 중국어는 제2 쓰기 체계(예컨대, 보포모포(bopomofo), 즉 주인 푸하오("Zhuyin"))를 사용하여 표현될 수 있다. 또 다른 실시예로서, 중국어는 제3 쓰기 체계(예컨대, 한지(Hanzi))를 사용하여 표현될 수도 있다.

발명의 내용

해결하려는 과제

[0004] 쓰기 체계과 언어 간의 복잡한 관계는 입력 텍스트로부터 언어를 자동적으로 검출하는데 어려움을 증가시킨다. 입력 텍스트로부터의 언어 검출의 정확성과 정밀도는 분류자를 훈련하기 위해 사용되는 훈련 데이터의 양과 품질에 따라 달라질 수 있다.

과제의 해결 수단

[0005] 본 명세서는 언어 검출에 관한 기술들을 설명한다.

[0006] 일반적으로, 이 명세서에 설명된 주제 중 한 양태는 방법들로 구현될 수 있다. 그 방법들은 텍스트를 수신하는 단계; 텍스트의 제1 세그먼트를 검출하는 단계?제1 세그먼트의 상당한 양이 제1 언어를 표현함?; 텍스트의 제2 세그먼트를 검출하는 단계?제2 세그먼트의 상당한 양이 제2 언어를 표현함?; 텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별하는 단계; 및 그 점수들의 변화량에 기초하여 텍스트에서 제1 언어에서 제2 언어로의 전환을 식별하는 에지(edge)를 검출하는 단계의 동작들을 포함한다. 본 양태의 다른 양태들은 상응하는 시스템들, 장치들, 및 컴퓨터 프로그램 제품들을 포함한다.

[0007] 이러한 및 다른 실시예들은 선택적으로 하나 이상의 후술하는 특징들을 포함할 수 있다. 점수는 n-그램이 특정 언어를 표현할 가능성을 나타낸다. 점수는 n-그램이 제1 언어를 표현할 가능성과 n-그램이 제2 언어를 표현할 가능성 간의 차이를 나타낸다. 에지를 검출하는 단계는 연속하는 n-그램의 제1 그룹에 대한 점수들의 제1 평균을 계산하는 단계를 포함하는데, 여기서 연속하는 n-그램은 제1 왼쪽 콘텍스트와 제1 오른쪽 콘텍스트를 포함하는 제3 n-그램과, 제2 왼쪽 콘텍스트와 제2 오른쪽 콘텍스트를 포함하는 제4 n-그램로서 정의되고, 제2 왼쪽 콘텍스트는 제1 오른쪽 콘텍스트이고, 연속하는 n-그램의 제1 그룹은 마지막 n-그램을 포함하는, 특정 개수의 연속하는 n-그램을 포함하는 것으로 정의된다. 에지를 검출하는 단계는 연속하는 n-그램의 제2 그룹에 대한 점수들의 제2 평균을 계산하는 단계?연속하는 n-그램의 제2 그룹은 개시 n-그램을 포함하는, 동일한 개수의 연속적인 n-그램을 포함하는 것으로서 정의되고, 마지막 n-그램은 개시 n-그램과 인접함?; 및 제1 평균과 제2 평균 간의 차이에 기초하여 에지를 식별하는 단계를 더 포함한다.

[0008] 전체적으로, 이 명세서에 설명된 주제의 다른 양태는 방법들로 구현될 수 있다. 그 방법들은 텍스트를 수신하는 단계; 텍스트의 제1 부분에서 표현된 쓰기 체계를 식별하는 단계?상기 쓰기 체계는 하나 이상의 제1 언어를 표현함?; 및 텍스트의 제1 부분에 표현된 하나 이상의 제1 언어에서만 특정 언어를 검출하는 단계의 동작들을 포함한다. 본 양태의 다른 실시예들은 상응하는 시스템들, 장치들, 및 컴퓨터 프로그램 제품들을 포함한다.

[0009] 이러한 및 다른 실시예들은 선택적으로 하나 이상의 후술하는 특징들을 포함할 수 있다. 쓰기 체계를 식별하는 단계는 텍스트의 제1 부분에 있는 인코딩(encoding)으로 된 문자들을 상응하는 쓰기 체계에 매핑하는 단계를 포

함한다. 이 인코딩은 유니코드이다.

[0010] 전체적으로, 본 명세서에 설명된 주제의 다른 양태는 방법들로 구현될 수 있다. 그 방법들은 문서를 수신하는 단계; 문서의 제1 부분을 식별하는 단계?제1 부분에 있는 텍스트의 상당한 양이 제1 쓰기 체계로 된 텍스트를 표현함?; 문서의 제1 부분에서 하나 이상의 세그먼트를 식별하는 단계?하나 이상의 세그먼트 각각에 있는 텍스트의 상당한 양이 제1 쓰기 체계의 언어로 표현됨?; 및 하나 이상의 세그먼트에서 텍스트의 상당한 양에 의해 표현되는 제1 쓰기 체계의 특정 언어를 검출하는 단계의 동작들을 포함한다. 본 양태의 다른 구현예들은 상응하는 시스템들, 장치들, 및 컴퓨터 프로그램 제품들을 포함한다.

[0011] 이러한 및 다른 구현예들은 선택적으로 하나 이상의 후술하는 특징들을 포함할 수 있다. 이 방법은 문서의 제2 부분을 식별하는 단계?제2 부분에 있는 텍스트의 상당한 양이 제2 쓰기 체계로 된 텍스트를 표현함?; 및 문서의 제2 부분에 있는 하나 이상의 세그먼트를 식별하는 단계?하나 이상의 세그먼트의 각 부분에 있는 텍스트의 상당한 양이 제2 쓰기 체계의 언어로 표현됨?를 더 포함한다. 제1 쓰기 체계는 중국어, 일본어, 및 한국어를 표현하기 위해 사용되는 쓰기 체계를 포함하는 병합된 쓰기 체계(merged writing system)이다.

발명의 효과

[0012] 본 명세서에서 설명된 주제의 특정 실시예들은 하나 이상의 후술하는 장점들을 실현하기 위하여 구현될 수 있다.

[0013] 언어 검출을 위해 개시된 시스템들과 기술들은 예를 들어, 노이즈 데이터를 제거하고, 특정 언어의 대표로서 훈련 데이터를 정확하게 분류함으로써 훈련 데이터의 품질을 향상시키는데 사용될 수 있고, 이로써 입력 텍스트로부터 언어들을 검출하는 것에 대한 정확성, 효율성, 및 정밀성을 향상시킬 수 있다. 특히, 반복적인 텍스트를 검출하고 제거하는 것은 언어들이 검출될 수 있는 문서(예컨대, 웹 페이지들, 블로그들, 및 이메일들과 같은 노이즈가 많은 문서)들의 유형을 증가시키고, 이로써 이용가능한 훈련 데이터의 양을 증가시킨다. 이뿐 아니라, 많은 문서들이 둘 이상의 언어들로 된 텍스트를 포함하기 때문에, 단일 문서에서 섞여있는 언어들을 검출하는 것 또한 이용가능한 훈련 데이터의 양을 증가시킨다.

[0014] 또한 언어를 검출하기 위한 시스템들 및 기술들은 예를 들어, 입력 텍스트로부터 노이즈 데이터를 제거하고, 입력 텍스트가 표현할 수 있는 유일한 쓰기 체계들의 특정 언어들에 분석을 집중하고, 유사 언어들을 구별하고, 입력 텍스트에서 사용된 언어들 간의 정밀한 전환(transition)들을 검출함으로써, 스트림라인 언어 검출(streamline language detecion)에 사용될 수 있고, 이로써 그 입력로부터 언어들을 검출하는 것에 대한 정확성, 효율성, 및 정밀성을 더욱 향상시킬 수 있다.

[0015] 본 명세서에서 설명된 주제에 대한 하나 이상의 실시예가 첨부 도면들과 후술하는 상세한 설명에서 개시된다. 본 주제에 대한 다른 특징들, 양태들, 및 장점들은 상세한 설명, 도면들, 및 청구항들로부터 명백해질 것이다.

도면의 간단한 설명

[0016] 도 1은 예시적 검출 시스템을 포함한다.

도 2a는 압축(compression)을 사용하여 반복을 검출하는 예시적 프로세스를 나타낸다.

도 2b는 반복 토큰들을 포함하는 예시적 토큰 시퀀스를 예시한다.

도 3은 쓰기 체계들과 언어들을 검출하기 위해 세그먼트들을 식별하는 예시적 프로세스를 나타낸다.

도 4a는 제1 언어로 된 텍스트를 표현하는 토큰들의 제1 시퀀스에 이어지는 제2 언어로 된 텍스트를 표현하는 토큰들의 제2 시퀀스를 포함하는 예시적 텍스트를 나타낸다.

도 4b는 텍스트에서 표현된 다른 언어들 간의 에지(edge)들을 검출하는 예시적 프로세스를 나타낸다.

도 5는 인공 언어(artificial language)를 생성하고, 그 인공 언어를 사용하여 언어들을 검출하는 예시적 프로세스를 나타낸다.

도 6a는 유사 언어들로부터의 용어들을 포함하는 텍스트의 예시적 시퀀스를 나타낸다.

도 6b는 유사 언어들 간을 구별하는 예시적 프로세스를 나타낸다.

도 7은 일반적인 컴퓨터 시스템의 개략도이다.

여러 도면에서 유사한 참조 번호와 명칭들은 유사한 구성요소들을 가리킨다.

발명을 실시하기 위한 구체적인 내용

[0017] **통계적 언어 검출 개관(Statistical Language Detection Overview)**

[0018] n-그램은 n개의 연속적인 토큰들(예컨대, 단어들 또는 문자들)의 시퀀스이다. n-그램은 n-그램에 있는 토큰의 개수인 차수 또는 크기를 갖는다. 예를 들어, 1-그램(즉, 유니그램(unigram))은 하나의 토큰을 포함하고, 2-그램(즉, 바이그램(bi-gram))은 두개의 토큰을 포함한다.

[0019] 주어진 n-그램은 n-그램의 다른 부분들에 따라서 설명될 수 있다. n-그램은 콘텍스트와 미래 토큰으로서 설명될 수 있는데(context,w), 여기서, 콘텍스트는 길이 n-1를 갖고, w는 미래 토큰을 표현한다. 예를 들어, 3-그램 "$c_1c_2c_3$"는 n-그램 콘텍스트와 미래 토큰으로서 설명될 수 있는데, 여기서 c_1,c_2,c_3는 각각 문자를 표현한다. n-그램 왼쪽 콘텍스트(n-gram left context)는 n-그램의 마지막 토큰의 앞에 있는 n-그램의 모든 토큰을 포함한다. 이 주어진 예시에서, "c_1c_2"이 콘텍스트이다. 콘텍스트에 있는 최좌 토큰(the left most token)이 왼쪽 토큰으로 불리운다. 미래 토큰은 이 예에서 "c_3"인, n-그램의 마지막 토큰이다. n-그램은 또한 오른쪽 콘텍스트에 관련되어 설명될 수 있다. 오른쪽 콘텍스트는 (n-1)그램으로서 표현되고, n-그램의 첫 번째 토큰을 따르는 n-그램의 모든 토큰들을 포함한다. 상술되어진 예시에서, "c_2c_3"는 오른쪽 콘텍스트이다.

[0020] 각 n-그램은 예를 들어, 대수확률(log-probability)과 같은 연관된 확률 추정치(associated probability estimate)을 가질 수 있는데, 연관된 확률 추정치는 훈련 데이터에서 전체 출현 총수(count of total occurrences)에 대한 훈련 데이터에 특정 언어로 된 출현 총수에 대한 함수로서 계산된다. 예를 들어, 언어 검출 시스템은 훈련 데이터에서 모든 4-그램들(쿼드그램)들을 식별함으로써 훈련 데이터를 분석(parse)할 수 있다. 각 쿼드그램에 대해서, 훈련 데이터에서 각 특정 언어로 된 출현 개수의 총수는 유지관리되고 기록될 수 있다. 각 쿼드그램은 또한 그 쿼드그램이 특정 언어를 식별할 가능성을 나타내는 확률 추정에 연관될 수 있다. 예를 들어, en-Latn(예컨대, 영어-라틴어)에 상응하는 제1 쿼드그램에 대한 항목은 제1 쿼드 그램이 영어 텍스트를 나타낼 가능성에 연관될 수 있다. 다른 예시로서, 제1 쿼드그램에 대한 다른 항목은 또한 fr-Latn(예컨대, 프랑스어-라틴어)에 상응할 수 있고, 제1 쿼드그램이 프랑스어 텍스트을 나타낼 가능성에 연관될 수 있다.

[0021] 일부 구현예들에서, 특정 언어로 된 단어들을 식별하는 n-그램의 확률들은 훈련 데이터에서 특정 언어로 표현되는 n-그램의 상대 빈도(relative frequency)를 사용하여 훈련된다. 추가적으로, 일부 구현예들에서, 분산된 훈련 환경은 대형 훈련 데이터(예컨대, 테라바이트 데이터)에 대하여 사용될 수 있다. 분산된 훈련을 위한 예시적 기술 중 하나가 맵리듀스(MapReduce)이다. 맵리듀스의 추가적인 상세 내용은, 2004년 12월 6일에 있었던 운영 체계 설계 및 구현에 대한 6차 심포지엄의 논문집 137-150 페이지에, 제이.딘과 에스.게마와트에 의해 개재된 "랩리듀스: 대형 클러스터상에서 단순화된 데이터 프로세싱"에 설명되어 있다.

[0022] n-그램, 연관된 확률 추정치, 및 각각의 총수는 입력 텍스트에서 언어들을 검출하는 분류자(예컨대, 베이지안 분류자)에 의한 사용을 위하여 분류 모델에 저장될 수 있다. 입력 텍스트가 특정 언어를 표현할 가능성을 나타내는 점수는 입력 텍스트에 포함된 n-그램을 특정 언어에 대해 연관된 확률 추정치에 매핑하고, 각 n-그램에 대해 연관된 확률 추정치들에 상응하는 로그-우도(log-likelihood)들을 합산함으로써 계산될 수 있다.

[0023] 예를 들어, 훈련 데이터의 2개의 그룹은 프랑스어 훈련 데이터로 된 제1 그룹(예컨대, 프랑스어로 된 웹페이지들의 코퍼스, 탐색 쿼리 로그들, 이메일들, 및 블로그들)과 영어 훈련 데이터로 된 제2 그룹을 포함한다. 종래 기술들은 훈련 데이터의 각 그룹을 처리하는데 사용되어, 각 n-그램에 대하여 상술한 총수들과 확률 추정들은 예컨대, 해쉬 테이블을 사용하여 식별되고, 분류 모델에 저장될 수 있다. 이어서, 확률 추정치들이 다른 텍스트에 있는 언어들을 검출하기 위해서 사용될 수 있다. 특히, 다른 텍스트는 특정 크기의 n-그램을 사용하여 분석될 수 있다. 특정 크기의 n-그램들 각각의 확률을 결정하기 위해, 특정 크기의 n-그램들이 분류 모델에 있는 항목들과 비교될 수 있다. 후술되는 바와 같이, 다른 기술들과 종래 기술들에 대한 개량 기술들이 있을 수 있다.

[0024] **예시적 언어 검출 시스템(Example Language Detection System)**

[0025] 도 1은 예시적 검출 시스템(100)을 포함한다. 검출 시스템(100)은 압축 모듈(105), 분할 모듈(110), 에지 검출 모듈(115), 언어 생성 모듈(120), 비교 모듈(125), 분류 모듈(130), 및 검출 모듈(135)을 포함하고 있다. 검출 시스템(100) 의 구성요소들은 서로와 하나 이상 통신적으로 연결된다. 다른 구현예들도 있을 수 있다. 예를 들

어, 분류 모듈(130)은 검출 시스템(100)과 분리된 구성요소일 수 있다. 또한, 검출 시스템(100)에서 식별된 구성요소들이 논리적으로 분리되었거나 별개인 것으로 설명되었지만, 검출 시스템(100)의 하나 이상의 구성요소가 조합되거나 또는 추가로 나뉠 수 있다.

[0026] 검출 시스템(100)은 도 2a 내지 도 6b와 관련하여 후술되는 동작들을 포함하는 동작들을 수행함으로써 텍스트로 표현되는 쓰기 체계와 언어들을 검출하는데 사용될 수 있다. 일부 구현예들에서, 검출 시스템(100)은 다른 텍스트의 언어들을 식별하는데 사용될 수 있는 특정 언어들에 대한 훈련 데이터의 그룹들을 식별한다. 예를 들어, 훈련 데이터의 식별된 그룹은 언어 검출에 사용되는 네이브 베이지안 분류자(naïve Bayesian classifier)를 훈련하는데 사용될 수 있다. 일부 구현예들에서, 검출 시스템(100)은 특정 언어들을 표현하는 훈련 데이터의 특정된 그룹으로부터 생성된 분류 모델에 기초하여 입력 텍스트의 언어들을 식별한다. 다른 구현들이 있을 수 있다. 예를 들어, 검출 시스템(100)은 종래 기술들, 후술되는 기술들, 또는 이것들의 조합을 사용하여, 특정 언어들에 대한 훈련 데이터의 그룹들을 식별하고, 다른 입력 텍스트의 언어들을 검출하기 위하여 그 훈련 데이터의 식별된 그룹들을 사용할 수 있다.

[0027] 개관에서와 같이, 압축 모듈(105)은 특정 언어를 식별할 수 없는 텍스트의 반복을 검출한다. 분할 모듈(110)은 텍스트의 부분들에서 표현되는 특정 언어들의 정밀한 식별을 가능하게 하기 위하여 텍스트를 부분(portion)들로 분할한다. 에지 검출 모듈(115)는 텍스트의 부분들에 있는 언어들간의 전환들의 정확한 식별을 가능하게 하기 위하여, 특정 언어로 표현되고 있는 텍스트 내 시퀀스들의 확률들을 식별하기 위하여 사용되는 점수의 에지들을 검출한다. 언어 생성 모듈(120)은 텍스트를 표현하는 특정 언어들을 식별하는데 사용되면 안 되는 텍스트의 부분들의 식별을 가능하게 하기 위하여, 예컨대 분류 모델의 형태로 인공 언어(artificial language)들을 생성한다. 비교 모듈(125), 분류 모듈(130), 및 검출 모듈(135)은 언어 검출을 제공하기 위해, 단독적으로 사용되거나 또는 다른 모듈과 결합되어 사용될 수 있다. 예를 들어, 비교 모듈(125) 및 분류 모듈(130)이 유사한 용어를 사용하는 언어들을 구별하기 위해 사용될 수 있다.

[0028] **반복 텍스트 검출(*Detecting Repetitive Text*)**

[0029] 반복 텍스트는 특정 언어들을 식별하는 텍스트의 시퀀스들에 대해 계산된 확률들을 왜곡하는 "노이즈"로 간주될 수 있다. 따라서 훈련 데이터 또는 언어가 검출될 입력 텍스트로부터 반복 텍스트를 제거하는 것이 유리할 수 있다.

[0030] 도 2a는 압축을 수행하여 반복을 검출하는 예시적인 프로세스(200)를 나타낸다. 편의상, 텍스트의 압축은 압축을 수행하는 시스템(예컨대, 도 1에 도시된 검출 시스템(100))에 관련되어 설명될 것이다. 시스템은 텍스트를 수신한다(202). 시스템은 텍스트의 부분들을 반복되지 않는 것으로서 식별한다.

[0031] 특히, 시스템은 텍스트의 제1 부분의 기본 데이터(underlying data)를 압축한다(204) . 예를 들어, 압축 모듈(105)이 입력 텍스트의 일부(예컨대, 입력 텍스트의 하나 이상의 라인 또는 단락)를 압축하기 위해 종래의 압축 기술을 사용할 수 있다. 예를 들어, 시스템은 무손실 데이터 압축 기술이나 손실 데이터 압축 기술을 사용할 수 있다. 기본 데이터를 압축한 후에, 시스템은 기본 데이터 압축량에 기초하여 데이터 압축 비율을 식별한다(206). 예를 들어, 압축 모듈(105)이 데이터 압축 비율을 식별할 수 있다.

[0032] 시스템은 데이터 압축 비율에 기초하여 텍스트의 제1 부분이 비반복인지 여부를 판단한다(208). 예를 들어, 압축 모듈(105)이 데이터 압축 비율을 임계값에 비교할 수 있다. 일반적으로, 텍스트는 압축될 수 있는 기본 데이터의 양이 증가할 때, 반복될 가능성이 더 높다. 따라서 사용되는 임계값은 수행되는 압축 기술에 따라 달라질 수 있다. 예를 들어, 무손실 데이터 압축 기술이 사용될 때는, 임계값이 손실 데이터 압축 기술이 사용될 때보다 낮은 값으로 설정될 수 있다. 일부 구현예들에서, 식별되는 데이터 압축 비율이 임계값보다 낮으면, 예를 들어 텍스트의 제1 부분이 반복적 텍스트를 표현하는 압축량보다 많이 압축될 수 없으면, 텍스트의 제1 부분이 비반복으로 판단된다. 마찬가지로, 식별된 데이터 압축 비율이 임계값이상이면, 텍스트의 제1 부분이 반복으로 판단된다.

[0033] 시스템은 비반복으로 판단된 텍스트의 부분들에 기초하여 언어 검출에서 사용하기 위한 후보 텍스트로서 텍스트의 제1 부분을 식별한다(210). 예를 들어, 분류 모듈(130)은 텍스트의 제1 부분이 비반복으로서 식별되면, 그 텍스트의 제1 부분을 언어 검출에서 사용하기 위한 후보 텍스트로서 식별할 수 있다. 일부 구현예들에서, 후보 텍스트는 언어들을 검출하는 하나 이상의 분류 모델들을 생성하기 위하여 사용될 수 있는 훈련 데이터로서 사용

될 수 있다. 일부 대안적 구현예들에서, 후보 텍스트는 하나 이상의 언어가 검출되는 입력 텍스트의 부분이다.

[0034] 일부 구현예들에서, 입력 텍스트의 고정-크기 블록(예컨대, 48바이트)은 예측 윈도우(prediction window)를 사용하여 분석된다. 예측 윈도우는 예를 들어 트라이그램을 따르는 다음 토큰을 예측하기 위하여, 트라이그램에 대해 12-비트 해쉬를 수행하기 위해 사용될 수 있다. 예측 윈도우는 정확한 예측(또는 반복)의 개수를 계산하기 위하여 텍스트에서 예를 들어, 한번에 하나의 토큰을 쉬프트하여 각 트라이그램에 걸쳐 쉬프트할 수 있다.

[0035] 도 2b는 반복 토큰을 포함하는 텍스트의 예시적 시퀀스를 나타낸다. 텍스트의 시퀀스는 토큰들 "X_1 X_2 X_3 X_4 X_1 X_2 X_5 X_1 X_6"을 나타낸다. 예를 들어, 각 토큰은 문자들을 나타낼 수 있다. 검출 시스템은 문자들의 시퀀스 "X_1 X_2 X_3 X_4 X_1 X_2 X_5 X_1 X_6"를 수신한다. 첫 번째 문자 X_1는 예를 들어, 해쉬 테이블과 같은 데이터 구조로 메모리에 저장될 수 있다. 또한 검출 시스템은 텍스트에 있는 첫 번째 문자 후에 바로 두번째 문자가 출현하였을 때, 첫 번째 문자와 두 번째 문자를 연관시킨다. 예를 들어, X_2가 X_1 후에 바로 출현하면, X_2는 X_1과 연관될 수 있다. 검출 시스템은 두 번째 문자가 이미 첫번째 문자에 연관되었을 때(예컨대, 두 번째 문자가 첫 번째 문자에 의해 예측됨), 첫 번째 문자에 이어 두 번째 문자가 오는 조합을 반복으로서 식별한다. 예를 들어, 문자들 "X_1 X_2"는 X_4의 출현 후에 반복되고, 반복으로서 검출된다.

[0036] 일부 구현예들에서, 첫 번째 문자(예컨대, X_1)는 첫 번째 문자 후에 바로 출현한 것으로 검출된 가장 최근 문자(the most recent character)에만 연관된다. 예를 들어, X_6은 X_1의 세 번째 출현 직후에 출현한다. 따라서 X_6은 X_1과 연관되며, X_2는 더이상 X_1과 연관되지 않는다. 그 결과로서, 문자 "X_1 X_2"의 다음 출현은 반복으로서 식별되지 않을 것이다. 오히려, "X_1, X_2"이 다음에 출현하면, X_1이 X_2에 연관될 것이다. 다시 말하면, 첫 번째 문자에 이어서 두 번째 문자와는 다른 제3 문자가 출현하기 전에 첫 번째 문자에 이어서 두 번째 문자가 오는 조합이 다시 출현하는 경우에만, 반복 문자들이 식별될 수 있다.

[0037] 일부 구현예들에서, 입력 텍스트의 고정-크기 블록에 대한 정확한 예측의 높은 비(예컨대, 60%)는 반복 텍스트를 나타내고, 입력 텍스트의 고정-크기 블록은 시스템이 언어 검출을 수행하기 전에 제거된다. 또 다른 구현예들에서, 텍스트는 예컨대, 상술되어진 기술을 수행하는 것에 의해서는 반복으로서 나타나지 않을 수도 있지만, 특정 언어에 의해 표현된 것으로서 낮은 신뢰도로 식별될 수 있다. 예를 들어, 텍스트로 표현될 가능성이 가장 높은 두 개의 언어의 확률들은 유사하거나, 또는 텍스트로 표현될 가능성이 가장 높은 언어에 대한 확률은 텍스트의 1/3 이하를 표현하는 것으로서 검출된다. 신뢰성이 낮고, 정확한 예측의 높은 비(예컨대, 50%)가 반복 텍스트를 나타낼 때, 그 반복 텍스트를 포함하는 단어는 단어 검출이 수행되기 전에 제거된다. 예를 들어, 도 2b에 도시된 바와 같이 X_4 후에 "X_1 X_2"의 출현은 언어 검출에서의 사용으로부터 제거될 수 있다.

[0038] 다른 구현예들도 있을 수 있다. 예를 들어, 반복 단어들은 신뢰성이 낮지 않을 때에도, 제거될 수 있다. 일부 구현예들에서, 반복 단어들이 삭제에 의해 제거될 수 있다. 또 다른 구현예들에서, 반복 단어들의 연관된 확률 추정치는 가중치를 이용하여 수정될 수 있다(예를 들어, 낮춰짐). 예를 들어, 제거를 위해, 연관된 확률 추정치에 0이 곱해질 수 있다. 다른 예시로서, 반복 단어를 완전하게 제거하지 않으면서 반복 단어들에 의해 발생하는 통계적 에러를 감소시키기 위해, 연관된 확률 추정치에 0과 1사이의 값을 갖는 가중치가 곱해질 수 있다.

[0039] **텍스트에서 세그먼트들을 식별(Identifying Segments in Text)**

[0040] 텍스트의 다른 부분들이 다른 쓰기 체계로 표현되기 때문에, 텍스트를 분할하는 것이 언어들을 검출할 때 유용할 수 있다. 이에 더하여, 특정 쓰기 체계로 된 부분의 다른 세그먼트들은 다른 언어로 표현될 수 있다. 예를 들어, 텍스트는 라틴어로 된 텍스트의 제1 부분과 키릴 문자로 된 텍스트의 제2 부분을 포함할 수 있다. 라틴어로 된 제1 부분은 영어와 스페인어를 나타내는 텍스트의 세그먼트들을 포함할 수 있다. 키릴 문자로 된 텍스트의 제2 부분은 불가리아어와 러시아어를 나타내는 텍스트의 세그먼트들을 포함할 수 있다. 텍스트의 제1 부분 또는 제2 부분이 중국어를 나타내는 세그먼트들을 포함할 가능성은 있을 수 없다. 또한, 텍스트의 제1 부분이 불가리아어를 나타내는 세그먼트들을 포함할 가능성도 있을 수 없다. 결과적으로, 쓰기 체계들에 의해 표현되는 언어들을 검출하기 전에, 텍스트에서 표현되는 쓰기 체계들을 먼저 검출하는 것이 유리할 수 있다.

[0041] 도 3은 쓰기 체계들과 언어들을 검출하기 위해 텍스트에서 세그먼트들을 식별하는 예시적 프로세스(300)를 나타낸다. 편의상, 세그먼트를 식별하는 것은 식별을 수행하는 시스템(예를 들어, 도 1에 도시된 검출 시스템(10

0))에 관련되어 설명될 것이다. 시스템은 텍스트를 수신한다(302). 예를 들어, 시스템은 입력 텍스트(예컨대, 텍스트 문서의 형식으로 됨)를 수신할 수 있다.

[0042] 시스템은 텍스트의 제1 부분에 표현된 쓰기 체계를 식별한다(304). 쓰기 체계는 하나 이상의 제1 언어를 표현한다. 예를 들어, 시스템은 쓰기 체계를 식별하기 위해, 종래 기술들, 본 명세서에서 설명된 기술들, 또는 그것들의 조합을 사용할 수 있다. 특정 예시로서, 쓰기 체계는 텍스트의 인코딩을 검출함으로써 식별될 수 있다.

[0043] 상술된 바와 같이, 쓰기 체계는 하나 이상의 언어에 상응할 수 있다. 시스템은 텍스트의 제1 부분에 표현된 하나 이상의 제1 언어에서만 특정 언어를 검출할 수 있다(306). 예를 들어, 시스템은 문서의 제1 부분을 식별할 수 있고, 이 제1 부분에 있는 텍스트의 상당한 양은 제1 쓰기 체계로 된 텍스트를 표현한다. 일반적으로, 입력 텍스트에 있는 각 문자는 특정 스크립트 또는 쓰기 체계에 속한다. 문자 테이블에 있는 문자들의 룩업은, 예를 들어 EUC-KR과 같은 입력 인코딩에 있는 문자들을 EUC-KR 문자 테이블에 있는 문자값에 매핑함으로써, 문자값과 쓰기 체계를 식별하기 위해 수행될 수 있다. 이 방식으로 각 문자를 매핑함으로써, 입력 텍스트의 인접 부분들이 식별될 수 있다.

[0044] 유사한 기술을 사용하여, 시스템은 또한 다른 쓰기 체계들을 식별하면서, 문서의 다른 부분(예컨대, 텍스트의 문단들 또는 라인들)을 식별할 수 있다. 다른 쓰기 체계로 된 텍스트를 표현하는 각 식별된 부분이 다른 쓰기 체계 각각에 상응하는 언어들을 식별하기 위해 별도로 처리될 수 있다. 예를 들어, 시스템은 문서의 제1 부분에 있는 하나 이상의 세그먼트(예컨대, 텍스트의 문단들 또는 라인들 내에 있는 문자들의 시퀀스들)을 식별할 수 있는데, 하나 이상의 세그먼트 각각에 있는 텍스트의 상당한 양은 제1 쓰기 체계의 언어로 표현되어 있다. 또한 시스템은 하나 이상의 세그먼트에 있는 텍스트의 상당한 양으로 표현되는 제1 쓰기 체계로 된 특정 언어를 검출할 수 있다. 예를 들어, 시스템은 문서 내 제1 문단이 라틴어로 표현되었다는 것을 식별할 수 있다. 시스템은 이어 제1 문단의 일부는 영어로 되어 있고, 제1 문단의 다른 일부는 프랑스어로 되어 있다는 것을 검출할 수 있다.

[0045] 다른 구현예들도 있을 수 있다. 일부 구현예들에서, 둘 이상의 쓰기 체계가 단일 쓰기 체계로 취급될 수 있다. 예를 들어, 중국어, 일본어, 및 한국어(CJK)를 표현하는 쓰기 체계는 언어 검출의 목적을 위해, 단일 쓰기 체계(예컨대, 병합된 쓰기 체계)로서 조합되고, 취급될 수 있다. 병합된 쓰기 체계을 사용하는 것은 둘 이상의 언어가 동일한 쓰기 체계으로부터의 문자들을 사용할 때, 유리할 수 있다. 특히, 중국어, 일본어, 및 한국어 각각은 한자(중국어 문자)를 사용한다. 일본어 텍스트가 한자 부분들, 가타카나 부분들, 및 히라가나 부분들로 분할된다면, 한자 부분들은 일본어보다는 중국어를 표현하는 것으로 잘못 식별될 수 있다. 예컨대 CJK를 위하여 병합된 쓰기 체계을 사용함으로써 부분들을 조합하는 것은, 섞여 있는 한자를 식별할 때 가타카나와 히라가나로부터의 콘텍스트가 고려될 수 있도록 하여, 일본어 식별에 이상적인 결과를 갖는다.

[0046] **에지들의 검출(Detecting Edges)**

[0047] 언어 검출을 개량하기 위한 다른 기술은 텍스트에 있는 한 언어를 다른 언어로 전환을 나타내는 에지를 검출하는 것에 관련된다. 특히, 점수들간의 변화량이 전환을 식별하기 위해 검출될 수 있다.

[0048] 도 4a는 제1 언어로 된 텍스트를 표현하는 토큰들의 제1 시퀀스에 이어서 제2 언어로 된 텍스트를 표현하는 토큰들의 제2 시퀀스를 포함하는 예시적 텍스트를 나타낸다. 특히, 텍스트 "hello bonjour"는 영어 단어 "hello"에 이어서 프랑스어 단어 "bonjour"(예컨대, 영어로 "hello")를 포함한다. 이 텍스트는 토큰들의 시퀀스 "h e l l o b o n j o u r"로서 표현될 수 있다. 점수들은 텍스트 내에 표현된 하나 이상의 언어들을 식별하기 위해 토큰들의 시퀀스에 있는 n-그램에 대해 계산될 수 있다.

[0049] 도 4b는 텍스트에 표현되는 다른 언어들 간의 에지들을 검출하기 위한 예시적 프로세스(400)를 나타낸다. 편의상, 에지를 검출하는 것은 검출을 수행하는 시스템(도 1에 도시된 검출 시스템(100))과 관련되어 설명될 것이다. 시스템은 텍스트를 수신한다(402). 시스템은 텍스트의 제1 세그먼트를 검출하는데, 제1 세그먼트의 상당한 양이 제1 언어를 표현한다(404). 시스템은 텍스트의 제2 세그먼트를 검출하는데, 제2 세그먼트의 상당한 양이 제2 언어를 표현한다(406). 예를 들어, 시스템은 도 1 내지 도 4b에 관련되어 상술된 기술들에 기초하여, 제1 언어를 표현하는 것으로 텍스트의 제1 세그먼트를 검출할 수 있고, 제2 언어를 표현하는 것으로 텍스트의 제2 세그먼트를 검출할 수 있다. 단지 예시적인 목적으로만, 시스템은 영어로 된 텍스트로서 "heool bon"을 처음에 식별하고, 프랑스어로 된 "jour"(영어로 "day")를 식별할 수 있다.

[0050] 시스템은 텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별한다(408). 도 4a에 도시된 바와 같이,

예를 들어, 사이즈 4(쿼디그램)의 n-그램에 대하여 점수들이 계산될 수 있다. 예를 들어, 쿼디그램들은 "hell", "ello", "llob", "lobo", "obon", "bonj", "onjo", "njou", 및 "jour"를 포함한다.

[0051] 일부 구현예들에서, 쿼디그램이 영어를 표현할 확률을 나타내는 각 쿼디그램에 대한 제1 점수가 계산된다. 이에 더하여, 쿼디그램이 프랑스어를 표현할 확률을 나타내는 각 쿼디그램에 대한 제2 점수가 식별된다. 예를 들어, "hell"이 영어를 표현할 확률을 나타내는 제1 점수 A가 식별될 수 있다. 이에 더하여, "hell"이 프랑스어를 표현할 확률을 나타내는 제2 점수 B가 식별될 수 있다. 제1 점수에서 제2 점수를 감산(예컨대, A-B)하여, "hell"에 대한 중간 점수를 생성할 수 있다. 중간 점수는 유사한 방식으로 쿼드그램들의 각각에 대하여 계산될 수 있다. 일부 구현예들에서, 프랑스어보다는 영어로 표현될 가능성이 더 있는 쿼디그램들은 플러스(positive) 중간 점수를 갖고, 영어보다는 프랑스어로 표현될 가능성이 높은 쿼디그램들은 마이너스(negative) 중간 점수를 갖는다.

[0052] 단일 n-그램에 대한 중간 점수들은 일반적으로 복수의 중간 점수들을 평균함으로써 제거될 수 있는 노이즈를 포함한다. 두 언어 사이의 전환을 나타내는 가장 가능성 있는 경계를 식별하기 위하여, 평균은 데이터를 평탄화(smooth)시킨다. 그것으로서, 다른 구현예에서는, 텍스트에 있는 시퀀스에서 출현한 쿼디그램들에 대하여 특정된 개수의 중간 점수들의 평균이 계산된다. 예를 들어, 특정된 개수가 4개이면, "hell", "ello", "llob", 및 "lobo"에 대한 중간 점수들의 평균이 계산된다. 이 예시에서, "(1) "hell", "ello", "llob", "iobo"; (2) "ello", "llob", "lobo", "obon"; (3) "llob", "lobo", "obon", "bonj"; (4)"lobo", "obon", "bonj", "onjo"; (5) "obon", "bonj", "onjo", "njou "; 및 (6)"bonj", "onjo", "njou", "jour"를 포함하는 쿼디그램들의 6개의 그룹에 대한 중간 점수의 평균이 계산된다. 텍스트에서 연속적으로 출현한 쿼디그램들의 2개 그룹에 대한 중간 점수 쌍 각각 간에 차이가 계산될 수 있다. 특히, 중간 점수들 간의 차이가 그룹 (1)과 (2), (2)와 (3), (3)과 (4), (4)와 (5), 및 (5)와 (6)에 대해서 계산될 수 있다.

[0053] 시스템은 점수들의 변화량에 기초하여 텍스트에서 제1 언어에서 제2 언어로의 전환을 식별하는 에지를 검출한다(410). 예를 들어, 점수들 간의 최대 차이가 에지를 검출하기 위해 사용될 수 있다. 이상적으로, 중간 점수들 간의 최대 차이는 "hello"와 "bonjour"간에 에지가 존재하는 것을 나타내는, 그룹들 (5)와 (6)에 상응할 것이다. 예를 들어, 에지는 쿼디그램들의 6개의 그룹에 대한 제1 점수만의 평균 간에 최대 변화량에 기초하여 식별될 수 있다.

[0054] **인공 언어(Artificial Language)**

[0055] 문학이나 신문과 같은 소스에서 발견되는 텍스트와는 달리, 웹 페이지로부터의 텍스트는 텍스트에서 어떤 자연 언어(예를 들어, 인간에 의해 말해지는 언어)가 표현되고 있는지에 대한 유용한 표시를 제공하지 않을 수 있다. 이러한 텍스트는,적어도 그것의 전체에서, 언어들을 검출하는 분류자를 훈련하기 위한 훈련 데이터로서 사용되면 안 된다. 예를 들어, "Copyright 2008"이 영어가 아닌 다른 언어로 쓰여진 웹 페이지에 출현한다. 그러므로 단어 "Copyright"은 언어들을 검출하기 위한 유용한 표시자가 되지 않을 것이다. 마찬가지로, 문자 "jpg"의 시퀀스(예컨대, 이미지 파일 형식에 대한 확장자를 나타냄)가 텍스트에 주기적으로 출현하고, 또한 언어들을 검출을 위한 유용한 표시를 제공하지 않는다. 실제로는, "copyright"와 "jpg"가 영어가 아닐 수 있는 특정 자연 언어에 속하는 것으로서 식별되어, 언어 검출 결과가 왜곡될 수 있다. 텍스트에서 언어들을 검출할 때 n-그램이 통계적 오류에 기여하지 않도록 하기 위해, 이러한 유형의 n-그램을 포함하는 인공 언어가 생성될 수 있다.

[0056] 도 5는 인공 언어를 생성하고, 그 인공 언어를 이용하여 언어들을 검출하는 예시적 프로세스(500)를 나타낸다. 편의상, 생성하는 과정과 검출하는 과정은 생성과 검출을 수행하는 시스템(예컨대, 도 1에 도시된 검출 시스템(100))에 관련하여 설명될 것이다. 시스템은 각각이 복수의 자연 언어를 식별하는 것에 대한 유사한 가능성에 연관되는 데이터를 훈련하는 하나 이상의 n-그램을 검출한다(502). 예를 들어, 시스템이 종래 기술, 본 명세서에서 상술된 기술들, 또는 그것들의 조합들을 사용하여 n-그램을 점수화할 수 있고, 둘 이상의 자연 언어(예컨대, 인간들이 말하는)를 식별하는 것에 대해 실질적으로 유사한 가능성을 갖는 하나 이상의 n-그램을 식별할 수 있다.

[0057] 시스템은 식별된 n-그램에 기초하여 인공 언어를 생성한다(504). 예를 들어, 시스템은 식별된 n-그램, n-그램이 인공 언어를 표현할 연관된 확률 추정치들, 및 개별적 총수를 포함하는 인공 언어들을 위한 분류 모델을 생성할 수 있다.

[0058] 일부 구현예들에서, 인공 언어들이 입력 텍스트에 의해 잠재적으로 표현될 언어로서, 자연적 언어처럼

처리된다. 예를 들어, 텍스트가 수신될 수 있다. 시스템은 수신된 언어가 인공 언어 또는 다른 자연 언어를 나타내는 텍스트를 포함하는지 여부를 검출할 수 있다. 특히, 시스템은 텍스트를 수신하고(506), 수신된 텍스트가 인공 언어를 표현할 제2 가능성과 비교하여 수신된 텍스트가 제1 자연 언어를 표현할 제1 가능성을 계산한다(508). 예를 들어, 시스템은 수신된 텍스트가 영어를 표현할 가능성 30%를 검출하고, 수신된 텍스트가 프랑스어를 표현할 가능성 40%를 검출하고, 수신된 텍스트가 인공 언어를 표현할 가능성 30%를 검출할 수 있다.

[0059] 예를 들어, 수신된 텍스트가 프랑스어 또는 다른 자연 언어와 비교하여 영어를 표현할 가능성을 표현하는 신뢰값(confidence value)을 식별하기 위해, 수신된 텍스트가 인공 언어를 표현할 가능성이 수신된 텍스트가 영어를 표현할 가능성과 비교될 수 있다.

[0060] 일부 구현예에서, 인공 언어를 표현하는 것으로 식별된 수신된 텍스트는 수정된 텍스트를 생성하기 위하여 그 수신된 텍스트로부터 제거될 수 있다. 상술된 바와 같이, 제거가 삭제에 의해 또는 가중치(예를 들어, 0 가중치)를 사용하여 연관된 확률 추정치들을 수정함으로써 수행될 수 있다. 그 결과, 시스템은 수정된 텍스트가 자연 언어들을 표현할 새로운 가능성을 검출한다. 예를 들어, 수정된 텍스트에 대한 두 번째 단계에서, 시스템은 영어에 대한 가능성 60%와 프랑스에 대한 가능성 40%를 검출할 수 있다.

[0061] **유사한 언어들(Similar Languages)**

[0062] 도 6a는 유사한 언어들 간을 구별하는 예시적 프로세스(600)를 나타낸다. 편의상, 유사한 언어들 간의 구별은 구별을 수행하는 시스템에 관련되어 설명될 것이다. 시스템(예컨대, 검출 시스템(100))은 텍스트를 수신한다(602). 시스템은 텍스트의 부분에 표현된 복수의 언어들(복수의 언어들의 각각은 실질적으로 유사함)을 검출한다(604). 예를 들어, 시스템은 종래 기술들, 상술되어진 기술들, 또는 그것들의 조합들을 사용하여 텍스트의 부분에서 표현된 복수의 언어(예컨대, 말레이시아어 및 인도네시아어와 같은 유사한 언어)들을 검출할 수 있다. 언어들은 그 언어들이 동일한 어족에 속할 때, 예를 들어, 또는 그들이 공통 언어학적 구조를 공유하면, 서로 실질적으로 유사하다고 고려될 수 있다. 유사한 언어들의 다른 예시로는 체코어와 슬로바키아어가 있다.

[0063] 일부 구현예들에서, 유사한 언어들은 둘 이상의 언어에서 자주 출현하는 특정 n-그램을 식별함으로써 식별될 수 있는데, 상기 특정 n-그램은 둘 이상의 언어들을 표현하는 것에 대해 실질적으로 유사한 가능성을 갖는다.

[0064] 시스템은 복수의 언어 중 제1 언어가 전체 텍스트를 표현할 제1 가능성을 식별한다(606). 예를 들어, 시스템은 말레이시아어가 전체 텍스트를 표현할 제1 가능성을 식별할 수 있다. 시스템은 복수의 언어 중 제2 언어가 전체 텍스트를 표현할 제2 가능성을 식별한다(608). 예를 들어, 시스템은 인도네시아어가 전체 텍스트를 표현할 제2 가능성을 식별할 수 있다. 시스템은 제1 가능성과 제2 가능성을 비교한다(610). 예를 들어, 시스템은 말레이시아어가 전체 텍스트를 표현할 가능성과 인도네시아어가 전체 텍스트를 표현할 가능성을 비교할 수 있다.

[0065] 시스템은 그 비교 결과에 기초하여 제1 언어로 표현되는 텍스트의 부분을 식별한다(612). 예를 들어, 말레이시아어가 전체 텍스트를 표현할 가능성이 인도네시아어가 전체 텍스트를 표현할 가능성보다 크면, 시스템은 텍스트의 부분이 말레이시아어로 표현된 것을 식별할 수 있다. 다른 구현예들도 있을 수 있다. 예를 들어, 제1 가능성과 제2 가능성은 전체 텍스트보다 적은 텍스트(예컨대, 복수의 언어들이 처음에 검출된 텍스트의 부분보다 큰 텍스트의 다른 부분에 기초함)에 기초하여 식별될 수 있다.

[0066] 또한, 언어들이 유사하더라도, 일부 구현예들에서는 유사한 언어들 간의 차이가 한 번에 더 많은 개수의 토큰들(예컨대, 8개의 토큰)을 처리함으로써 쉽게 식별될 수 있다. 한 번에 많은 수의 토큰을 처리하는 것이 모든 언어들에 대해 수행될 수 있긴 하지만, 많은 언어들이 더 적은 개수의 토큰(예컨대, 4개의 토큰)들에 대해 처리되는 동안 식별될 수 있기 때문에, 유사한 언어들에 대해서만 이 처리를 수행하는 것이 언어 검출의 효율성을 증가시킬 수 있다.

[0067] 예를 들어, 유사한 언어들은 크기 x의 n-그램 검사(examination)에 기초하여, 텍스트의 시퀀스를 잠재적으로 표현하는 것으로서 검출될 수 있다. n-그램의 크기를 검사된 크기가 보다 큰 y(y>x)로 증가시키는 것은 n-그램이 한 언어에서 하나 이상의 완전한 단어로 매핑될 가능성을 증가시키고, 이로써 다른 것들로부터 하나의 유사한 언어를 구별할 가능성을 증가시킨다. 예를 들어, "keuangan"는 인도네시아어로 될 높은 확률을 갖고, "kewangan"은 말레이시아어로 될 높은 확률을 갖지만, "keua", "uang", "ngan", "kewa", "wang", 및 "ngan"은 인도네시아어 또는 말레이시아어로 될 유사한 가능성을 갖는다.

[0068] 도 6b는 유사한 언어들을 구현하는 다른 예시적 프로세스(650)를 나타낸다. 편의상, 유사한 언어들 간의 구별은

구별을 수행하는 시스템에 관련하여 설명될 것이다. 시스템(예컨대, 검출 시스템(100))은 텍스트를 수신한다(652). 시스템은 텍스트를 크기 x의 n-그램으로 분할함으로써 텍스트에서 표현되는 제1 언어와 제2 언어를 검출한다(654). 예를 들어, 시스템은 크기 x(예컨대, 크기 4)의 n-그램을 사용하여 텍스트를 분석한다.

[0069] 시스템은 제1 언어가 제2 언어와 실질적으로 유사한지를 판단한다(656). 제1 언어가 제2 언어와 실질적으로 유사할 때, 시스템은 제1 언어가 제2 언어와 실질적으로 유사하다는 식별에 기초하여 텍스트에서 표현되는 특정 언어를 식별하기 위해, 텍스트를 크기 y(여기서, y>x)의 n-그램으로 분할함으로써, 그 텍스트를 처리한다(658). 예를 들어, 시스템은 크기 y(예컨대, 크기 8)의 n-그램을 사용하여 텍스트를 분석한다.

[0070] 다른 구현예들이 있을 수 있다. 예를 들어, 시스템이 대량의 훈련 데이터(예컨대, 수 백만의 웹 페이지)에 대하여 오직 한 언어를 지속적으로 식별할 때, n-그램의 크기는 감소될 수 있다.

[0071] 상술되어진 기술들은 런-타임동안, 예컨대 입력 텍스트를 수신한 것에 응답하여 실시간으로 수행되거나, 또는 그것들의 조합들로 오프라인으로 수행될 수 있다. 오프라인 기술들을 수행하는 예시로는 입력 텍스트에서 표현되는 언어들에 대한 식별에 사용하기 위한 훈련 데이터를 생성하는 과정이 포함된다. 런-타임 동안 이 기술을 수행하는 예시로는 반복적 부분들을 제거하기 위하여 입력 텍스트를 압축하는 과정, 남은 부분들을 분할하는 과정, 및 검출된 에지들에 기초하여 분할된 부분에 있는 언어들을 식별하는 과정들이 있다. 다른 구현예들도 가능하다.

[0072] 도 7은 일반적 컴퓨터 시스템(700)의 개략도이다. 시스템(700)은 상술된 기술들(예컨대, 프로세스들(200, 220, 300, 400, 500, 600, 및 650))과 연관되어 설명된 동작들을 실행하기 위해 사용될 수 있다. 본 시스템(700)은 프로세서(710), 메모리(720), 저장 디바이스(730), 및 입력/출력 디바이스(740)를 포함할 수 있다. 컴포넌트(710, 720, 730 및 740) 각각은 예컨대 시스템 버스(750)를 이용하여 상호 접속될 수 있다. 프로세서(710)는 시스템(700) 내에서의 실행을 위한 명령어들을 처리할 수 있다. 일 구현예에서, 프로세서(710)는 싱글 쓰레드(Single-threaded) 프로세서이다. 이와 같이 실행되는 명령들은 예를 들어, 도 1 내지 6b를 참조하여 설명된 것처럼, 언어들을 검출하는 컴퓨터의 하나 이상의 구성요소들과 구현될 수 있다. 일 구현예에서, 프로세서(710)은 단일-쓰레드(single-threaded) 프로세서이다. 다른 구현예에서, 프로세서(710)는 멀티 쓰레드(Multi-threaded) 프로세서이다. 프로세서(710)는 입력/출력 디바이스(740) 상의 사용자 인터페이스에 그래픽 정보를 디스플레이하기 위하여, 메모리(720) 혹은 저장 디바이스(730) 상에 저장된 명령들을 처리할 수 있다.

[0073] 메모리(720)는 예를 들어, 시스템(700) 내에서 정보를 저장하는 휘발성 메모리 또는 비휘발성 메모리를 포함하는 컴퓨터 판독가능 매체이다. 메모리(720)는 예를 들어, 분류 모델들을 저장할 수 있다. 저장 디바이스(730)는 본 시스템(700)에 대한 지속성 저장부를 제공할 수 있다. 저장 디바이스(730)는 플로피 디스크 디바이스, 하드 디스크 디바이스, 광학 디스크 디바이스, 테입 디바이스, 또는 다른 적합한 지속적 저장 수단일 수 있다. 입력/출력 디바이스(740)는 시스템(700)에 대한 입/출력 동작들을 제공한다. 일 구현예에서, 입력/출력 디바이스(740)는 키보드 및/또는 포인팅 디바이스를 포함한다. 다른 구현예에서, 입력/출력 디바이스(740)는 그래픽 사용자 인터페이스를 디스플레이하는 디스플레이 유닛을 포함한다.

[0074] 입력/출력 디바이스(740)는 시스템(예컨대, 도 1에 도시된 검출 시스템(100))에 대한 입력/출력 동작들을 제공할 수 있다. 검출 시스템(100)은 예를 들어, 모듈들(105, 110, 115, 120, 125)를 구현하는 컴퓨터 소프트웨어 구성요소들을 포함할 수 있다. 몇 가지만 예로 들자면, 이러한 소프트웨어 구성요소들은 저장 디바이스(7230), 메모리(720) 내에 존재하거나, 네트워크 연결을 통해 얻어질 수 있다.

[0075] 본 명세서에 기재된 요지와 기능적 동작들의 실시예들은 디지털 전자 회로로 구현되거나, 또는 상세한 설명에 기재된 구조 및 그들의 구조적 등가물을 포함하는 컴퓨터 소프트웨어, 펌웨어, 또는 하드웨어로 구현되거나, 또는 이들 중 하나 이상의 조합으로 구현될 수 있다. 본 명세서에 기재된 요지의 실시예들은 하나 이상의 컴퓨터 프로그램 제품, 즉, 데이터 프로세싱 장치에 의해 실행되거나 또는 그 장치의 동작을 제어하도록, 컴퓨터 저장 미디어에 부호화된 컴퓨터 프로그램 명령의 하나 이상의 모듈로서 구현될 수 있다. 컴퓨터 저장 매체는 컴퓨터-판독가능 저장 디바이스, 컴퓨터 판독가능 저장 기판(substrate), 랜덤 또는 시리얼 액세스 메모리 어레이 또는 디바이스, 이들 중 하나 이상의 조합일 수 있다.

[0076] 본 명세서에서 설명된 동작들은 하나 이상의 컴퓨터 판독가능 매체 디바이스에 저장되거나 다른 소스로부터 수신된 데이터에 대해 데이터 처리 장치에 의해 동작들이 실행됨으로써 구현될 수 있다.

[0077] "데이터 프로세싱 장치"라는 용어는 데이터를 처리하기 위한 모든 장치, 디바이스 및 기계를 포괄하며, 예를 들어, 프로그래머블 프로세서, 컴퓨터, 또는 다중 프로세서 또는 컴퓨터를 포함한다. 이 장치는 또한 하드웨어 외

에도, 당해 컴퓨터 프로그램에 대한 실행 환경을 생성하는 코드를 포함하고, 코드는 예를 들어, 프로세서 펌웨어, 프로토콜 스택, 데이터베이스 관리 시스템, 운영 시스템, 또는 이들 중 하나 이상의 조합을 구성한다.

[0078] 컴퓨터 프로그램(프로그램, 소프트웨어, 소프트웨어 애플리케이션, 스크립트 또는 코드로도 알려짐)은 컴파일 또는 인터프리터 언어나 선언적 또는 절차적 언어를 포함하는 모든 형태의 프로그래밍 언어로 작성될 수 있으며, 독립형 프로그램이나 모듈, 컴포넌트, 서브루틴 또는 컴퓨터 환경에서 사용하기에 적합한 그 밖의 유닛을 포함하는 임의의 형태로도 배치될 수 있다. 컴퓨터 프로그램은 파일 시스템의 파일에 반드시 상응해야 하는 것은 아니다. 프로그램은 다른 프로그램 또는 데이터를 보유하는 파일의 일부에 저장되거나(예를 들어, 마크업 언어 문서 내에 저장되는 하나 이상의 스크립트), 당해 프로그램 전용의 단일 파일에 저장되거나, 또는 다수의 조화된(coordinated) 파일들(예를 들어, 하나 이상의 모듈, 서브프로그램, 코드의 부분을 저장하는 파일)에 저장될 수 있다. 컴퓨터 프로그램은 하나의 컴퓨터에서, 또는 한 위치에 배치되거나 또는 다수의 위치에 걸쳐서 분산되고 통신 네트워크에 의해 접속된 다수의 컴퓨터에서 실행되도록 배치될 수 있다.

[0079] 본 명세서에 설명된 프로세스와 논리 흐름은 하나 이상의 프로그래머블 프로세서에 의해 수행될 수 있고, 이 프로그래머블 프로세서는 입력 데이터에 작용하여 출력을 생성함으로써 기능들을 수행하는 하나 이상의 컴퓨터 프로그램들을 실행한다. 예를 들어, FPGA(field programmable gate array) 또는 ASIC(application specific integrated circuit)과 같은 전용 논리 회로가 프로세스와 논리 흐름을 수행하거나, 장치를 구현할 수 있다.

[0080] 컴퓨터 프로그램의 실행에 적합한 프로세서에는, 예를 들어, 범용 및 전용 마이크로프로세서, 및 임의 종류의 디지털 컴퓨터 중 하나 이상의 프로세서가 있다. 일반적으로, 프로세서는 판독 전용 메모리(ROM), 또는 랜덤 액세스 메모리(RAM), 또는 양자로부터 명령과 데이터를 수신한다. 컴퓨터의 필수 구성요소는 명령들을 실행하는 프로세서, 및 명령과 데이터를 저장하는 하나 이상의 메모리 디바이스이다. 일반적으로, 컴퓨터는 데이터를 저장하기 위한 하나 이상의 대용량 저장 디바이스(예를 들어, 자기 디스크, 광자기 디스크, 또는 광디스크)를 포함하거나, 또는 이 디바이스와 데이터를 송수신하기 위하여 동작적으로(operatively) 결합될 수 있다. 하지만 컴퓨터는 이러한 디바이스를 반드시 구비할 필요는 없다. 더욱이, 컴퓨터는 예를 들어, 모바일 전화기, 개인 정보 단말(PDA), 모바일 오디오 또는 비디오 재생기, 게임 콘솔, GPS(global positioning system) 수신기 등과 같은 다른 디바이스에 내장될 수 있다.

[0081] 컴퓨터 프로그램 명령어와 데이터를 저장하기 적합한 컴퓨터-판독가능 미디어에는, 예를 들어, 반도체 메모리 디바이스(예를 들어, EPROM, EEPROM, 플래시 메모리 디바이스); 자기 디스크(예를 들어, 내부 하드디스크, 착탈식 디스크); 광자기 디스크; 및 CD ROM과 DVD-ROM 디스크를 포함하는 모든 형태의 비휘발성 메모리, 매체 및 메모리 디바이스가 포함된다. 프로세서와 메모리는 전용 논리 회로에 의해 보완되거나 또는 전용 논리 회로에 통합될 수 있다.

[0082] 사용자와의 상호동작을 제공하기 위하여, 본 명세서에 기술된 요지의 실시예는, 정보를 사용자에게 디스플레이하기 위한 디스플레이 디바이스(예를 들어, CRT(cathode ray tube) 또는 LCD(liquid crystal display) 모니터), 키보드 및 포인팅 디바이스(예를 들어, 마우스 또는 트랙볼)를 구비한 컴퓨터에 구현될 수 있다. 사용자는 키보드와 포인팅 디바이스를 이용하여 컴퓨터에 입력을 제공할 수 있다. 사용자와의 상호동작을 제공하기 위하여 다른 종류의 디바이스가 또한 사용될 수 있다. 예를 들어, 사용자에게 제공되는 피드백(feedback)은 예를 들어, 시각 피드백, 청각 피드백 또는 촉각 피드백인 임의 형태의 감각 피드백일 수 있고, 사용자로부터의 입력은 음향, 음성 또는 촉각 입력을 포함하는 임의의 형태로 수신될 수 있다.

[0083] 본 명세서에 기술된 요지의 실시예는, 예를 들어, 데이터 서버와 같은 백엔드(back-end) 구성요소를 구비하는 컴퓨팅 시스템; 또는 예를 들어, 애플리케이션 서버와 같은 미들웨어 구성요소를 구비하는 컴퓨팅 시스템; 또는 예를 들어, 사용자가 본 명세서에 기술된 요지의 구현예와 상호동작할 수 있는 그래픽 사용자 인터페이스 또는 웹브라우저를 구비한 클라이언트 컴퓨터와 같은 프론트엔드(front-end) 구성요소를 구비하는 컴퓨터 시스템; 또는 이러한 백엔드, 미들웨어 또는 프론트엔드 구성요소들의 임의 조합을 구비하는 컴퓨팅 시스템으로 구현될 수 있다. 시스템의 구성요소는 디지털 데이터 통신의 임의 형태 또는 매체(예를 들어, 통신 네트워크)에 의해 상호접속될 수 있다. 통신 네트워크의 예에는 근거리 네트워크(LAN)와 인터넷과 같은 광역 네트워크(WAN)가 포함된다.

[0084] 컴퓨팅 시스템은 클라이언트와 서버를 포함할 수 있다. 클라이언트와 서버는 보통 서로 떨어져 있으며, 일반적으로는 통신 네트워크를 통하여 상호동작한다. 클라이언트와 서버의 관계는 각각의 컴퓨터상에서 실행되고 상호 클라이언트-서버 관계를 갖는 컴퓨터 프로그램에 의하여 발생한다.

[0085] 본 명세서가 다수의 특정한 구현 세부사항을 포함하고 있지만, 이는 임의 구현예의 범위나 청구할 사항의 범위에 대한 어떠한 제약으로서도 이해되어서는 안 되며, 특정 구현예들의 특정한 실시예에 고유할 수 있는 특징의 설명으로서 이해되어야 한다. 별개의 실시예의 문맥으로 본 명세서에서 설명된 소정 특징은 조합되어 단일 실시예로 구현될 수 있다. 반대로, 단일 실시예의 문맥에서 설명한 다양한 특징은 복수의 실시예에서 별개로 구현되거나 어떤 적당한 하위 조합으로서도 구현 가능하다. 또한, 앞에서 특징이 소정 조합에서 동작하는 것으로서 설명되고 그와 같이 청구되었지만, 청구된 조합으로부터의 하나 이상의 특징은 일부 경우에 해당 조합으로부터 삭제될 수 있으며, 청구된 조합은 하위 조합이나 하위 조합의 변형으로 될 수 있다.

[0086] 마찬가지로, 도면에서 특정한 순서로 동작을 묘사하고 있지만, 그러한 동작이 바람직한 결과를 얻기 위해, 도시한 특정 순서나 순차적인 순서로 수행되어야 한다거나, 설명한 모든 동작이 수행되어야 한다는 것을 의미하는 것은 아니다. 소정 환경에서, 멀티태스킹 및 병렬 프로세싱이 바람직할 수 있다. 또한, 상술한 실시예에 있어서 다양한 시스템 구성요소의 분리는 모든 실시예에서 그러한 분리를 요구하는 것으로 이해되어서는 안 되며, 설명한 프로그램 구성요소와 시스템은 단일 소프트웨어 제품으로 통합되거나 또는 복수의 소프트웨어 제품으로 패키지될 수 있다는 점을 이해되어야 한다.

[0087] 본 명세서에서 설명한 요지의 특정 실시예가 기술되었다. 그 밖의 실시예는 후술하는 청구범위 내에 속한다. 예를 들어, 청구항에 인용된 동작들은 상이한 순서로 수행될 수 있지만, 여전히 바람직한 결과를 달성한다. 일 실시예로서, 첨부 도면에서 묘사된 프로세스들은, 바람직한 결과를 얻기 위해, 도시된 특정 순서나 순차적인 순서를 반드시 요구하는 것은 아니다. 소정 구현예에서, 멀티태스킹과 병렬 프로세싱이 효과적일 수 있다.

부호의 설명

[0088] 100: 검출 시스템

105: 압축 모듈

110: 분할 모듈

115: 에지 검출 모듈

120: 언어 생성 모듈

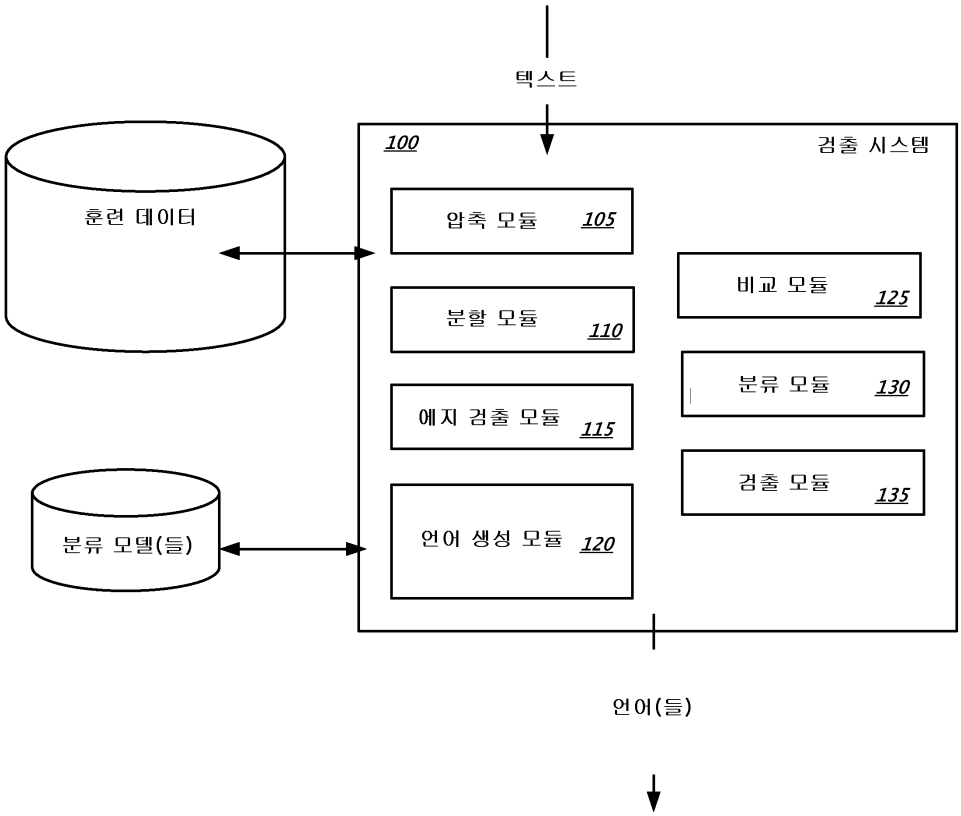
125: 비교 모듈

130: 분류 모듈

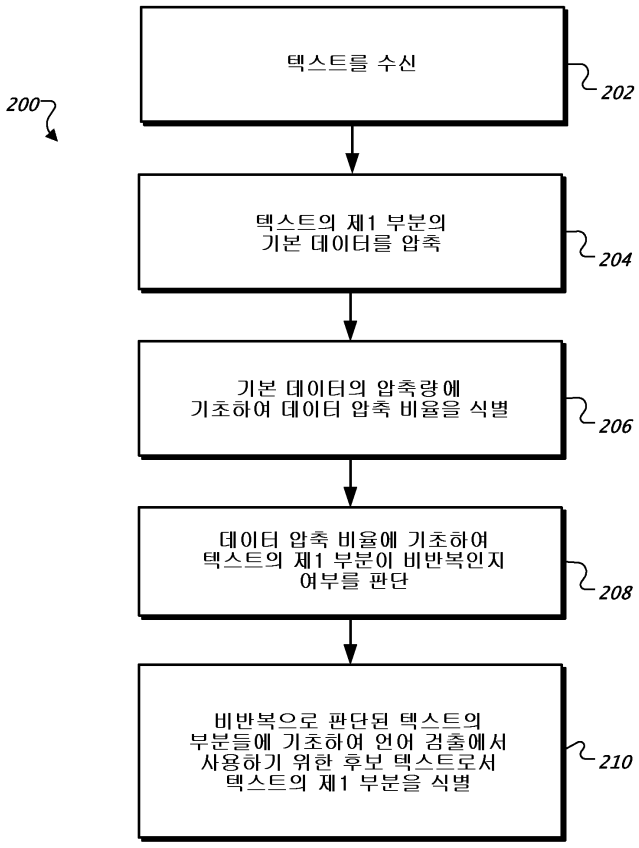
135: 검출 모듈

도면

도면1


텍스트
100
검출 시스템
훈련 데이터
압축 모듈 105
분할 모듈 110
에지 검출 모듈 115
언어 생성 모듈 120
비교 모듈 125
분류 모듈 130
검출 모듈 135
분류 모델(들)
언어(들)

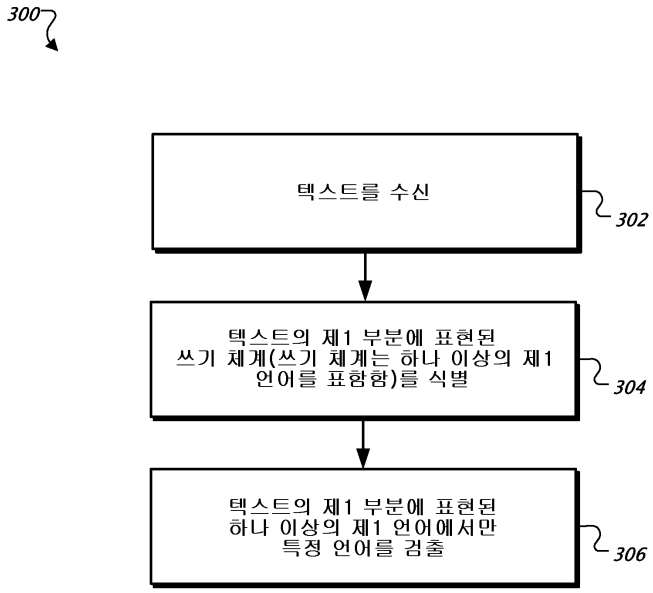
도면2a



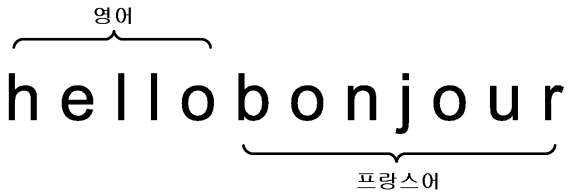
도면2b

$$X_1\ X_2\ X_3\ X_4\ X_1\ X_2\ X_5\ X_1\ X_6$$

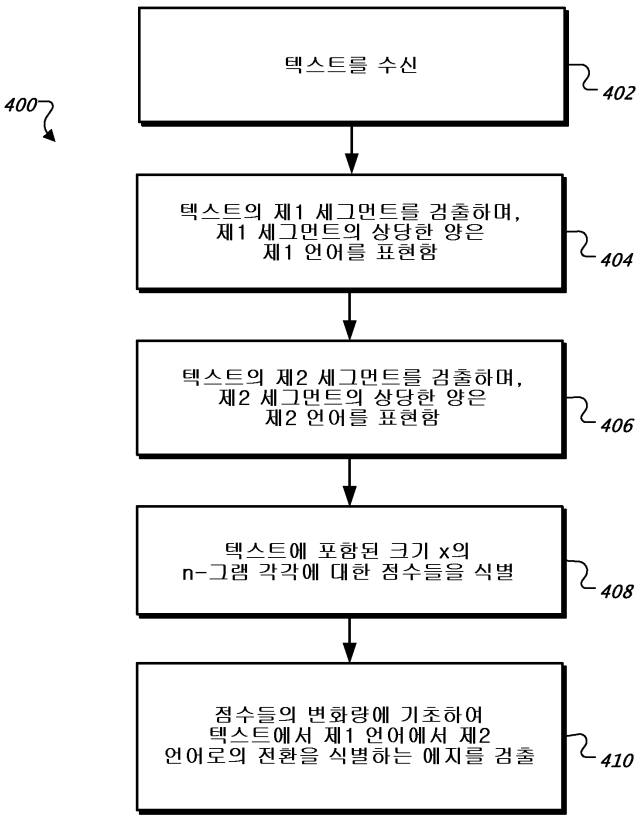
도면3


300
텍스트를 수신
302
텍스트의 제1 부분에 표현된 쓰기 체계(쓰기 체계는 하나 이상의 제1 언어를 표함함)를 식별
304
텍스트의 제1 부분에 표현된 하나 이상의 제1 언어에서만 특정 언어를 검출
306

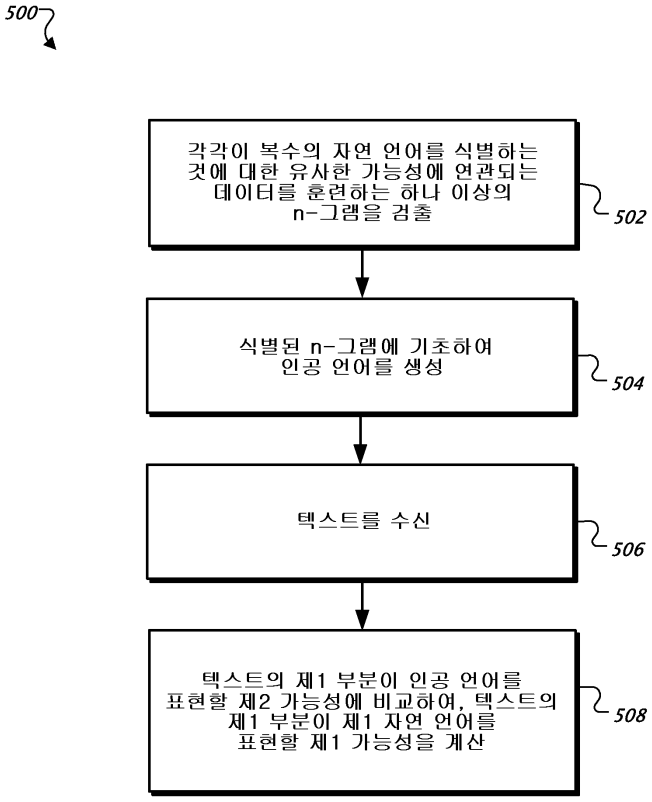
도면4a


영어
hellobonjour
프랑스어

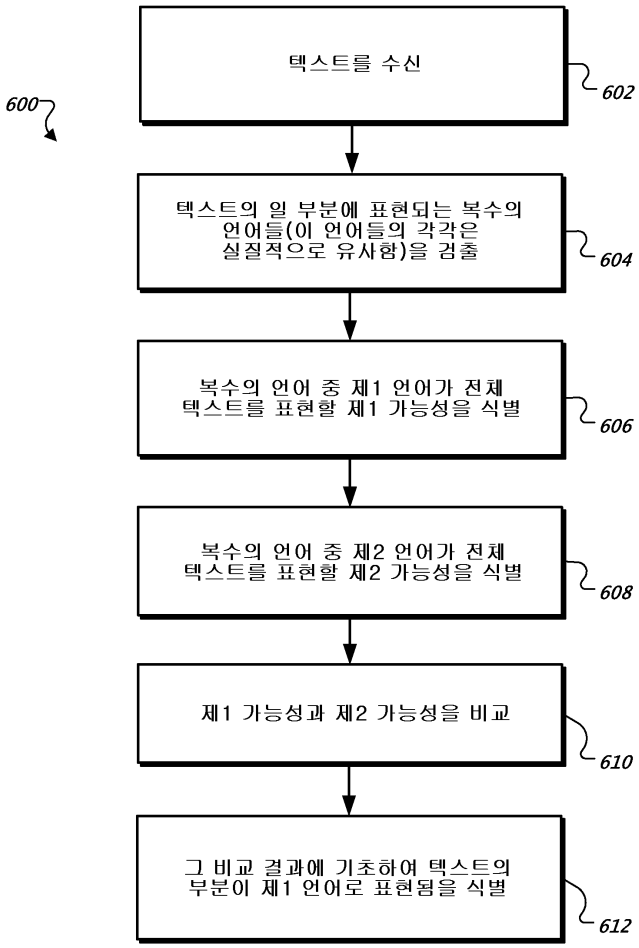
도면4b


400
텍스트를 수신
402
텍스트의 제1 세그먼트를 검출하며, 제1 세그먼트의 상당한 양은 제1 언어를 표현함
404
텍스트의 제2 세그먼트를 검출하며, 제2 세그먼트의 상당한 양은 제2 언어를 표현함
406
텍스트에 포함된 크기 x의 n-그램 각각에 대한 점수들을 식별
408
점수들의 변화량에 기초하여 텍스트에서 제1 언어에서 제2 언어로의 전환을 식별하는 에지를 검출
410

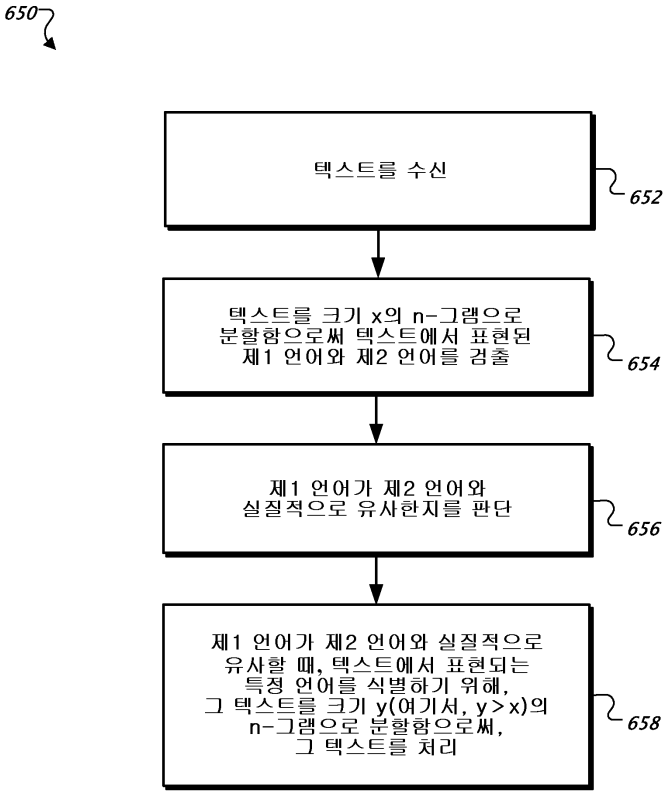
도면5


500
각각이 복수의 자연 언어를 식별하는 것에 대한 유사한 가능성에 연관되는 데이터를 훈련하는 하나 이상의 n-그램을 검출
502
식별된 n-그램에 기초하여 인공 언어를 생성
504
텍스트를 수신
506
텍스트의 제1 부분이 인공 언어를 표현할 제2 가능성에 비교하여, 텍스트의 제1 부분이 제1 자연 언어를 표현할 제1 가능성을 계산
508

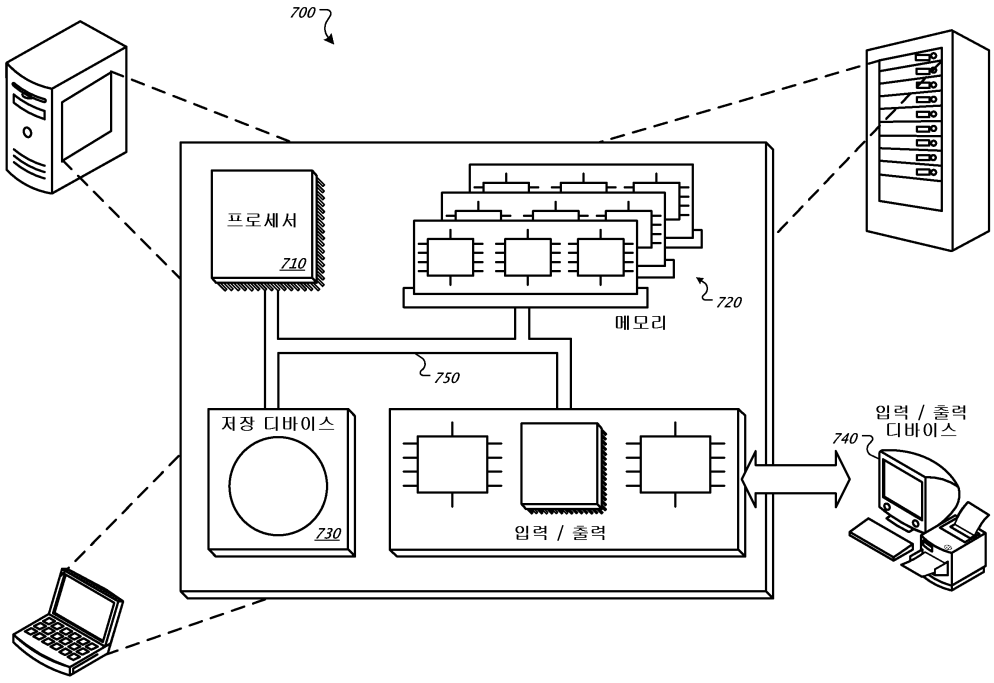
도면6a


600
텍스트를 수신
602
텍스트의 일 부분에 표현되는 복수의 언어들(이 언어들의 각각은 실질적으로 유사함)을 검출
604
복수의 언어 중 제1 언어가 전체 텍스트를 표현할 제1 가능성을 식별
606
복수의 언어 중 제2 언어가 전체 텍스트를 표현할 제2 가능성을 식별
608
제1 가능성과 제2 가능성을 비교
610
그 비교 결과에 기초하여 텍스트의 부분이 제1 언어로 표현됨을 식별
612

도면6b


650
텍스트를 수신
652
텍스트를 크기 x의 n-그램으로 분할함으로써 텍스트에서 표현된 제1 언어와 제2 언어를 검출
654
제1 언어가 제2 언어와 실질적으로 유사한지를 판단
656
제1 언어가 제2 언어와 실질적으로 유사할 때, 텍스트에서 표현되는 특정 언어를 식별하기 위해, 그 텍스트를 크기 y(여기서, y>x)의 n-그램으로 분할함으로써, 그 텍스트를 처리
658

도면7


700
프로세서
710
720
메모리
750
저장 디바이스
730
입력 / 출력
입력 / 출력 디바이스
740