(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년08월01일
(11) 등록번호 10-1292404
(24) 등록일자 2013년07월26일

(51) 국제특허분류(Int. Cl.)
G06F 17/20 (2006.01) *G06F 17/27* (2006.01)
G06F 17/30 (2006.01)
(21) 출원번호 10-2007-7024524
(22) 출원일자(국제) 2006년03월14일
심사청구일자 2011년02월23일
(85) 번역문제출일자 2007년10월24일
(65) 공개번호 10-2008-0003364
(43) 공개일자 2008년01월07일
(86) 국제출원번호 PCT/US2006/009147
(87) 국제공개번호 WO 2006/115598
국제공개일자 2006년11월02일
(30) 우선권주장
11/113,612 2005년04월25일 미국(US)
(56) 선행기술조사문헌
US6889361 B1
US6424983 B1
전체 청구항 수 : 총 21 항

(73) 특허권자
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
포터, 더글라스 더블유.
미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이
하트, 주니어, 에드워드 씨.
미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이
(뒷면에 계속)
(74) 대리인
제일특허법인

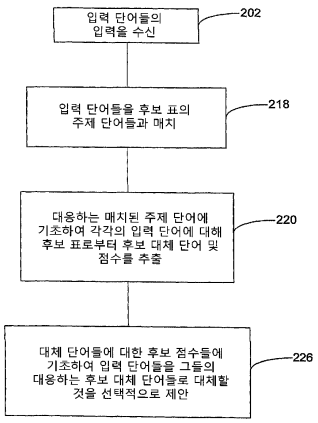
심사관 : 박태식

(54) 발명의 명칭 **철자 제안을 생성하기 위한 방법 및 시스템**

(57) 요 약

문자열의 단어들의 대체 단어들을 제안하는 컴퓨터로 구현되는 방법이 제공된다. 이 방법에서, 입력 단어들의 입력 문자열이 수신된다. 그리고, 입력 단어들은 후보 표의 주제 단어들에 매치된다. 다음에, 매치된 주제 단어들에 대응하는 후보 표로부터의 후보 대체 단어 및 점수들이 추출된다. 각각의 점수는 입력 단어가 대응하는 후보 대체 단어로 대체되어야 할 확률을 나타낸다. 마지막으로, 입력 단어들을 그들의 대응하는 후보 대체 단어들로 대체하는 것이 대체 단어들에 대한 점수들에 기초하여 선택적으로 제안된다. 본 발명의 다른 한 양태는 방법을 구현하기 위해 구성된 철자 검사 시스템에 관한 것이다.

대 표 도 - 도2



(72) 발명자

이가라시, 히사까즈

미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이

쉬미드, 패트리샤 엠.

미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이

람지, 윌리암 디.

미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이

특허청구의 범위

청구항 1

2개 이상의 입력 단어들을 포함하는 입력 문자열을 수신하는 단계;

컴퓨터 저장 매체에 저장된 명령어들을 프로세서에 의해 실행하는 단계에 응답하여 상기 입력 단어들 중 하나의 입력 단어에 대한 대체 단어를 제안하는 단계 - 상기 제안하는 단계는,

상기 입력 단어들 중 하나의 입력 단어에 대한 후보 대체 단어들을 획득하는 단계,

각각의 후보 대체 단어에 대한 후보 점수를 획득하는 단계 - 각각의 후보 점수는, 상기 입력 단어가 대응하는 후보 대체 단어로 대체되어야 할 확률을 나타냄 - ,

각각의 후보 대체 단어에 대한 후보 대체 문자열을 생성하는 단계 - 각각의 후보 대체 문자열은 상기 후보 대체 단어, 및 상기 입력 문자열에서 상기 후보 대체 단어에 대응하는 상기 입력 단어를 뺀 입력 단어들을 포함함 - ,

통계 데이터에 기초하여 후보 대체 문자열들의 각각에 대한 확률 점수를 생성하는 단계,

상기 후보 대체 단어에 대한 상기 후보 점수 및 상기 후보 대체 단어를 포함하는 상기 후보 대체 문자열에 대한 상기 확률 점수에 기초하여 상기 후보 대체 단어들의 각각에 대한 최종 점수를 계산하는 단계, 및

상기 후보 대체 단어들의 최종 점수들에 기초하여 상기 입력 단어를 그에 대응하는 후보 대체 단어들 중 하나로 대체할 것을 제안하는 단계를 포함함 -

를 포함하는 철자 검사 방법.

청구항 2

제1항에 있어서,

상기 입력 단어들 중 하나의 입력 단어에 대한 후보 대체 단어들을 획득하는 단계는,

상기 입력 단어를 후보 테이블의 주제 단어들(subject words)에 매칭시키는 단계, 및

매칭된 주제 단어에 대응하는 하나 이상의 후보 대체 단어들을 상기 후보 테이블로부터 획득하는 단계를 포함하는 철자 검사 방법.

청구항 3

제2항에 있어서,

상기 확률 점수들은 상기 후보 대체 문자열의 단어들이 함께 나타날 가능성에 기초하는 철자 검사 방법.

청구항 4

제2항에 있어서,

상기 후보 테이블의 대응하는 후보 대체 단어들 및 상기 주제 단어들은 각각 어휘 목록에서 단어들을 식별하는 어휘 목록 식별자의 형태인 철자 검사 방법.

청구항 5

제2항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 임계치를 충족하는 그들의 대응하는 주제 단어들로부터의 편집 거리(edit distance)를 갖는 단어들을 포함하는 철자 검사 방법.

청구항 6

제2항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들과 유사한 의미를 갖는 단어들을 포함하는 철자 검사 방법.

청구항 7

제2항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들에 대해 음성학적으로 매치하는 단어들을 포함하는 철자 검사 방법.

청구항 8

제2항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들을 통상적으로 교정한 단어들을 포함하는 철자 검사 방법.

청구항 9

철자 검사 시스템으로서,

프로세서를 포함하는 컴퓨터;

입력 문자열의 입력 단어들에 대한 대체 단어들을 제안하기 위한, 상기 프로세서에 의해 실행가능한 명령어들을 포함하는 프로그램 모듈들을 포함하는 컴퓨터 판독가능 기록 매체를 포함하고, 상기 모듈들은,

상기 입력 단어들 중 후보 테이블의 주제 단어와 매칭되는 하나의 입
력 단어에 대한 하나 이상의 후보 대체 단어들 및 대응하는 후보 점수
들의 출력을 포함하는 후보 발생기 - 각각의 후보 점수는 상기 입력
단어가 대응하는 후보 대체 단어로 대체되어야 할 확률을 나타냄 - ,
하나 이상의 후보 대체 단어들의 출력을 수신하고 각각의 후보 대체
단어에 대한 후보 대체 문자열을 생성하는 컨텍스추얼 철자 엔진
(contextual spelling engine) - 각각의 후보 대체 문자열은 후보
대체 단어, 및 상기 입력 문자열에서 상기 후보 대체 단어에 대응하는
상기 입력 단어를 뺀 입력 단어들을 포함함 - , 및
후보 대체 문자열들의 각각에 대한 확률 점수들의 출력을 갖는 언어
모델 - 상기 확률 점수들은 통계 데이터에 기초함 -

을 포함하며,

상기 컨텍스추얼 철자 엔진은 상기 후보 대체 단어에 대한 후보 점수 및 상기 후보 대체 단어를 포함하는 상기 후보 대체 문자열에 대한 확률 점수에 기초하여 상기 후보 대체 단어들의 각각에 대한 최종 점수를 계산하고,

최종 점수들에 기초하여 상기 후보 대체 단어들 중 하나를 출력하는 철자 검사 시스템.

청구항 10

제9항에 있어서,

상기 모듈들은 상기 후보 발생기의 출력과 상기 컨텍스추얼 철자 엔진의 출력 중 하나로부터 제외된 후보 대체 단어들의 목록을 포함하는 후보 제외 테이블을 포함하는 철자 검사 시스템.

청구항 11

제9항에 있어서,

상기 확률 점수들은 상기 후보 대체 문자열의 단어들이 함께 나타날 가능성에 기초하는 철자 검사 시스템.

청구항 12

제9항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 임계치를 충족하는 그들의 대응하는 주제 단어들로부터의 편집 거리를 갖는 단어들을 포함하는 철자 검사 시스템.

청구항 13

제9항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들과 유사한 의미를 갖는 단어들을 포함하는 철자 검사 시스템.

청구항 14

제9항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들에 대해 음성학적으로 매치하는 단어들을 포함하는 철자 검사 시스템.

청구항 15

제9항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들을 통상적으로 교정한 단어들을 포함하는 철자 검사 시스템.

청구항 16

2개 이상의 입력 단어들의 오리지널(original) 입력 문자열을 수신하는 단계 - 상기 입력 단어들 중 적어도 하나는 철자에 오류가 있음 - ;

컴퓨터 저장 매체에 저장된 명령어들을 프로세서에 의해 실행하는 단계에 응답하여, 상기 입력 단어들 중 하나에 대한 대체 단어를 제안하는 단계 - 상기 제안하는 단계는,

상기 오리지널 입력 문자열 중 철자에 오류가 있는 입력 단어들을

교정함으로써, 올바른 철자의 입력 단어들만을 포함하는 교정된 입력

문자열을 생성하는 단계,

상기 교정된 입력 문자열의 상기 입력 단어들을 후보 테이블의 주제

단어들에 매칭시키는 단계,

각각이 상기 매칭된 주제 단어들에 대응하는 후보 대체 단어들 및

대응하는 후보 점수들을 상기 후보 테이블로부터 추출하는 단계 - 각

각의 후보 점수는, 상기 교정된 입력 문자열의 상기 입력 단어가 대응하는 후보 대체 단어로 대체되어야 할 확률을 나타냄 - ,

각각의 후보 대체 단어에 대한 후보 대체 문자열을 생성하는 단계 - 각각의 후보 대체 문자열은 상기 후보 대체 단어, 및 상기 교정된 입력 문자열에서 상기 후보 대체 단어에 대응하는 상기 입력 단어를 뺀 입력 단어들을 포함함 - ,

통계 데이터에 기초하여 후보 대체 문자열들의 각각에 대한 확률 점수들을 생성하는 단계,

상기 후보 대체 단어에 대한 상기 후보 점수 및 상기 후보 대체 단어를 포함하는 상기 후보 대체 문자열에 대한 상기 확률 점수에 기초하여 상기 후보 대체 단어들의 각각에 대한 최종 점수를 계산하는 단계, 및

상기 후보 대체 단어들에 대한 최종 점수들에 기초하여 상기 교정된 입력 문자열의 입력 단어들을 그에 대응하는 후보 대체 단어들로 대체할 것을 선택적으로 제안하는 단계를 포함함 -

를 포함하는 철자 검사 방법.

청구항 17

제16항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 임계치를 충족하는 그들의 대응하는 주제 단어들로부터의 편집 거리를 갖는 단어들을 포함하는 철자 검사 방법.

청구항 18

제16항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들과 유사한 의미를 갖는 단어들을 포함하는 철자 검사 방법.

청구항 19

제16항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들에 대해 음성학적으로 매치하는 단어들을 포함하는 철자 검사 방법.

청구항 20

제16항에 있어서,

상기 후보 테이블의 상기 후보 대체 단어들은 그들의 대응하는 주제 단어들을 통상적으로 교정한 단어들을 포함하는 철자 검사 방법.

청구항 21

제16항에 있어서,

상기 확률 점수들은 상기 후보 대체 문자열의 단어들이 함께 나타날 가능성에 기초하는 철자 검사 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 일반적으로 철자 검사 방법 및 시스템에 관한 것이며, 좀더 자세하게는, 입력 문자열의 단어들에 대한 대체 단어를 입력 문자열의 단어에 기초하여 제안하도록 구성된 철자 검사 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 문서에서 워드 프로세싱 애플리케이션을 이용하여 발생된 것 등과 같은 텍스트 입력들은 철자 오류를 포함하는 다양한 종류의 많은 오류들을 포함할 수 있다. 무효한 단어로 귀결되는 철자 오류들은 일반적으로 어휘 목록 기반(lexicon-based) 철자 검사기에 의해 처리될 수 있다. 그러한 철자 오류들은 단어의 철자의 무지 또는 오타로 인해 발생할 수 있다.

[0003] 어휘 목록 기반 철자 검사기들은 텍스트 입력에 있는 단어들을 단어들의 어휘 목록에 비교하고 어휘 목록에서 발견되지 않는 텍스트 입력의 단어들을 식별해낸다. 철자 오류된 단어에 대해 흔히 하나 이상의 대체 단어들이 제안된다. 예를 들어, "fly frm Boston"라는 텍스트 입력에서 철자 검사기는 "frm"을 철자 오류된 것으로 식별해낼 것이다.

[0004] 종래의 철자 검사 애플리케이션들을 이용하면 일반적으로 탐지 가능하지 않은 다른 종류의 철자 오류들은 유효한 단어로 귀결된다. 실례로, 의도된 단어의 철자의 무지 또는 오타의 결과로서 워드 프로세싱 애플리케이션의 사용자가 의도하지 않은 유효한 단어를 입력할 수 있다. 예를 들어, "fly form Boston"이라는 텍스트 입력에서, "form"이라는 단어는, 그 단어가 의도된 단어 "from"의 철자 오류일지라도, 종래의 철자 검사 애플리케이션들에 의해 플래그(flag)되지 않을 유효한 단어이다. 이러한 종류의 철자 오류들의 교정은 일반적으로 단어가 이용되는 문맥의 분석을 요구한다.

[0005] 종래의 철자 검사 애플리케이션들은 일반적으로 편집 거리(edit distance)에 기초하여 식별되는 무효한 단어들을 위해 대체단어들을 제안하는 것이다. 편집 거리는 유효한 대안적 단어를 형성하기 위해 요구되는 변화를 나타낸다. 어휘 목록 내에서 타이핑된 무효한 단어로부터 최단 편집 거리를 갖는 단어가 사용자에게 제안되는 제1 대체 단어이다. 예를 들어, "fly frm Boston"라는 구에서 대부분의 철자 검사 애플리케이션들은, 제안할 때 단어의 문맥을 고려하지 않기 때문에, 올바른 단어 "from"을 제안하기 전에 "form"을 대체 단어로서 제안할 것이다. 철자 오류에 대해 가장 적절한 대체 단어를 제안하기 위해서는, 철자 오류가 발견되는 문맥의 분석이 이루어져야 한다.

[0006] 따라서, 철자 오류된 단어에 대한 더 좋은 제안, 및 부적절하게 이용되는 유효한 단어들의 개선된 탐지를 제공하기 위해 단어들이 이용되는 문맥을 분석할 수 있는 개선된 철자 검사 방법 및 시스템에 대한 요구가 있다.

[0007] 본 발명의 실시예들은 상기 및 기타의 문제들에 대한 해결책을 제공하고, 종래 기술을 능가하는 다른 이점을 제공한다.

발명의 상세한 설명

[0008] 본 발명은 일반적으로 주제 단어와 후보 대체 단어 쌍 및 각각의 쌍에 대한 후보 점수를 포함하며 단어 대체 제안들의 기초가 되는 후보 표를 이용하는 철자 검사 방법 및 시스템에 관한 것이다.

[0009] 본 발명의 한 양태는 문자열의 단어들에 대한 대체 단어들을 제안하는 컴퓨터로 구현되는 방법에 관한 것이다. 이 방법에서는, 입력 단어들의 입력 문자열이 수신된다. 그 후, 입력 단어들은 후보 표의 주제 단어들에 매치된다. 다음에, 매치된 주제 단어들에 대응하는 후보 표로부터의 후보 대체 단어 및 후보 점수들이 추출된다. 각각의 후보 점수는 입력 단어가 대응하는 후보 대체 단어로 대체되어야 할 확률을 나타낸다. 마지막으로, 입력 단어들을 그들의 대응하는 후보 대체 단어들로 대체하는 것이 대체 단어들에 대한 후보 점수들에 기초하여 선택적으로 제안된다.

[0010] 본 발명의 다른 한 양태는 입력 문자열의 입력 단어들에 대한 대체 단어들을 제안하는 철자 검사 시스템에 관한 것이다. 이 시스템은 후보 발생기 및 컨텍스추얼(contextual) 철자 엔진을 포함한다. 후보 발생기는 입력 단어들의 각각에 대해 후보 표의 주제 단어에 매치하는 후보 대체 단어 및 대응하는 후보 점수를 출력한다. 각각의 후보 점수는 입력 단어들이 대응하는 후보 대체 단어로 대체되어야 할 확률을 나타낸다. 컨텍스추얼 철자 엔진은 대응하는 후보 점수들에 기초하여 입력 단어들에 대한 후보 대체 단어들을 선택적으로 출력한다.

[0011] 본 발명의 또다른 한 양태는 입력 문자열의 입력 단어들의 대체 단어들을 제안하기 위해 철자 검사 시스템에서

이용되는 후보 표를 형성하는 방법에 관한 것이다. 이 방법에서는, 단어들의 어휘 목록이 제공된다. 다음에, 어휘 목록의 주제 단어들이 어휘 목록의 다른 단어들에 비교된다. 그 후, 후보 대체 단어들은 비교에 기초하여 주제 단어들에 대해 식별된다. 그 후, 식별된 주제 단어들과 그들의 대응하는 후보 대체 단어들의 쌍들을 포함하는 후보 표가 형성된다. 마지막으로, 후보 표가 컴퓨터 판독가능 매체에 저장된다.

[0012] 본 발명의 다른 실시예들의 특징을 이루는 특징 및 이점들은 아래의 상세한 설명을 읽고 관련 도면을 살펴보면 명백해질 것이다.

실 시 예

[0017] 본 발명은 일반적으로 무효한 입력 문자열의 입력 단어들에 대한 정확한 대체 단어 제안들을 제공하는 철자 검사 방법 및 시스템에 관한 것이다. 또한, 본 발명의 철자 검사 방법 및 시스템은 부적절하게 이용되는 입력 문자열들의 유효한 입력 단어들에 대한 대체 단어 제안들을 제공할 수 있다. 본 발명의 실시예들은 입력 단어들이 이용되는 문맥에 기초하여 대체 단어들을 제안하고 있다.

[0018] 본 발명을 상세하게 기술하기 전에, 본 발명이 이용될 수 있는 예시적 컴퓨팅 환경들의 설명이 제공될 것이다.

[0019] **예시적 컴퓨팅 환경**

[0020] 도 1은 본 발명이 구현되기에 적합한 컴퓨팅 시스템 환경(100)의 일례를 도시하고 있다. 컴퓨팅 시스템 환경(100)은 적합한 컴퓨팅 환경의 일례에 불과하며, 본 발명의 용도 또는 기능성의 범위에 관해 어떤 제한을 암시하고자 하는 것이 아니다. 컴퓨팅 환경(100)이 예시적인 운영 환경(100)에 도시된 컴포넌트들 중 임의의 하나 또는 그 컴포넌트들의 임의의 조합과 관련하여 어떤 의존성 또는 요구사항을 갖는 것으로 해석되어서는 안된다.

[0021] 본 발명은 많은 기타 범용 또는 특수 목적의 컴퓨팅 시스템 환경 또는 구성에서 동작할 수 있다. 본 발명에서 사용하는 데 적합할 수 있는 잘 알려진 컴퓨팅 시스템, 환경 및/또는 구성의 예로는 퍼스널 컴퓨터, 서버 컴퓨터, 핸드-헬드 또는 랩톱 장치, 멀티프로세서 시스템, 마이크로프로세서 기반 시스템, 셋톱 박스, 프로그램가능한 가전제품, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터, 상기 시스템들이나 장치들 중 임의의 것을 포함하는 분산 컴퓨팅 환경, 기타 등등이 있지만 이에 제한되는 것은 아니다.

[0022] 본 발명은 일반적으로 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터 실행가능 명령어와 관련하여 기술될 것이다. 일반적으로, 프로그램 모듈은 특정 태스크를 수행하거나 특정 추상 데이터 유형을 구현하는 루틴, 프로그램, 개체, 컴포넌트, 데이터 구조 등을 포함한다. 본 발명은 또한 통신 네트워크를 통해 연결되어 있는 원격 처리 장치들에 의해 태스크가 수행되는 분산 컴퓨팅 환경에서 실시되도록 설계된다. 분산 컴퓨팅 환경에서, 프로그램 모듈은 메모리 저장 장치를 비롯한 로컬 및 원격 컴퓨터 저장 매체 둘다에 위치할 수 있다.

[0023] 도 1과 관련하여, 본 발명을 구현하는 예시적인 시스템은 컴퓨터(110) 형태의 범용 컴퓨팅 장치를 포함한다. 컴퓨터(110)의 컴포넌트들은 처리 장치(120), 시스템 메모리(130), 및 시스템 메모리를 비롯한 각종 시스템 컴포넌트들을 처리 장치(120)에 연결시키는 시스템 버스(121)를 포함하지만 이에 제한되는 것은 아니다. 시스템 버스(121)는 메모리 버스 또는 메모리 컨트롤러, 주변 장치 버스 및 각종 버스 아키텍처 중 임의의 것을 이용하는 로컬 버스를 비롯한 몇몇 유형의 버스 구조 중 어느 것이라도 될 수 있다. 예로서, 이러한 아키텍처는 ISA(industry standard architecture) 버스, MCA(micro channel architecture) 버스, EISA(Enhanced ISA) 버스, VESA(video electronics standard association) 로컬 버스, 그리고 메자닌 버스(mezzanine bus)로도 알려진 PCI(peripheral component interconnect) 버스 등을 포함하지만 이에 제한되는 것은 아니다.

[0024] 컴퓨터(110)는 통상적으로 각종 컴퓨터 판독가능 매체를 포함한다. 컴퓨터(110)에 의해 액세스 가능한 매체는 그 어떤 것이든지 컴퓨터 판독가능 매체가 될 수 있고, 이러한 컴퓨터 판독가능 매체는 휘발성 및 비휘발성 매체, 이동식 및 비이동식 매체를 포함한다. 예로서, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 포함하지만 이에 제한되는 것은 아니다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보를 저장하는 임의의 방법 또는 기술로 구현되는 휘발성 및 비휘발성, 이동식 및 비이동식 매체를 포함한다. 컴퓨터 저장 매체는 RAM, ROM, EEPROM, 플래시 메모리 또는 기타 메모리 기술, CD-ROM, DVD(digital versatile disk) 또는 기타 광 디스크 저장 장치, 자기 카세트, 자기 테이프, 자기 디스크 저장 장치 또는 기타 자기 저장 장치, 또는 컴퓨터(110)에 의해 액세스되고 원하는 정보를 저장할 수 있는 임의의 기타 매체를 포함하지만 이에 제한되는 것은 아니다. 통신 매체는 통상적으로 반송파(carrier wave) 또는 기타 전송 메커니즘(transport mechanism)과 같은 피변조 데이터 신호(modulated data signal)에 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터 등을 구현하고 모든 정보 전달 매체를 포함한다.

"피변조 데이터 신호"라는 용어는, 신호 내에 정보를 인코딩하도록 그 신호의 특성들 중 하나 이상을 설정 또는 변경시킨 신호를 의미한다. 예로서, 통신 매체는 유선 네트워크 또는 직접 배선 접속(direct-wired connection)과 같은 유선 매체, 그리고 음향, RF, 적외선, 기타 무선 매체와 같은 무선 매체를 포함한다. 상술된 매체들의 모든 조합이 또한 컴퓨터 판독가능 매체의 영역 안에 포함되는 것으로 한다.

[0025] 시스템 메모리(130)는 판독 전용 메모리(ROM)(131) 및 랜덤 액세스 메모리(RAM)(132)와 같은 휘발성 및/또는 비휘발성 메모리 형태의 컴퓨터 저장 매체를 포함한다. 시동 중과 같은 때에, 컴퓨터(110) 내의 구성요소들 사이의 정보 전송을 돕는 기본 루틴을 포함하는 기본 입/출력 시스템(BIOS)(133)은 통상적으로 ROM(131)에 저장되어 있다. RAM(132)은 통상적으로 처리 장치(120)가 즉시 액세스 할 수 있고 및/또는 현재 동작시키고 있는 데이터 및/또는 프로그램 모듈을 포함한다. 예로서, 도 1은 운영 체제(134), 애플리케이션 프로그램(135), 기타 프로그램 모듈(136) 및 프로그램 데이터(137)를 도시하고 있지만 이에 제한되는 것은 아니다.

[0026] 컴퓨터(110)는 또한 기타 이동식/비이동식, 휘발성/비휘발성 컴퓨터 저장매체를 포함한다. 단지 예로서, 도 1은 비이동식·비휘발성 자기 매체에 기록을 하거나 그로부터 판독을 하는 하드 디스크 드라이브(141), 이동식·비휘발성 자기 디스크(152)에 기록을 하거나 그로부터 판독을 하는 자기 디스크 드라이브(151), CD-ROM 또는 기타 광 매체 등의 이동식·비휘발성 광 디스크(156)에 기록을 하거나 그로부터 판독을 하는 광 디스크 드라이브(155)를 포함한다. 예시적인 운영 환경에서 사용될 수 있는 기타 이동식/비이동식, 휘발성/비휘발성 컴퓨터 기억 매체로는 자기 테이프 카세트, 플래시 메모리 카드, DVD, 디지털 비디오 테이프, 고상(solid state) RAM, 고상 ROM 등이 있지만 이에 제한되는 것은 아니다. 하드 디스크 드라이브(141)는 통상적으로 인터페이스(140)와 같은 비이동식 메모리 인터페이스를 통해 시스템 버스(121)에 접속되고, 자기 디스크 드라이브(151) 및 광 디스크 드라이브(155)는 통상적으로 인터페이스(150)와 같은 이동식 메모리 인터페이스에 의해 시스템 버스(121)에 접속된다.

[0027] 위에서 설명되고 도 1에 도시된 드라이브들 및 이들과 관련된 컴퓨터 저장 매체는, 컴퓨터(110)를 위해, 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 및 기타 데이터를 저장한다. 도 1에서, 예를 들어, 하드 디스크 드라이브(141)는 운영 체제(144), 애플리케이션 프로그램(145), 기타 프로그램 모듈(146), 및 프로그램 데이터(147)를 저장하는 것으로 도시되어 있다. 여기서 주의할 점은 이들 컴포넌트가 운영 체제(134), 애플리케이션 프로그램(135), 기타 프로그램 모듈(136), 및 프로그램 데이터(137)와 동일하거나 그와 다를 수 있다는 것이다. 이에 관해, 운영 체제(144), 애플리케이션 프로그램(145), 기타 프로그램 모듈(146) 및 프로그램 데이터(147)에 다른 번호가 부여되어 있다는 것은 적어도 이들이 다른 사본(copy)이라는 것을 나타내기 위한 것이다.

[0028] 사용자는 키보드(162), 마이크(163) 및 마우스, 트랙볼(trackball) 또는 터치 패드와 같은 포인팅 장치(161) 등의 입력 장치를 통해 명령 및 정보를 컴퓨터(110)에 입력할 수 있다. 다른 입력 장치(도시 생략)로는 마이크, 조이스틱, 게임 패드, 위성 안테나, 스캐너 등을 포함할 수 있다. 이들 및 기타 입력 장치는 종종 시스템 버스에 결합된 사용자 입력 인터페이스(160)를 통해 처리 장치(120)에 접속되지만, 병렬 포트, 게임 포트 또는 USB(universal serial bus) 등의 다른 인터페이스 및 버스 구조에 의해 접속될 수도 있다. 모니터(191) 또는 다른 유형의 디스플레이 장치도 비디오 인터페이스(190) 등의 인터페이스를 통해 시스템 버스(121)에 접속될 수 있다. 모니터 외에, 컴퓨터는 스피커(197) 및 프린터(196) 등의 기타 주변 출력 장치를 포함할 수 있고, 이들은 출력 주변장치 인터페이스(195)를 통해 접속될 수 있다.

[0029] 컴퓨터(110)는 원격 컴퓨터(180)와 같은 하나 이상의 원격 컴퓨터로의 논리적 접속을 사용하여 네트워크화된 환경에서 동작할 수 있다. 원격 컴퓨터(180)는 또 하나의 퍼스널 컴퓨터, 핸드-헬드 장치, 서버, 라우터, 네트워크 PC, 피어 장치 또는 기타 통상의 네트워크 노드일 수 있고, 통상적으로 컴퓨터(110)와 관련하여 상술된 구성요소들의 대부분 또는 그 전부를 포함한다. 도 1에 도시된 논리적 접속으로는 LAN(171) 및 WAN(173)이 있지만, 기타 네트워크를 포함할 수도 있다. 이러한 네트워킹 환경은 사무실, 전사적 컴퓨터 네트워크(enterprise-wide computer network), 인트라넷, 및 인터넷에서 일반적인 것이다.

[0030] LAN 네트워킹 환경에서 사용될 때, 컴퓨터(110)는 네트워크 인터페이스 또는 어댑터(170)를 통해 LAN(171)에 접속된다. WAN 네트워킹 환경에서 사용될 때, 컴퓨터(110)는 통상적으로 인터넷과 같은 WAN(173)을 통해 통신을 설정하기 위한 모뎀(172) 또는 기타 수단을 포함한다. 내장형 또는 외장형일 수 있는 모뎀(172)은 사용자 입력 인터페이스(160) 또는 기타 적절한 메커니즘을 통해 시스템 버스(121)에 접속된다. 네트워크화된 환경에서, 컴퓨터(110) 또는 그의 일부와 관련하여 기술된 프로그램 모듈은 원격 메모리 저장 장치에 저장될 수 있다. 예로서, 도 1은 원격 애플리케이션 프로그램(185)이 원격 컴퓨터(180)에 있는 것으로 도시하고 있지만 이에 제한되는 것은 아니다. 도시된 네트워크 접속은 예시적인 것이며 이 컴퓨터들 사이에 통신 링크를 설정하는 기타 수

단이 사용될 수 있다는 것을 이해할 것이다.

[0031] 본 발명은 도 1과 관련하여 도시된 바와 같은 컴퓨터 시스템에서 수행될 수 있음을 주목해야 한다. 그러나, 본 발명은 서버, 메시지 처리에 사용되는 컴퓨터 또는 본 발명의 다양한 부분들이 분산 컴퓨팅 시스템의 다양한 부분에서 수행되는 분산 시스템에서 수행될 수 있다.

[0032] **철자 검사 방법 및 시스템**

[0033] 앞서 말했듯이, 본 발명은 일반적으로 무효한 입력 문자열의 입력 단어들에 대한 정확한 대체 단어 제안들을 제공하는 철자 검사 방법 및 시스템에 관한 것이다. 또한, 본 발명의 철자 검사 방법 및 시스템은 부적절하게 이용되는 입력 문자열의 유효한 입력 단어들을 위한 대체 단어 제안들을 제공할 수 있다. 본 발명에 의해 제공된 대체 단어 제안들은 일반적으로 입력 단어들이 이용되는 문맥에 의존한다.

[0034] 또한, 본 발명은 철자 검사에 대한 데이터 구동적 접근 방법을 제공한다. 결과적으로, 철자 검사 방법 및 시스템의 실시예들은 품사 또는 규칙 기반 문법 검사기들과 동일한 종류의 이론적 언어학 지식을 요구하지 않는다. 구현도 규칙 기반 시스템들보다 한층 더 단순하며 유지 및 보안 비용이 더 저렴하다. 또한, 제안들을 발생시키는 메커니즘은 일반적으로 언어 종속적이며 다수의 언어들로 용이하게 스케일할 수 있다.

[0035] 본 발명의 실시예들은 도 2 및 3을 참조하여 설명될 수 있다. 도 2는 본 발명의 실시예들에 따른 문자열의 단어들을 위한 대체 단어들을 제안하는 방법을 예시하는 플로우차트이다. 도 3은 본 발명의 실시예들에 따른 방법을 구현하게 구성된 철자 검사 시스템(200)의 블록 다이어그램이다.

[0036] 이 방법의 단계 202에서는, 입력 단어들의 입력 문자열(204)이 철자 검사 시스템(200)에 의해 수신된다. 입력 문자열(204)은 처음에 키보드, 마이크로폰(즉, 구술된) 또는 다른 통상의 방법에 의해 워드 프로세싱 애플리케이션(206)의 사용자가 입력할 수 있다. 대안적으로, 입력 문자열(204)은 기존의 문서, 웹 페이지 또는 다른 소스로부터 검색될 수 있다.

[0037] 애플리케이션(206)이 입력 문자열(204)을 시스템(200)에 완전한 문장들로 제공하는 것이 바람직하다. 또한, 입력 문자열(204)은 철자 검사 시스템(200)에 토큰화된 형태 또는 다른 인식할 수 있는 포맷으로 제공될 수 있으며, 그렇지 않으면 시스템(200)에 의해 상기 포맷으로 변환될 수 있다.

[0038] 입력 문자열은 시스템(200)의 컨텍스추얼 철자 엔진(210)으로부터 후보 발생기(208)로 제공된다. 입력 문자열(204)은 유효한(즉, 올바르게 철자된) 단어들만을 포함하는 것이 바람직하다.

[0039] 이 방법의 한 실시예에 따라, 입력 문자열(204)의 철자 오류된 입력 단어들은 어휘 목록 기반 철자 검사기(212)를 이용하여 교정된다. 철자 검사기(212)는 입력 문자열(204)의 각각의 입력 단어를 어휘 목록(214)의 단어들과 비교한다. 어휘 목록(214)에 포함되지 않은 입력 단어들에 대해 제안되는 대체들은 철자 검사기(212)가 통상의 방법에 따라 발생시킨다. 철자 검사기(212)가 발생키시는 제안된 대체들은 사용자가 철자 검사기(212)에 의해 식별된 철자 오류된 단어들에 대한 대체를 선택하도록 제공될 수 있으며, 그에 따라 입력 문자열들이 변경된다. 대안적으로, 철자 검사기(212)에 의해 철자 오류된 입력 단어들에 대해 발생되는 제안된 교정들은 유효한 단어들만을 포함하는 입력 문자열을 형성하는 후보 발생기(208)에 제공된다. 그 후, 유효한 단어들만을 포함하는 입력 문자열이 후보 발생기(208)에 의해 분석된다.

[0040] 후보 발생기(208)의 목적은 입력 문자열(204)의 입력 단어들과 유사하거나 아니면 관련된 후보 대체 단어들을 식별시키려는 것이다. 후보 대체 단어들은 나중에 입력 단어들에 대해 제안된 대체 단어들로 애플리케이션(206)에 제공될 수 있다. 후보 대체 단어 및 입력 단어 쌍들은 후보 표(216)에 포함되어 있다.

[0041] 표 1은 본 발명의 실시예들에 따른 후보 표(216)의 일부의 예시이다. 후보 표(216)는 후보 대체 단어에 각각 결합되어 있는 주제 단어들 및 주제 단어가 후보 대체 단어로 대체되어야 하는 확률을 나타내는 후보 점수를 포함한다.

표 1

[0042] **예시적 후보 표**

주제 단어	후보 대체 단어	후보 점수	편집
aback	alack	0.543	b:l
aback	back	0.023	a:

abalones	abalone’s	0.870	A
abandoned	abandoner	0.765	d:r
break	brake	0.689	H

[0043] 후보 표(216)의 한 실시예는 후보 대체 단어들을 형성하기 위해 주제 단어에 대해 수행되어야 할 변환을 기술하는 각각의 주제 단어와 후보 대체 단어 쌍을 위한 편집 입력을 포함한다. 예를 들어, "aback"을 "alack"으로 변환하기 위해 "aback"의 "b"가 "l"로 변화되어야 하며, 그것은 "b:l"로 나타내어진다. 마찬가지로, 후보 대체 단어 "back"을 형성하기 위해 "aback"에서 첫번째 "a"를 삭제하는 것은 "a:"로 나타내어질 수 있다. Abalone’s에 대한 “’ ” 의 부가는 단순히 "A"로 나타내어질 수 있다. 주제 단어 "break"에 대한 "brake" 등과 같은 동음이의어 후보 대체 단어들은 "H"로 나타내어진다. 주제 단어를 대응하는 후보 대체 단어로 변환하기 위해 수행되어야 할 다양한 편집들을 식별하는 다른 방법도 이용될 수 있다.

[0044] 후보 표의 편집 입력은 컨텍스추얼 철자 엔진(210)이 입력 문자열의 분석 과정에서 후보 대체 단어들에 대한 점수를 변경하기 위해 이용할 수 있다. 편집 입력들은 클래스별로 그루핑될 수 있고 편집 종류들의 클래스의 빈도 등과 같은 것을 반영하는 상이한 값들이 할당될 수 있다. 값이 높을수록, 어떤 후보 대체 단어의 점수에 대한 그 편집 종류의 영향이 크다. 이러한 클래스 또는 편집 종류 점수들은 후보 점수에 더해지거나 포함될 수 있다. 예를 들어, 후보 대체 단어를 형성하기 위해 주제 단어의 첫문자를 삭제하는 편집은 후보 대체 단어에 대한 점수를 증가시키는 것으로 귀결될 수 있다.

[0045] 본 발명의 한 실시예에 따라, 후보 표(216)가 컴퓨터 판독가능 매체에 바이너리 파일로 저장되어 있고, 그것은 철자 검사 시스템(200)이 신속하게 액세스하기 위해 컴퓨팅 환경의 메모리에 로드된다. 본 발명의 한 실시예에 따라, 후보 표(216)는 바이너리 파일에 해시 표로서 저장된다.

[0046] 후보 표(216)의 다른 한 실시예에 따라, 어휘 목록 식별자들이 주제 단어와 후보 대체 단어들을 식별시키기 위해 이용된다. 어휘 목록 식별자들은 도 3에 도시된 어휘 목록(214) 등과 같은 어휘 목록의 대응하는 단어들에 연결할 수 있게 한다. 본 발명의 이 실시예는 큰 후보 표(216)를 저장하기 위해 요구되는 메모리의 양을 줄이도록 작동한다.

[0047] 방법의 단계 218에서, 후보 발생기(208)는 입력 문자열(204)의 입력 단어들을 후보 표(216)의 주제 단어들에 매치한다. 그 후, 단계 220에서 후보 대체 단어(222) 및 대응하는 후보 점수(224)들이 대응하는 매치된 주제 단어에 기초하여 입력 문자열(204)의 각각의 입력 단어에 대해 후보 표(216)로부터 추출된다. 동일한 주제 또는 입력 단어에 결합된 많은 후보 대체 단어들이 있을 수 있으므로, 후보 발생기(208)는 각각의 매치된 입력 단어에 대해 하나 이상의 후보 대체 단어 및 점수 쌍을 생성할 수 있다. 후보 대체 단어(222) 및 점수(224) 쌍들은 부가 처리를 위해 컨텍스추얼 철자 엔진(210)에 출력된다.

[0048] 후보 표(216)의 후보 점수(224)는 일반적으로 입력 단어가 대응하는 후보 대체 단어로 대체되어야 하는 확률을 나타낸다. 또한, 후보 점수(224)는 주제 단어와 대응하는 후보 대체 단어(222) 사이의 오류의 양을 반영할 수 있다.

[0049] 본 발명의 한 실시예에 따라, 후보 점수(224)는 하나 이상의 인자들에 기초하며, 양호하게는 그 각각이 후보 표(216)의 주제 단어와 후보 대체 단어 쌍을 위한 후보 점수를 형성하도록 서로 곱해진다. 그러한 인자들은 주제 단어와 후보 대체 단어 사이의 편집 또는 타이핑 거리, 발견적 학습법(heuristics), 주제 단어와 후보 대체 단어 사이의 음성학적 차이 및 후보 대체 단어가 주제 단어를 대체해야 할 확률과 관련될 수 있는 다른 인자들을 포함한다.

[0050] 편집 종류 및 발견적 학습법에 기초한 후보 점수(224)들은 주제 단어와 후보 대체 단어 사이의 차이에 의존하여 변할 수 있다. 예를 들어, 독특한 주제 단어가 다수의 후보 대체 단어로 변환되어야 할 확률은 매우 드문 것일 수 있다. 결과적으로, 그러한 후보 대체 단어들은 낮은 후보 점수가 주어져야 한다. 다른 한편으로는, 주제 단어로부터 후보 대체 단어로의 변환이 매우 보편적으로 마주치는 철자 오류에 관한 것인 첫번째 문자의 변경을 필요로 할 때는, 그러한 주제 단어와 후보 대체 단어 쌍들은 높은 후보 점수를 받는다.

[0051] 주제 단어에 공백을 부가함으로써 주제 단어를 두 개의 단어들로 분리하는 것에 기초하는 편집 거리와 점수들도 관계된다. 그러한 편집들은 그들의 상대적으로 높은 빈도로 인해 일반적으로 높은 후보 점수가 주어진다.

[0052] 후보 대체 단어가 주제 단어의 동음이의어 또는 거의 동음이의어일 때는, 후보 대체 단어가 음성학적으로 주제

단어를 거의 닮지 않았을 때보다 더 높은 점수가 단어 쌍에게 주어진다.

[0053] 이 방법의 단계 226에서, 컨텍스추얼 철자 엔진(210)은 후보 대체 단어(222)의 대응하는 점수(224)에 기초하여 입력 문자열(204)의 입력 단어들에 대한 후보 대체 단어(222)들의 출력(230)을 선택적으로 생성한다. 본 발명의 한 실시예에 따라, 시스템(200)은 컨텍스추얼 철자 엔진(210)으로부터 후보 대체 문자열(242)들을 수신하는 언어 모델(240)을 포함한다. 후보 대체 문자열(242)들은 입력 문자열(204)들로 변경되며, 그 각각은 대응하는 입력 단어의 대신에 후보 대체 단어(222)를 포함한다.

[0054] 언어 모델(240)은 각각의 후보 대체 문자열(242)에 대해 확률 점수(244)를 출력하도록 작동한다. 확률 점수(244)는 일반적으로 문장들의 큰 로그에 대한 통계적 데이터(246)(즉, 다른 단어들에 대한 상대적 단어 발생 빈도 데이터)에 기초하여 특수한 후보 대체 문자열(242)을 보여줄 확률을 측정할 수 있게 한다. 일반적으로, 단어들의 특수한 문자열을 더 많이 보여줄수록 문자열에 포함된 단어들이 좀더 올바른 방식으로 이용될 것으로 추측된다. 그래서, 각각의 후보 대체 문자열(242)에 대한 확률 점수(244)는 그 문자열의 후보 대체 단어와 입력 단어들의 결합의 정확성을 반영한다.

[0055] 단어(즉, w_1, w_2, w_3,...w_N)들의 주어진 문자열에 대한 확률 점수(244)(P(문맥)으로 나타냄))는 통계적 데이터(246)를 이용하여 수학식 1에 따라 계산될 수 있다. 일반적으로, 문자열의 확률은 다른 단어들이 주어진 문자열에서의 각각의 단어의 확률과 등가이다. 그래서, 문자열의 확률은 제1 단어의 확률($P(w_1)$)을 제1 단어가 주어진 제2 단어의 확률($P(w_2|w_1)$)과 곱하고, 제1 및 제2 단어들이 주어진 제3 단어의 확률 ($P(w_3|w_2,w_1)$)과 곱하는 등등의 것과 등가이다.

수학식 1

[0056] $$P(\text{문맥})= P(w_1)*P(w_2|w_1)*P(w_3|w_2,w_1)* \ ...*P(w_N|w_{N-1},w_{N-2}...\ \ w_2,w_1)$$

[0057] 본 발명의 한 실시예에 따라, 수학식 2에서 제공되는, 수학식 1의 트라이그램 근사법(trigram approximation)이 이용된다. 문자열의 각각의 단어에 대해, 트라이그램 근사법은 문자열의 N 개의 모든 단어들보다는 두 개의 선행하는 단어(만일 있다면)들을 이용한다.

수학식 2

[0058] $$P(\text{문맥})\sim P(w_1)*P(w_2|w_1)*P(w_3|w_2,w_1)*...*P(w_N|w_{N-1},w_{N-2})$$

[0059] 본 발명의 한 실시예에 따라, 컨텍스추얼 철자 엔진(210)은 각각의 후보 대체 문자열의 최종 점수에 기초하여 제안된 후보 대체 단어 또는 문자열(230)들을 애플리케이션(206)에 출력하도록 선택한다. 최종 점수(P(후보입력 단어, 문맥)으로 나타냄))는 후보 대체 문자열(P(문맥, 후보로 나타냄))에 대한 확률 점수를 후보 대체 문자열(242)을 형성하기 위해 입력 단어를 대체한 후보 대체 단어(222)(후보)에 대응하는 후보 점수(224)로 곱함으로써 수학식 3에 따라 각각의 계산된다.

[0060] 예를 들어, "too", "tot" 및 "two"의 후보 대체 단어(222)들이 "I see you to"의 입력 문자열(204)의 입력 단어 "to"에 대한 후보 발생기(208)에 의한 그들의 대응하는 점수(224)들을 따라 발생되는 것을 말하기로 하자. 그 후, 대응하는 후보 대체 문자열(242)들은 "I see you too", "I see you tot" 및 "I see you two"로 된다. 그 후, 컨텍스추얼 철자 엔진(210)은 후보 대체 문자열(242)의 각각에 대한 확률 점수들을 계산하는 언어 모델(240)을 여러 번 호출한다.

[0061] 그 후, 후보 대체 문자열(242)에 대한 최종 점수들이 그들의 확률 점수(244)를 그들의 대응하는 후보 점수(224)로 곱함으로써 컨텍스추얼 철자 엔진(210)에 의해 계산된다. 그래서 후보 대체 문자열(242) "I see you too"에 대한 최종 점수가 후보 표(216)로부터 얻어진 입력 단어 "to"에 대한 후보 대체 단어 "too"에 대응하는 점수(224)로 곱해진 문자열 "I see you too"의 확률과 등가이다.

[0062] 본 발명의 한 실시예에 따라, 최고 최종 점수를 갖는 후보 대체 문자열의 후보 단어(222)가 컨텍스추얼 철자 엔진(210)에 의해 애플리케이션(206)에 대해 출력(230)으로 제안된다. 대안적으로, 컨텍스추얼 철자 엔진(210)은, 그것이 어떤 임계치를 초과한다는 전제하에, 최고 최종 점수를 갖는 후보 대체 문자열(242)에 대응하는 후보 대체 단어(222)만을 제안할 수 있다. 본 발명의 다른 한 실시예에 따라, 임계치를 초과하는 최종 점수들을 갖는 다수의 후보 대체 단어(222)들이 컨텍스추얼 철자 엔진(210)에 의해 애플리케이션(206)에 대해

출력(230)으로 제안된다.

[0063] 임계치는 예정되어 있거나 또는 주제 단어들과 후보 단어들의 확률의 함수로서 동적으로 계산될 수 있다. 한 실시예에서는, 임계치가 임계치 = α P(주제 단어들) + β P(후보 단어들) + ɣ P(입력 단어들) - P(후보 단어들)로부터 동적으로 판단된다.

[0064] **후보 표 발생**

[0065] 도 4는 본 발명의 실시예들에 따른 철자 검사 시스템(200)이 이용하기 위한 후보 표(216)를 발생시키는 방법을 예시하는 플로우차트이다. 이 방법의 단계 250에서는, 단어들의 어휘 목록이 제공된다. 양호하게는 어휘 목록이 매우 크다 (예를 들어, 100,000 단어 이상). 다음에, 단계 252에서, 어휘 목록의 주제 단어들이 어휘 목록의 다른 단어들과 비교된다. 양호하게는, 어휘 목록의 각각의 단어, 또는 적어도 어휘 목록의 가장 자주 이용된 단어들의 각각은 어휘 목록의 다른 단어들과 비교되는 주제 단어들이 된다. 단계 254에서, 후보 대체 단어들은 단계 252에서의 비교에 기초하여 주제 단어들을 위해 식별된다.

[0066] 본 발명의 한 실시예에 따라, 어휘 목록의 다른 단어들에 대한 주제 단어들의 비교(단계 252)는 주제 단어에 대한 어휘 목록의 각각의 단어들 사이의 편집 또는 타이핑 거리를 계산하고 편집 거리들을 임계치 편집 거리와 비교하는 것을 필요로 한다. 임계치 편집 거리를 충족하는 편집 거리를 갖는 후보 대체 단어들이 주제 단어들에 대한 후보 대체 단어로서 식별된다. 임계치를 "충족한다"는 것은, 편집 거리들이 어떻게 계산되느냐에 따라, 임계치에 도달하거나, 임계치를 초과하거나 또는 임계치 미만으로 되는 것에 의해 충족되는 것으로 하려는 것임을 이해해야 한다.

[0067] 비교 단계 252의 다른 한 실시예에 따라, 어휘 목록의 각각의 단어들의 의미가 주제 단어들과 비교된다. 식별 단계 254는 주제 단어와 의미가 유사한 어휘 목록의 단어들을 후보 대체 단어들로서 식별하는 것을 포함한다. 예를 들어, 주제 단어들의 동의어들은 후보 대체 단어로서 식별될 수 있다. 본 발명의 한 실시예에 따라, 어휘 목록의 주제 단어들은 기준어 사전(thesaurus) 데이터에 대해 검사되며, 그로부터 유사한 의미를 갖는 후보 대체 단어들이 이 방법의 단계 254에서 후보 대체 단어들로 식별된다.

[0068] 비교 단계 252의 다른 한 실시예에 따라, 어휘 목록의 단어들의 음성학적 표시들이 어휘 목록의 주제 단어들과 비교된다. 어휘 목록의 단어들의 음성학적 표시들은 양호하게는 종래의 텍스트 음성 변환(text-to-speech) 엔진에 단어를 제출하는 것을 통해 자동으로 발생된다. 주제 단어의 음성학적 표시와 매치하는 음성학적 표시들을 갖는 어휘 목록의 단어들은 식별 단계 254에서 주제 단어에 대한 후보 대체 단어들로서 식별된다. 이러한 쌍들의 예시들은 "bear"와 "bare" 및 "which"와 "witch"를 포함한다. 그래서, 주제 단어의 동음이의어들이 후보 대체 단어들로서 식별된다. 본 발명의 다른 한 실시예에 따라, 거의 동음이의어(즉, 임계치를 충족하는 것들)인 어휘 목록의 단어들도 후보 대체 단어들로서 식별된다.

[0069] 본 발명의 다른 한 실시예는 문장들의 큰 로그에서 발견되는 바이그램(bigrams)(즉, 단어 쌍)들의 분석을 포함한다. 그 분석은 바이그램의 제1 및 제2 단어들 사이에 배치된 공백을 이동 또는 삭제하는 것이 적어도 하나의 유효한 단어의 발생으로 귀결될 수 있는지를 판단하는 것이 필요하다. 비교 단계 252의 한 실시예는 주제 단어들을 공백 이동 분석으로부터 발생된 유효한 단어들과 비교하는 것을 포함한다. 양호하게는, 제1 단어의 끝 문자 앞 또는 제2 단어의 첫문자 뒤의 공백을 이동함으로써, 또는 그 공백을 삭제함으로써 형성되는 유효한 단어들만, 그들이 매우 보편적인 활자 오류들에 대응하는 것이기 때문에, 비교 단계 252에서 이용된다. 그리고, 주제 단어들과 매치하는 새롭게 형성된 유효한 단어들은 단계 254에서 주제 단어들에 대한 후보 대체 단어들로서 식별된다. 예를 들어, "use swords"이라는 단어 쌍에 대한 후보 대체 단어들은 "uses words"일 수 있고, "dog sand"라는 단어 쌍에 대한 후보 대체 단어들은 "dogs and"일 수 있다. 마찬가지로, "any one"이라는 단어 쌍에 대한 후보 대체 단어는 "anyone"일 수 있고, 역으로, "anyone"이라는 단어 쌍에 대한 후보 대체 단어는 "any one"일 수 있다.

[0070] 이 방법의 단계 256에서, 대응하는 식별된 후보 대체 단어들과 쌍을 이룬 주제 단어들을 포함하는 후보 목록(216)이 형성된다. 마지막으로, 단계 258에서, 후보 표(216)가 도 1과 관련하여 앞서 기술한 바와 같은 컴퓨터 판독가능 매체에 저장된다.

[0071] 본 발명의 다른 한 실시예에 따라, 앞서 설명했듯이 후보 대체 단어가 대응하는 주제 단어로 대체되어야 하는 확률에 기초하여 단계 256에서 주제 단어와 후보 대체 단어들의 각각의 쌍들에 대해 후보 점수(224)가 발생된다. 후보 점수들은 양호하게는 비교 단계 252에서 분석된 하나 이상의 인자들 및 앞서 설명한 것들에 기초한다. 후보 점수는 단계 258에서 컴퓨터 판독가능 매체에 저장되는 후보 표(216)에 포함되어 있다.

[0072] **후보 제외 표**

[0073] 본 발명의 다른 한 실시예에 따라, 후보 발생기(208)가 후보 대체 단어(222)로서 컨텍스추얼 철자 엔진(210)에 제출되지 않아야 하는 특정한 후보 대체 단어들을 식별하는 후보 제외 표(260)가 발생된다. 그래서, 후보 제외 표(260)는 부적절하거나 바람직하지 않은 후보 대체 단어들이 컨텍스추얼 철자 엔진(210)에 의해 애플리케이션(206)에 제안되는 것을 방지한다. 양호하게는, 후보 제외 표(260)가 불쾌감을 주는 후보 대체 단어들을 포함한다. 또한, "rough"와 "tough" 등과 같이 명확하게 하기 어렵거나 유사한 문맥들에서 흔히 발생하는 단어들이 후보 제외 표에 포함될 수도 있다. "color"와 "colour", 또는 "goodbye"와 "good-bye" 등과 같이 동일한 주제 단어의 용인할 만한 철자 변형들을 후보 제외 표(260)에 포함시켜 제외할 수 있다. 또한, 독특한 입력 단어가 다수의 형태로 귀결되는 활자 오류들은 드물기 때문에, 양호하게는, 독특한 주제 단어들 및 그들에 대응하는 다수의 단어들이 후보 제외 표(260)에 포함된다.

[0074] 본 발명의 한 실시예에 따라, 후보 표(216)는 후보 제외 표(260)에서 매칭 단어 쌍들을 갖는 주제 단어와 후보 대체 단어 쌍들을 제외시키기 위해 주기적으로 업데이트된다. 또한, 후보 제외 표(260)의 단어들과 매치하는 후보 대체 단어들을 갖는 후보 표(216)의 주제 단어와 후보 대체 단어 쌍들이 제외될 수도 있다. 그렇게 산출되는 후보 표(216)의 크기의 감축은 철자 검사 시스템(200)이 좀더 효율적으로 작동하게 한다.

[0075] 본 발명이 특수한 실시예들을 참조하여 기술되었지만, 이 분야에 숙련된 작업자들은 본 발명의 정신 및 범위로부터 벗어남이 없는 형태 및 상세사항들에서의 변화가 이루어질 수 있음을 알 것이다.

도면의 간단한 설명

[0013] 도 1은 본 발명이 실시될 수 있는 컴퓨팅 환경의 블록 다이어그램이다.

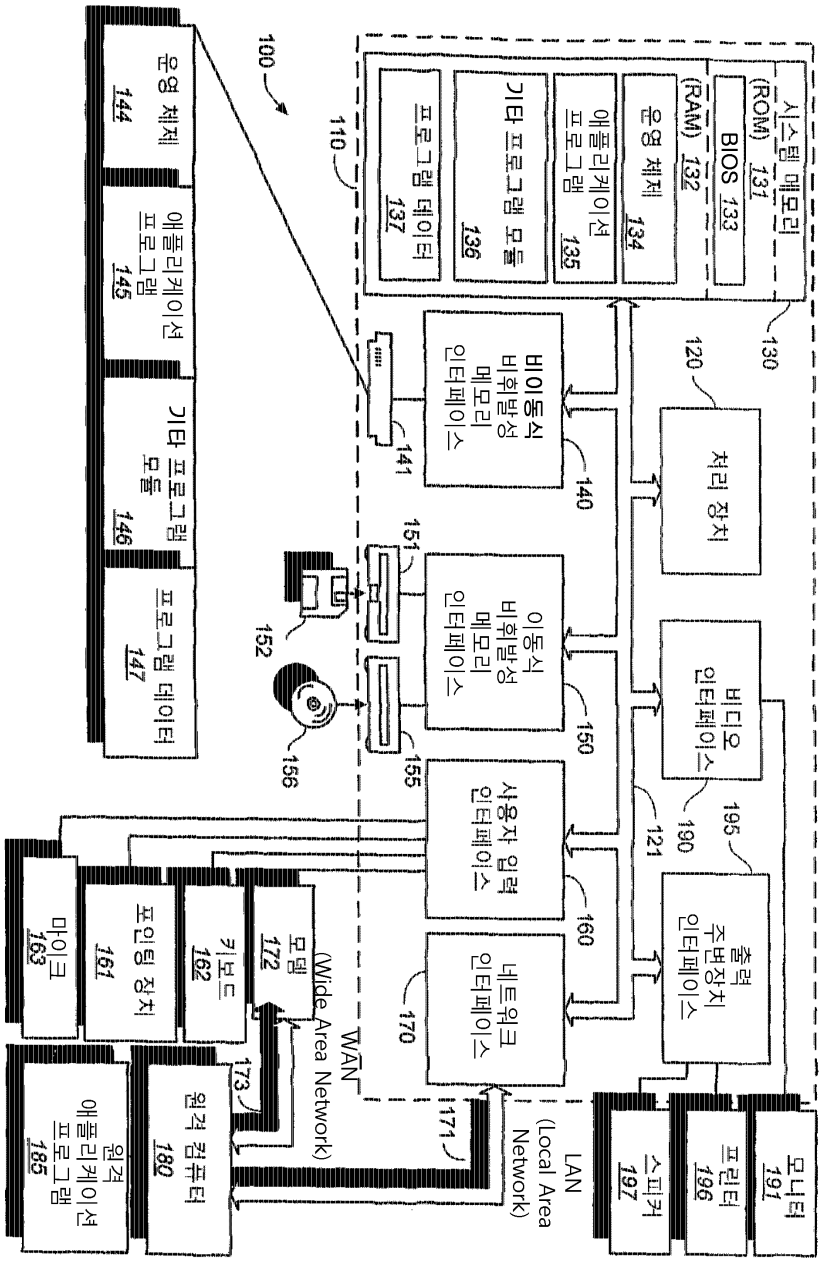
[0014] 도 2는 본 발명의 실시예들에 따라 문자열의 단어들에 대한 대체 단어들을 제안하는 방법을 예시하는 플로우차트이다.

[0015] 도 3은 본 발명의 실시예들에 따른 철자 검사 시스템의 블록 다이어그램이다.

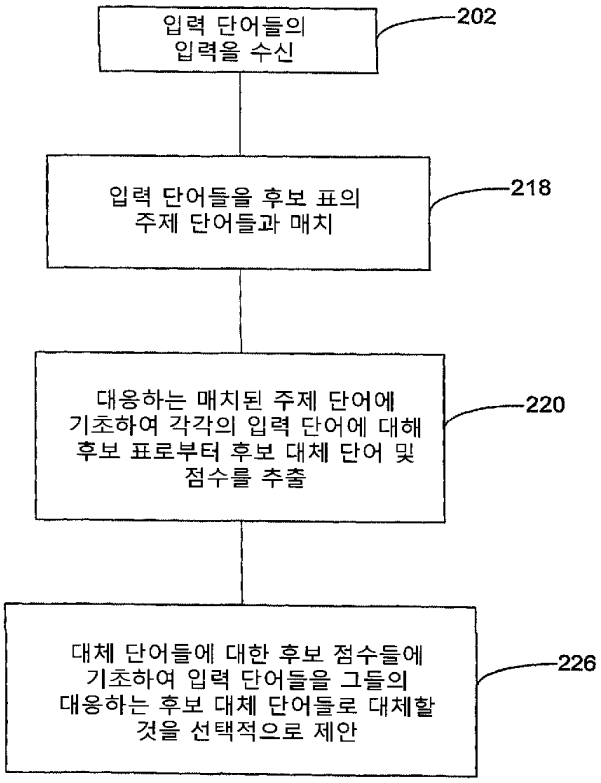
[0016] 도 4는 철자 검사 시스템이 본 발명의 실시예들에 따라 입력 문자열의 입력 단어들에 대한 대체 단어들을 제안하기 위해 이용하는 후보 표를 발생시키는 방법을 예시하는 플로우차트이다.

도면

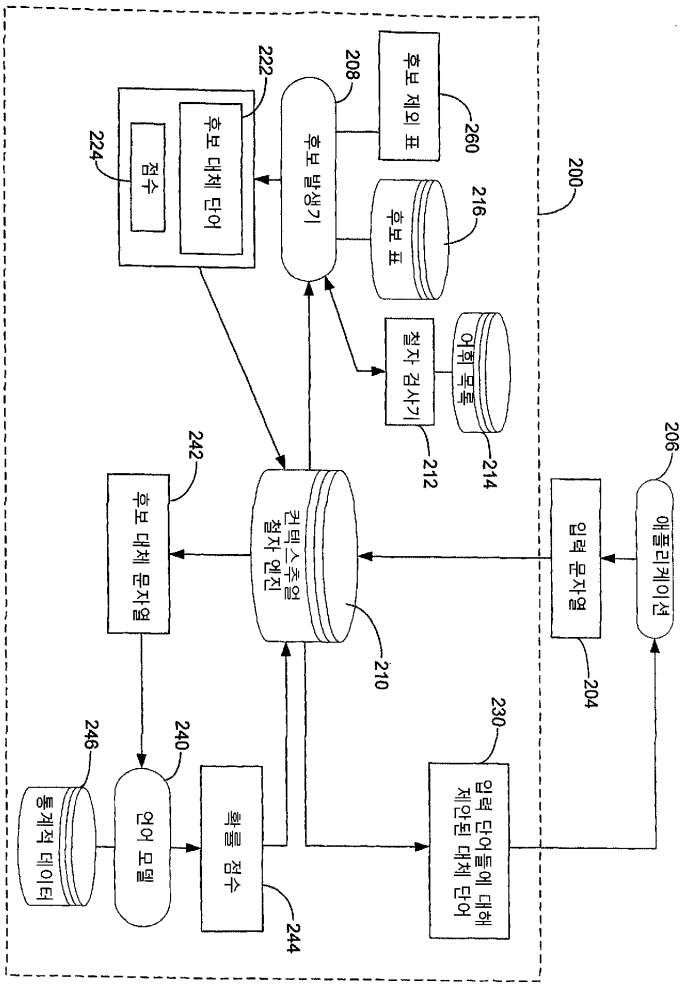
도면1


100
110
시스템 메모리 (ROM) 131
BIOS 133
(RAM) 132
운영 체제 134
애플리케이션 프로그램 135
기타 프로그램 모듈 136
프로그램 데이터 137
130
120
처리 장치
비이동식 비휘발성 메모리 인터페이스
140
141
이동식 비휘발성 메모리 인터페이스
150
151
155
152
156
비디오 인터페이스
190
195
121
출력 주변장치 인터페이스
사용자 입력 인터페이스
160
네트워크 인터페이스
170
171
LAN (Local Area Network)
WAN (Wide Area Network)
173
모뎀 172
키보드 162
포인팅 장치 161
마이크 163
원격 컴퓨터 180
원격 애플리케이션 프로그램 185
운영 체제 144
애플리케이션 프로그램 145
기타 프로그램 모듈 146
프로그램 데이터 147
모니터 191
프린터 196
스피커 197

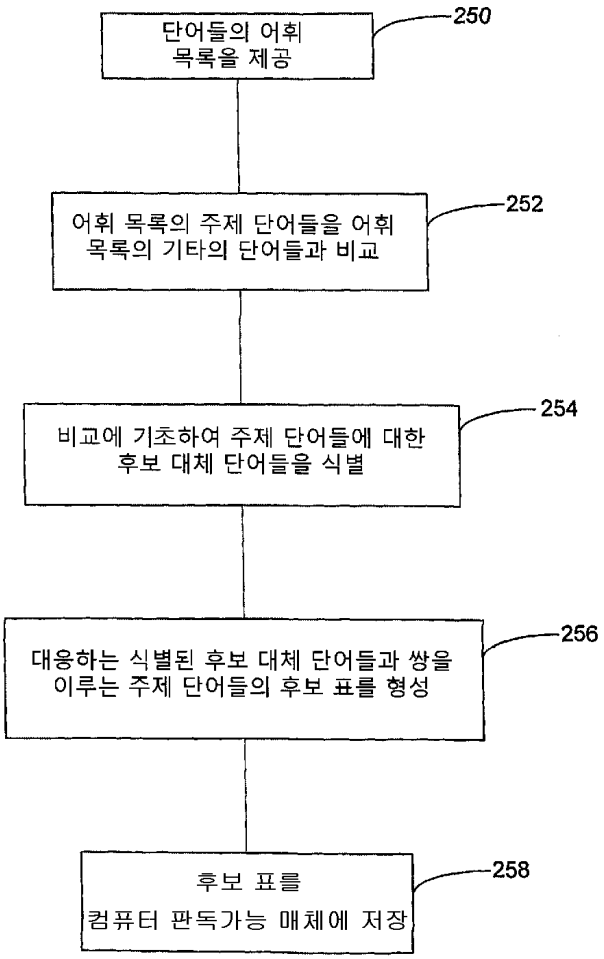
도면2


입력 단어들의 입력을 수신
202
입력 단어들을 후보 표의 주제 단어들과 매치
218
대응하는 매치된 주제 단어에 기초하여 각각의 입력 단어에 대해 후보 표로부터 후보 대체 단어 및 점수를 추출
220
대체 단어들에 대한 후보 점수들에 기초하여 입력 단어들을 그들의 대응하는 후보 대체 단어들로 대체할 것을 선택적으로 제안
226

도면3


200
후보 제외 표
260
208
후보 발생기
222
후보 대체 단어
224
점수
216
후보 표
철자 검사기
212
어휘 목록
214
206
애플리케이션
입력 문자열
204
242
후보 대체 문자열
컨텍스추얼 철자 엔진
210
230
입력 단어들에 대해 제안된 대체 단어
246
통계적 데이터
240
언어 모델
확률 점수
244

도면4


단어들의 어휘 목록을 제공
250
어휘 목록의 주제 단어들을 어휘 목록의 기타의 단어들과 비교
252
비교에 기초하여 주제 단어들에 대한 후보 대체 단어들을 식별
254
대응하는 식별된 후보 대체 단어들과 쌍을 이루는 주제 단어들의 후보 표를 형성
256
후보 표를 컴퓨터 판독가능 매체에 저장
258