(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0015933
(43) 공개일자 2016년02월15일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) *G06F 17/30* (2006.01)
(21) 출원번호 10-2014-0098936
(22) 출원일자 2014년08월01일
심사청구일자 2014년08월01일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임해창
서울특별시 동대문구 제기로4길 45, 401호(제기동, 더리빙플러스)
고대옥
서울특별시 동작구 상도로30길 78,마르마 404호 (상도동)
(74) 대리인
송인호, 민영준, 최관락

전체 청구항 수 : 총 10 항

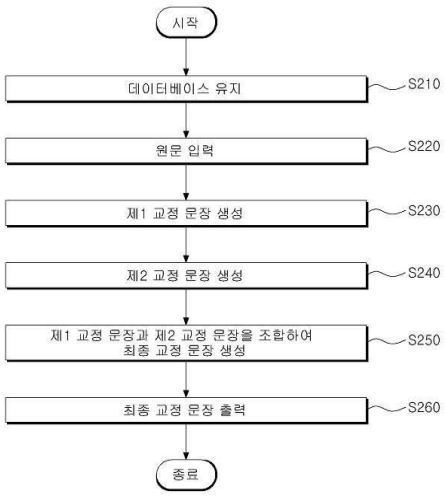
(54) 발명의 명칭 **소셜 텍스트를 위한 철자 오류 교정 방법 및 장치**

(57) 요 약

본 발명은 규칙 기반 교정 방식과 통계 기반 교정 방식을 혼합하여 철자 오류를 교정하는 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치에 관한 것이다

본 발명의 일실시예에 의한 소셜 텍스트를 위한 철자 오류 교정 방법은 규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장된 데이터베이스를 유지하는 단계; 원문의 입력을 수신하는 단계; 상기 원문에 대해 상기 규칙 기반 교정 방식에 대한 정보를 이용하여 제1교정 문장을 생성하는 단계; 상기 원문에 대해 상기 통계 기반 교정 방식에 대한 정보를 이용하여 제2교정 문장을 생성하는 단계; 및 상기 제1교정 문장과 제2교정 문장을 조합하여 최종 교정 문장을 생성하는 단계를 포함할 수 있다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업
과제고유번호 2012M3C4A7033344
부처명 미래창조과학부
연구관리전문기관 한국연구재단
연구사업명 차세대 정보컴퓨팅 기술개발사업
연구과제명 비정형 빅 데이터 분석 및 개인화 기술 개발
기 여 율 1/1
주관기관 포항공과대학교
연구기간 2012.07.01 ~ 2017.06.30

특허청구의 범위

청구항 1

규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장된 데이터베이스를 유지하는 단계;

원문의 입력을 수신하는 단계;

상기 원문에 대해 상기 규칙 기반 교정 방식에 대한 정보를 이용하여 제1교정 문장을 생성하는 단계;

상기 원문에 대해 상기 통계 기반 교정 방식에 대한 정보를 이용하여 제2교정 문장을 생성하는 단계; 및

상기 제1교정 문장과 제2교정 문장을 조합하여 최종 교정 문장을 생성하는 단계를 포함하는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 방법.

청구항 2

제 1 항에 있어서, 상기 최종 교정 문장 생성 단계는

상기 제1교정 문장에서 교정이 이루어지지 않은 철자는 상기 제2교정 문장의 대응되는 철자로 대체하는 단계를 포함하는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 방법.

청구항 3

제 2 항에 있어서, 상기 규칙 기반 교정 방식에 대한 정보는

오류 패턴 및 상기 오류 패턴과 쌍으로 매핑된 교정 패턴을 포함하는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 방법.

청구항 4

제 3 항에 있어서, 상기 제2교정 문장 생성 단계는

복수의 후보 문장 중 점수가 가장 높은 문장을 선택하는 단계를 포함하되,

상기 점수는 상기 통계 기반 교정 방식에 대한 정보에 포함된 오류 패턴에 대응되는 후보 교정 패턴의 변환 확률, 및 상기 후보 교정 패턴의 출현 확률을 고려하여 산정되는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 방법.

청구항 5

제 2 항에 있어서,

상기 규칙 기반 교정 방식은 상기 통계 기반 교정 방식에 비해 교정 정확률을 높고, 교정 재현율은 낮은 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 방법.

청구항 6

규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장된 데이터베이스;

원문의 입력을 수신하는 입력부;

상기 원문에 대해 상기 규칙 기반 교정 방식에 대한 정보를 이용하여 제1교정 문장을 생성하는 제1교정부;

상기 원문에 대해 상기 통계 기반 교정 방식에 대한 정보를 이용하여 제2교정 문장을 생성하는 제2교정부;

상기 제1교정 문장과 제2교정 문장을 조합하여 최종 교정 문장을 생성하는 결과 혼합부; 및

상기 데이터베이스, 상기 입력부, 상기 제1교정부, 상기 제2교정부 및 상기 결과 혼합부를 제어하는 제어부를 포함하는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 장치.

청구항 7

제 6 항에 있어서, 상기 결과 혼합부는

상기 제1교정 문장에서 교정이 이루어지지 않은 철자는 상기 제2교정 문장의 대응되는 철자로 대체하는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 장치.

청구항 8

제 7 항에 있어서, 상기 규칙 기반 교정 방식에 대한 정보는

오류 패턴 및 상기 오류 패턴과 쌍으로 매핑된 교정 패턴을 포함하는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 장치.

청구항 9

제 8 항에 있어서,

상기 제2교정부는 복수의 후보 문장 중 점수가 가장 높은 문장을 선택함으로써, 상기 제2교정 문장을 생성하고,

상기 점수는 상기 통계 기반 교정 방식에 대한 정보에 포함된 오류 패턴에 대응되는 후보 교정 패턴의 변환 확률, 및 상기 후보 교정 패턴의 출현 확률을 고려하여 산정되는 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 장치.

청구항 10

제 7 항에 있어서,

상기 규칙 기반 교정 방식은 상기 통계 기반 교정 방식에 비해 교정 정확률을 높고, 교정 재현율은 낮은 것을 특징으로 하는 소셜 텍스트를 위한 철자 오류 교정 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치에 관한 것으로, 보다 상세하게는 규칙 기반 교정 방식과 통계 기반 교정 방식을 혼합하여 철자 오류를 교정하는 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 스마트폰, 태블릿 등의 모바일 기기의 사용이 증가함에 따라, 실시간으로 자신의 상황을 알리거나 생각을 표현할 수 있는 소셜 네트워크 서비스가 각광을 받고 있다. 트위터는 대표적인 소셜 네트워크 서비스의 하나로서, 트위터 사용자들이 작성하는 트윗 메시지는 폭발적으로 증가하고 있는 추세이다.

[0003] 이러한 대량의 트윗은 최신 트렌드, 실시간 이슈 탐지, 또는 의사 결정을 위한 빅데이터 마이닝에 매우 활용가치가 높은 자원이다.

[0004] 그러나, 트윗에서는 많은 철자 오류가 발견된다. 예를 들어, 발음이 비슷한 단어로 바꿔 쓰는 경우(요 → 여, 고 → 구)나 오탈자(있 → 잇), 음절의 의도적 변형(다→ 당)이 있다. 이러한 철자 오류는 텍스트 마이닝 과정에서 필수적인 통계 분석 과정에 오류를 전파함으로써, 텍스트 분석의 신뢰도를 떨어뜨린다. 더 정확하고 신뢰성 있는 트윗 데이터 분석을 위해서는 이런 트윗 철자 오류를 자동으로 교정해 줄 수 있는 시스템의 개발이 필수적으로 요구된다.

[0005] 철자 오류 교정 태스크는 사용자가 작성한 문장을 입력으로 받아 그 문장 내의 철자 오류를 자동으로 탐지하고, 그것을 올바른 철자로 수정한 결과를 출력하는 태스크이다. 예를 들면 아래와 같다.

[0006] 원본 문장: 처음부터 여유롭개 할 걸...

[0007] 교정 문장: 처음부터 여유롭게 할 걸...

[0008] 기존 철자 오류 교정 연구는 주로 SMS 메시지를 대상으로 이루어져 왔다. 관련 선행문헌으로 대한민국 등록특허

제10-0897718호가 있다.

[0009] SMS 메시지를 대상으로 철자 오류 교정 코퍼스를 구축한 후, 그 코퍼스를 바탕으로 규칙을 추출하거나 통계 모델을 학습시키는 방식을 주로 사용하였다. SMS 메시지가 아닌 트윗에서의 철자 오류 교정을 위해서는 SMS 교정 코퍼스와 마찬가지로 트윗에서 자주 나타나는 오류를 학습할 수 있는 트윗 데이터에 대한 교정 코퍼스가 필요하지만, 현재 사용 가능한 자원을 구하기 어려운 실정이다. 이로 인해, 트윗에 적합한 철자 오류 교정 연구 역시 미흡한 실정이다.

[0010] 따라서 트윗 등의 소셜 미디어에서 사용자가 생성한 텍스트의 철자 오류 교정 기술에 대한 연구가 필요한 실정이다.

발명의 내용

해결하려는 과제

[0011] 본 발명의 목적은 규칙 기반 교정 방식과 통계 기반 교정 방식을 혼합 사용함으로써, 많은 철자를 정확하게 교정할 수 있는 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치를 제공하는 데 있다.

과제의 해결 수단

[0012] 상기 목적을 달성하기 위해 본 발명의 일실시예에 의하면, 규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장된 데이터베이스를 유지하는 단계; 원문의 입력을 수신하는 단계; 상기 원문에 대해 상기 규칙 기반 교정 방식에 대한 정보를 이용하여 제1교정 문장을 생성하는 단계; 상기 원문에 대해 상기 통계 기반 교정 방식에 대한 정보를 이용하여 제2교정 문장을 생성하는 단계; 및 상기 제1교정 문장과 제2교정 문장을 조합하여 최종 교정 문장을 생성하는 단계를 포함하는 소셜 텍스트를 위한 철자 오류 교정 방법이 개시된다.

[0013] 상기 목적을 달성하기 위해 본 발명의 일실시예에 의하면, 규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장된 데이터베이스; 원문의 입력을 수신하는 입력부; 상기 원문에 대해 상기 규칙 기반 교정 방식에 대한 정보를 이용하여 제1교정 문장을 생성하는 제1교정부; 상기 원문에 대해 상기 통계 기반 교정 방식에 대한 정보를 이용하여 제2교정 문장을 생성하는 제2교정부; 상기 제1교정 문장과 제2교정 문장을 조합하여 최종 교정 문장을 생성하는 결과 혼합부; 및 상기 데이터베이스, 상기 입력부, 상기 제1교정부, 상기 제2교정부 및 상기 결과 혼합부를 제어하는 제어부를 포함하는 소셜 텍스트를 위한 철자 오류 교정 장치가 개시된다.

발명의 효과

[0014] 본 발명의 일실시예에 의한 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치는 많은 철자를 정확하게 교정할 수 있다.

[0015] 본 발명의 일실시예에 의하면, 기존 방식에 비해 오류 교정이 더 잘 된 문장을 생성할 수 있기에, 다른 자연어 처리 기술의 전처리 수단으로 유용하게 사용될 수 있다.

도면의 간단한 설명

[0016] 도 1은 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 장치의 블록도이다.

도 2는 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법을 나타내는 흐름도이다.

도 3은 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법 중 규칙 기반 교정 방식을 먼저 적용한 후에 통계 기반 통계 방식을 적용하는 철자 오류 교정 방법의 개념도이다.

도 4 내지 도 5는 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법에 따른 성능 변화를 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0017] 이하, 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치에 대해 도면을 참조하여 설명하도록 하겠다.

[0018] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들

을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다.

[0019] 본 명세서에서 소셜 텍스트라 함은 소셜 미디어에서 사용자가 생성한 텍스트를 의미한다.

[0020] 도 1은 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 장치의 블록도이다.

[0021] 도시된 바와 같이, 철자 오류 교정 장치(100)는 데이터베이스(110), 입력부(120), 제1교정부(130), 제2교정부(140), 결과 혼합부(150), 출력부(160), 및 제어부(170)를 포함할 수 있다.

[0022] 상기 데이터베이스(110)에는 규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장될 수 있다.

[0023] 상기 규칙 기반 교정 방식은 교정 규칙을 생성한 후, 교정 규칙에 맞는 유형의 오류에 해당 규칙을 적용하여 철자 오류 교정을 하는 교정 방법이다. 교정 규칙의 생성에는 자동 생성과 수동 생성으로 나뉘며, 자동 생성의 경우 원본 문장과 정답 문장 쌍으로 이루어진 교정 말뭉치로부터 교정 내역을 추출하는 방식이며, 수동 생성은 교정 전문가가 직접 교정 규칙을 구성하는 방식이다. 상기 규칙 기반 교정 방식에 대한 정보는 오류 패턴 및 상기 오류 패턴과 쌍으로 매핑된 교정 패턴을 포함할 수 있다.

[0024] 통계 기반 교정 방식은 원본 문장에 대해 변환될 수 있는 모든 후보 문장들을 생성한 후, 가장 점수가 높은 후보 문장을 선택하는 방식이다. 상기 점수는 '후보 문장이 실제로 작성될 확률'과 '원본 문장으로부터 후보 문장으로 변환될 확률'을 고려하여 책정될 수 있다. 예를 들어, 상기 후보 문장이 실제로 작성될 확률과 원본 문장으로부터 후보 문장으로 변환될 확률이 곱을 이용하여 상기 점수가 산정될 수 있다.

[0025] 입력부(120)는 사용자가 작성한 소셜 텍스트(원문)를 입력 받을 수 있다.

[0026] 제1교정부(130)는 상기 규칙 기반 교정 방식에 대한 정보를 이용하여 제1교정 문장을 생성할 수 있다.

[0027] 제2교정부(130)는 통계 기반 교정 방식에 대한 정보를 이용하여 제2교정 문장을 생성할 수 있다.

[0028] 결과 혼합부(150)는 상기 제1교정부(130)에서 생성된 제1교정 문장과 상기 제2교정부(140)에서 생성된 제2교정 문장을 조합하여 최종 교정 문장을 생성할 수 있다.

[0029] 출력부(160)는 상기 최종 교정 문장을 출력할 수 있다.

[0030] 제어부(170)는 데이터베이스(110), 입력부(120), 제1교정부(130), 제2교정부(140), 결과 혼합부(150) 및 출력부(160)를 전반적으로 제어할 수 있다.

[0031] 도 2는 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법을 나타내는 흐름도이다.

[0032] 제어부(170)는 규칙 기반 교정 방식에 대한 정보 및 통계 기반 교정 방식에 대한 정보가 저장된 데이터베이스(110)를 유지할 수 있다(S210).

[0033] 입력부(120)는 원문 입력을 수신할 수 있다(S220). 상기 원문은 소셜 텍스트를 포함할 수 있다.

[0034] 제1교정부(130)는 상기 원문에 대해 규칙 기반 교정 방식에 따라 철자 오류를 교정하여 제1교정 문장을 생성할 수 있다(S230). 제1교정부(130)는 규칙 기반 교정 방식에 대한 정보에 포함된 오류 패턴 및 상기 오류 패턴과 쌍으로 매핑된 교정 패턴을 이용하여 제1교정 문장을 생성할 수 있다.

[0035] 제2교정부(140)는 상기 원문에 대해 통계 기반 교정 방식에 따라 철자 오류를 교정하여 제2교정 문장을 생성할 수 있다(S240). 상기 제2교정부(140)는 통계 기반 교정 방식에 대한 정보에 포함된 오류 패턴에 대응되는 후보 교정 패턴, 상기 후보 교정 패턴의 변환 확률, 및 상기 후보 교정 패턴의 출현 확률을 이용하여 원문의 철자 오류를 교정할 수 있다. 예를 들어, 제2교정부(140)는 후보 교정 패턴의 변환 확률과 상기 후보 교정 패턴의 출현 확률의 곱을 이용하여 점수를 책정하고, 책정된 점수가 가장 높은 교정 패턴을 포함하는 문장을 제2교정 문장으로 선택할 수 있다.

[0036] 상기 제1교정부(130)와 상기 제2교정부(140)는 서로 독립적으로 병렬적으로 동작될 수도 있다.

[0037] 결과 혼합부(150)는 상기 제1교정문과 제2교정문을 조합하여 최종 교정문을 생성할 수 있다(S250).

[0038] 규칙 기반 교정 방식과 통계 기반 교정 방식을 비교하였을 때, 규칙 기반 교정 방식은 통계 기반 교정 방식보다 교정 정확률이 높고, 교정 재현률은 상대적으로 낮다.

[0039] 따라서 상기 결과 혼합부(150)는 규칙 기반 교정 방식에 의한 교정 결과와 통계 기반 교정 방식에 의한 교정 결과가 충돌할 경우, 규칙 기반 교정 방식에 의한 교정 내역이 우선적으로 적용되게 함으로써, 교정 정확률을 최대한 유지할 수 있다. 즉, 결과 혼합부(150)는 상기 제2교정문을 적용하여 교정할 경우에는 제1교정문에서 교정된 부분이 아닌 부분만이 교정되게 할 수 있다.

[0040] 이하에서 규칙 기반 교정 방식에 의한 교정 결과를 먼저 적용하고, 차후에 통계 기반 교정 방식에 의한 교정 결과를 적용하는 방식을 선 규칙 후 통계 방법이라 한다.

[0041] 출력부(160)는 최종 교정 문장을 사용자 인터페이스를 통해 출력할 수 있다(S260).

[0042] 도 3은 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법 중 선 규칙 후 통계 방법의 개념도이다.

[0043] 도시된 바와 같이, 결과 혼합부(140)는 먼저 입력 문장에 대해 규칙 기반 철자 오류 교정 모델 이용하는 규칙 기반 교정 방식과 통계 기반 철자 오류 교정 모델을 이용하는 통계 기반 교정 방식에 따라 각각 교정 문장을 생성한 후에, 두 교정 문장을 조합할 수 있다. 이 때, 결과 혼합부(140)는 입력 문장과 비교했을 때 규칙 기반 교정 방식에 의해 교정이 이루어진 부분은 통계 기반 교정 방식에 의한 교정 시 교정 여부와 무관하게 무조건 적용할 수 있다. 결과 혼합부(140)는 그런 다음 규칙 기반 교정 방식에 의해 교정이 이루어지지 않은 부분에 대해 통계 기반 교정 방식에 의해 교정이 이루어진 부분을 마저 적용할 수 있다.

[0044] 다음은 선 규칙 후 통계 방법에 의한 교정의 예시이다.

[0045] 오류 포함 원문: 결쿠 후회스럽지 않는 선택임니다.

[0046] 원문에는 총 3개의 철자 오류가 포함되어 있다.

[0047] 규칙 기반 교정: 결코 후회스럽지 않는 선택입니다.

[0048] 교정 규칙 - 쿠 → 코, 임 → 입

[0049] 규칙 기반 교정에서는 2개의 철자 오류를 교정하였다. 이 2개의 교정은 혼합 방식 교정에도 모두 적용된다.

[0050] 통계 기반 교정: 결국 후회스럽지 않은 선택입니다.

[0051] 교정 내역 - 쿠 → 국, 는 → 은, 임 → 입

[0052] 통계 기반 교정에서는 3개의 철자 오류를 교정하였다. 이 중에서 2개의 교정(쿠 → 국, 임 →입)은 이미 규칙 기반 교정에 의해 교정된 부분(쿠→ 코, 임 → 입)이기 때문에 제외되어, 나머지 1개의 교정(는 → 은)만이 혼합 방식 교정에 적용되었다.

[0053] 혼합 방식 교정(선 규칙 후 통계 방법): 결코 후회스럽지 않은 선택입니다.

[0054] 적용된 규칙 기반 교정: 쿠 → 코, 임 → 입

[0055] 적용된 통계 기반 교정: 는 → 은

[0056] 따라서 혼합 방식 교정(선 규칙 후 통계 방법)은 규칙 기반 교정 방식의 장점인 높은 교정 정확률과 통계 기반 교정 방식의 장점인 교정 재현율을 적절히 취할 수 있다.

[0057] 이하에서는 각각의 말뭉치를 10-fold 교차 검증법을 통해 철자 오류 교정 방식에 대한 평가(실험)에 대해 설명하도록 하겠다.

[0058] 본 실험에서는 아래 수학식에 기재된 평가 척도들을 사용하였다.

수학식 1

[0059]
$$\text{교정 정확률} = \frac{\text{올바르게 교정된 어절 수}}{\text{시스템이 교정한 어절 수}}$$

수학식 2

[0060]
$$\text{교정 재현율} = \frac{\text{올바르게 교정된 어절 수}}{\text{올바른 교정 어절 수}}$$

수학식 3

[0061]
$$\text{F-measure} = \frac{2 \times \text{정확률} \times \text{재현율}}{(\text{정확률} + \text{재현율})}$$

수학식 4

[0062]
$$\text{교정 오적용율} = \frac{\text{맞았던 어절을 틀리게 교정한 어절 수}}{\text{시스템이 교정한 어절 수}}$$

[0063] 교정 정확률과 교정 재현율은 시스템이 교정한 어절수와 교정 코퍼스의 정답 어절 수를 기준으로 삼았다. 그리고 F-measure(또는 F1)는 교정 정확률과 교정 재현율을 하나로 표현하기 위해 사용하였다. F1 척도는 교정 정확률과 교정 재현율을 조합한 평가 척도이다.

[0064] 소셜 텍스트로는 소셜 미디어인 SMS와 트위터로부터 생성한 말뭉치를 실험에 사용하였으며, 해당 말뭉치에 대한 통계는 표 1과 같다.

표 1

[0065]

말뭉치 종류	총 문장 수	총 어절 수	문장 당 어절 수	오류 어절 수
SMS	109,084	336,564	3.09	70,226
트윗	13,534	85,837	6.34	3,018

[0066] 도 4 내지 도 5는 본 발명의 일실시예와 관련된 소셜 텍스트를 위한 철자 오류 교정 방법에 따른 성능 변화를 나타내는 그래프이다. 도 4는 SMS 말뭉치에 대한 선 규칙 후 통계 방법의 성능 변화를 나타낸 그래프이고, 도 5는 트윗 말뭉치에 대한 선 규칙 후 통계 방법의 성능 변화를 나타낸 그래프이다.

[0067] 도 4에서는 SMS 말뭉치에 대해 파라미터를 변경한 실험 결과를 그래프로 표현하였으며, 규칙 기반 교정 방식의 교정 규칙의 정확도 임계값을 70%로, 통계 기반 교정 방식의 교정 변환 확률 임계값을 0% 설정하였을 경우 가장 높은 F1 척도 값(76.18)을 보였다.

[0068] 반면, 도 5는 트윗 말뭉치에 대해 파라미터를 변경한 실험 결과를 그래프로 표현하였으며, 규칙 기반 교정 방식의 교정 규칙의 정확도 임계값을 60%로, 통계 기반 교정 방식의 교정 변환 확률 임계값을 10% 설정하였을 경우 가장 높은 F1 척도 값(45.28)을 보였다.

[0069] 두 말뭉치 상에서 최고 성능을 보인 혼합 방식이 변환 확률 임계값 설정에서 차이를 보인 이유는 트윗 말뭉치의 경우 전체 어절 대비 교정 어절 비율이 낮아 교정 쌍 변환 후부가 매우 적었기 때문이다.

[0070] 전술한 바와 같이, 본 발명의 일실시예에 의한 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치는 많은 철자를 정확하게 교정할 수 있다.

[0071] 본 발명의 일실시예에 의하면, 기존 방식에 비해 오류 교정이 더 잘 된 문장을 생성할 수 있기에, 다른 자연어

처리 기술의 전처리 수단으로 유용하게 사용될 수 있다.

[0072] 상술한 소셜 텍스트를 위한 철자 오류 교정 방법 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터로 판독 가능한 기록 매체에 기록될 수 있다. 이때, 컴퓨터로 판독 가능한 기록매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 한편, 기록매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0073] 컴퓨터로 판독 가능한 기록매체에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0074] 한편, 이러한 기록매체는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수도 있다.

[0075] 또한, 프로그램 명령에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0076] 상기와 같이 설명된 소셜 텍스트를 위한 철자 오류 교정 방법 및 장치 상기 설명된 실시예들의 구성과 방법이 한정되게 적용될 수 있는 것이 아니라, 상기 실시예들은 다양한 변형이 이루어질 수 있도록 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성될 수도 있다.

부호의 설명

[0077] 100: 철자 오류 교정 장치

110: 데이터베이스

120: 입력부

130: 제1교정부

140: 제2교정부

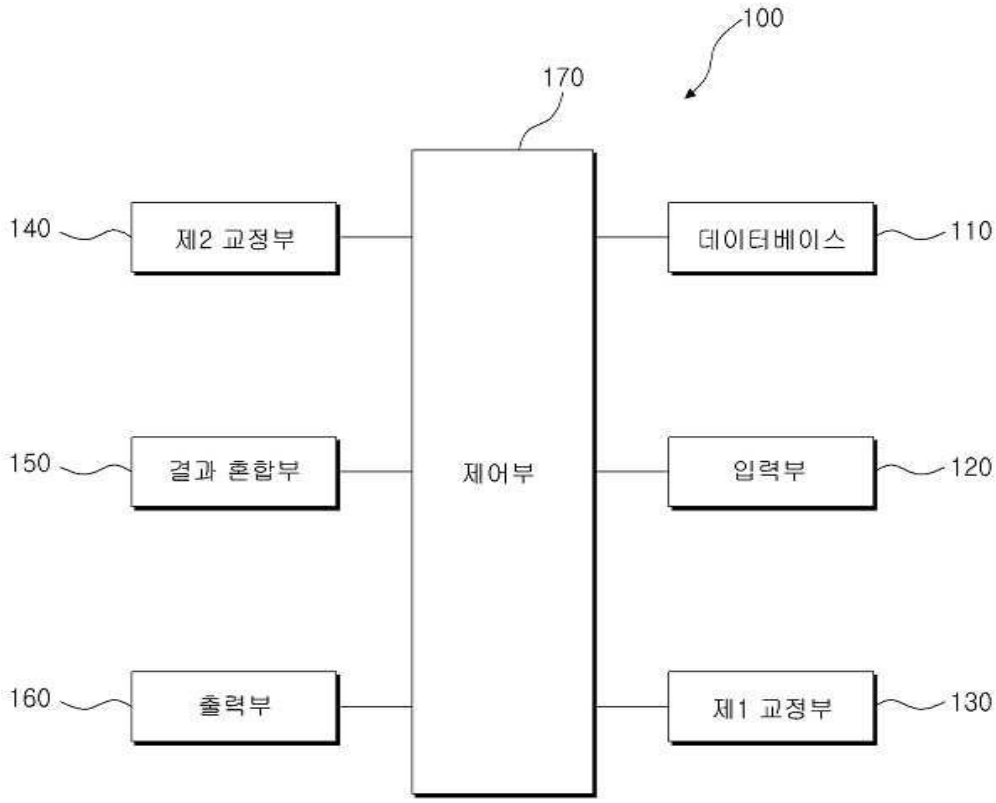
150: 결과 혼합부

160: 출력부

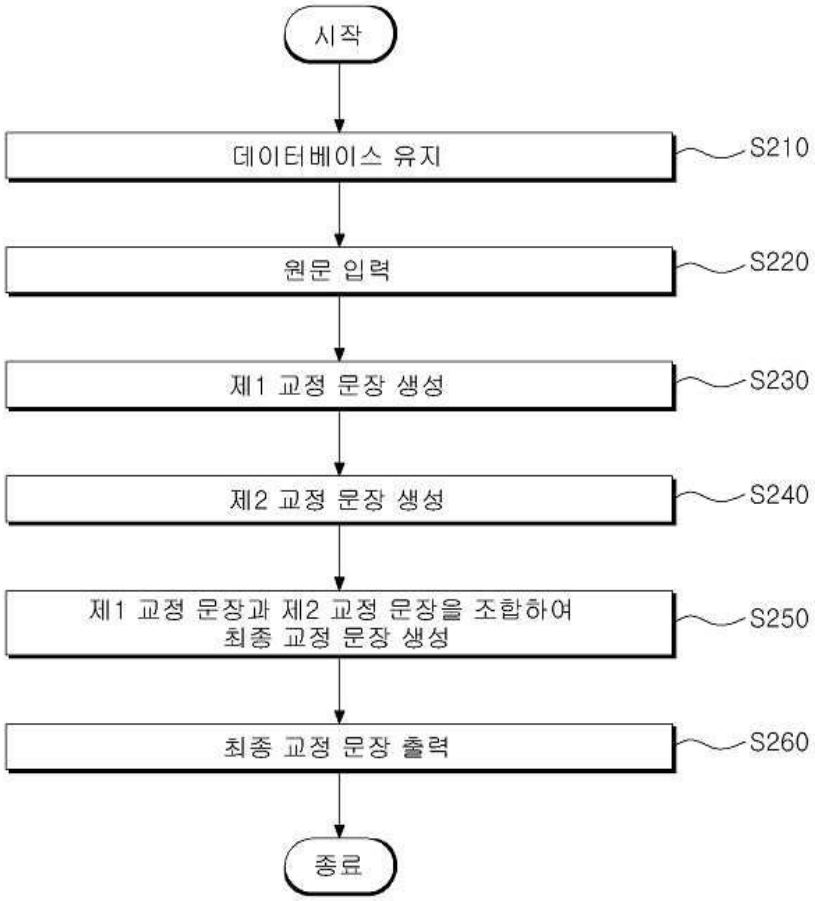
170: 제어부

도면

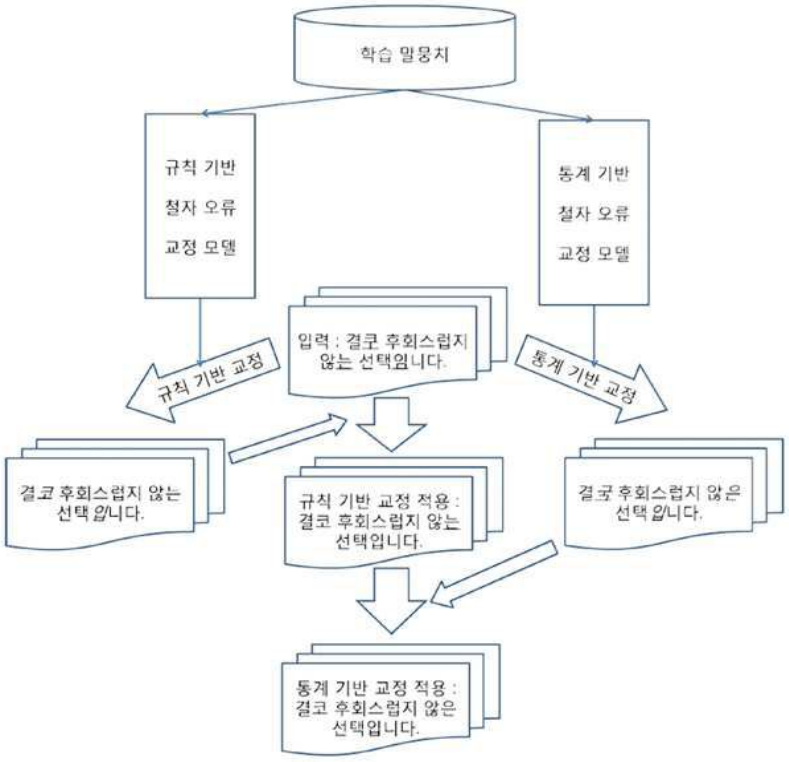
도면1


100
170
140
제2 교정부
데이터베이스
110
150
결과 혼합부
제어부
입력부
120
160
출력부
제1 교정부
130

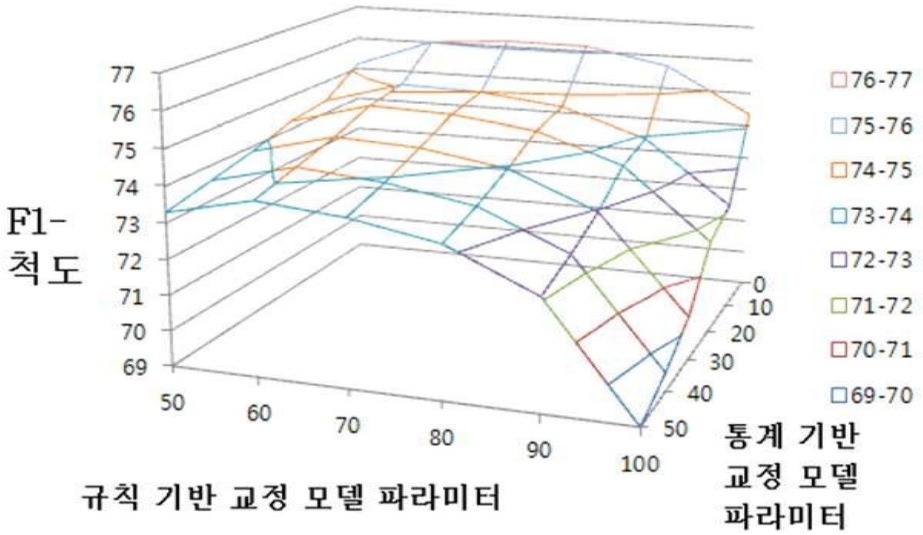
도면2


시작
데이터베이스 유지
S210
원문 입력
S220
제1 교정 문장 생성
S230
제2 교정 문장 생성
S240
제1 교정 문장과 제2 교정 문장을 조합하여
최종 교정 문장 생성
S250
최종 교정 문장 출력
S260
종료

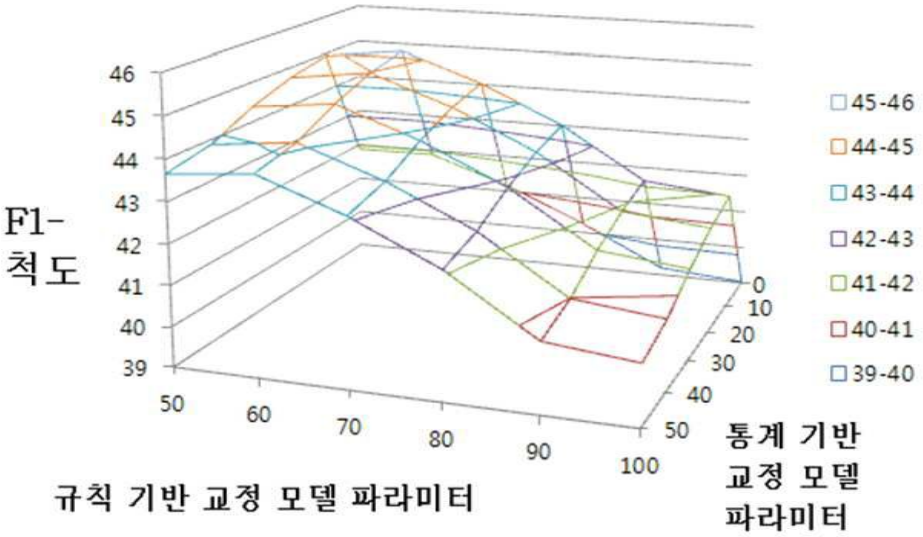
도면3


학습 말뭉치
규칙 기반
철자 오류
교정 모델
통계 기반
철자 오류
교정 모델
입력 : 결코 후회스럽지 않는 선택임니다.
규칙 기반 교정
통계 기반 교정
결코 후회스럽지 않는 선택입니다.
규칙 기반 교정 적용 : 결코 후회스럽지 않는 선택입니다.
결코 후회스럽지 않은 선택입니다.
통계 기반 교정 적용 : 결코 후회스럽지 않은 선택입니다.

도면4


F1-척도
77
76
75
74
73
72
71
70
69
50
60
70
80
90
100
규칙 기반 교정 모델 파라미터
0
10
20
30
40
50
통계 기반 교정 모델 파라미터
76-77
75-76
74-75
73-74
72-73
71-72
70-71
69-70

도면5


F1-척도
46
45
44
43
42
41
40
39
50
60
70
80
90
100
0
10
20
30
40
50
45-46
44-45
43-44
42-43
41-42
40-41
39-40
규칙 기반 교정 모델 파라미터
통계 기반 교정 모델 파라미터