(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0083111
(43) 공개일자 2020년07월08일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01) *G06N 20/00* (2019.01)
(52) CPC특허분류
G06F 40/205 (2020.01)
G06N 20/00 (2019.01)
(21) 출원번호 10-2019-0030688
(22) 출원일자 2019년03월18일
심사청구일자 2019년03월18일
(30) 우선권주장
1020180174248 2018년12월31일 대한민국(KR)

(71) 출원인
주식회사 엘솔루
서울특별시 서초구 마방로 10길 5, 14층 (양재동, 태석빌딩)
(72) 발명자
최종근
경기도 화성시 영통로26번길 24, 303동 802호(반월동, 반달마을 푸르지오)
이수미
서울특별시 서초구 언남7길 5, 101호(양재동, 아원데코빌)
김동필
서울특별시 강남구 논현로67길 13-4(역삼동)
(74) 대리인
유미특허법인

전체 청구항 수 : 총 22 항

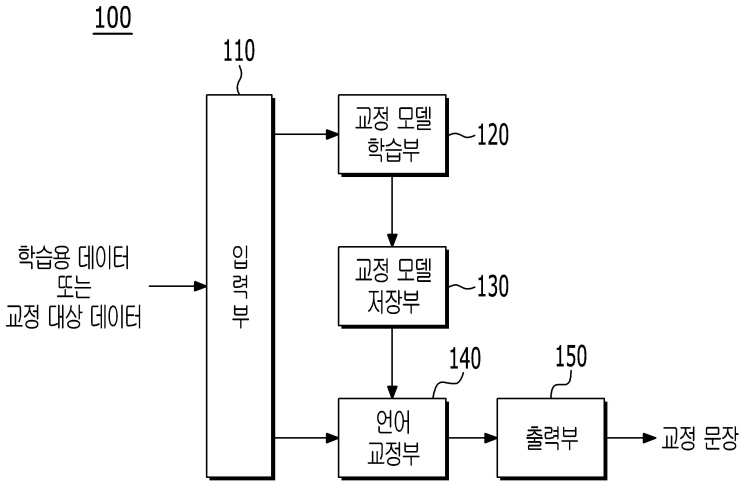
(54) 발명의 명칭 **언어 교정 시스템 및 그 방법과, 그 시스템에서의 언어 교정 모델 학습 방법**

(57) 요 약

언어 교정 시스템 및 그 방법과, 그 시스템에서의 언어 교정 모델 학습 방법이 개시된다.

이 시스템은 교정 모델 학습부 및 언어 교정부를 포함한다. 교정 모델 학습부는 비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합을 기계학습하여 교정 대상의 비문 데이터에 대응되는 정문 데이터를 검출하기 위한 교정 모델을 생성한다. 언어 교정부는 교정 대상의 문장에 대해 상기 교정 모델 학습부에 의해 생성된 교정 모델을 사용하여 대응되는 교정 문장을 생성하고, 생성되는 교정 문장과 함께 교정된 부분을 표시하여 출력한다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

기계학습 기반의 언어 교정 시스템으로서,

비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합을 기계학습하여 교정 대상의 비문 데이터에 대응되는 정문 데이터를 검출하기 위한 교정 모델을 생성하는 교정 모델 학습부; 및

교정 대상의 문장에 대해 상기 교정 모델 학습부에 의해 생성된 교정 모델을 사용하여 대응되는 교정 문장을 생성하고, 생성되는 교정 문장과 함께 교정된 부분을 표시하여 출력하는 언어 교정부

를 포함하는 언어 교정 시스템.

청구항 2

제1항에 있어서,

상기 교정 모델 학습부는,

상기 비문 데이터에 대해 언어 감지를 수행하여 단일어 문장으로의 필터링 작업, 데이터의 정제 및 정규화 작업을 수행하는 전처리부;

상기 전처리부에 필터링된 복수의 데이터 집합에 대해, 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장 작업 및 기계학습용 병렬 데이터 구축 작업을 수행하는 학습 가공부;

상기 학습 가공부에 의해 가공된 복수의 데이터 집합에 대해 지도 학습 기반의 기계학습을 수행하여 대응되는 상기 교정 모델을 생성하는 교정 학습부; 및

상기 학습 가공부에서의 지도 학습 데이터 레이블링 작업시 추가된 태그 부가 정보를 통해 오류 및 오류 카테고리 정보를 출력한 후 해당 태그 부가 정보를 제거하는 제1 후처리부

를 포함하는, 언어 교정 시스템.

청구항 3

제2항에 있어서,

상기 학습 가공부에서의 기계학습 데이터 확장 작업은,

상기 비문 데이터에 포함된 문자를 타이핑하기 위한 키보드의 정위치를 기준으로 주변의 오타 문자로 형성되는 글자를 사용한 데이터 확장 작업

을 포함하는, 언어 교정 시스템.

청구항 4

제2항에 있어서,

상기 학습 가공부에서의 기계학습용 병렬 데이터 구축 작업은,

교정이 불필요한 비문 문장과 이에 대응되는 정문 문장을 쌍으로 하는 병렬 코퍼스로 병렬 데이터를 구축하는 작업

을 포함하는, 언어 교정 시스템.

청구항 5

제2항에 있어서,

상기 교정 학습부는 상기 지도 학습 기반의 기계학습시 학습 결과에 대한 오류 출현 확률값을 비문 데이터와 정문 데이터와의 어텐션(attention) 가중치 정보로서 제공하는,

언어 교정 시스템.

청구항 6

제2항에 있어서,

입력 문장에 대해 미리 설정된 언어로 번역을 수행하는 번역 엔진을 더 포함하고,

상기 전처리부는,

상기 복수의 데이터 집합 내의 다량의 비문 데이터에 대해 상기 번역 엔진을 통한 번역을 수행하면서 상기 번역 엔진이 사용하는 사전에 등록되지 않은 단어에 대해 미리 설정된 마커를 사용하여 표시하고,

상기 다량의 비문 데이터에 대한 번역 완료 후, 상기 미리 설정된 마커에 의해 표시된 단어들을 추출하여 오류가 없는 단어로 일괄적으로 교정하는 선(先)교정을 수행하는,

언어 교정 시스템.

청구항 7

제6항에 있어서,

상기 전처리부는,

상기 미리 설정된 마커에 의해 표시된 단어들을 추출하면서 빈도수를 파악하고, 파악된 빈도수 기반으로 상기 미리 설정된 마커에 의해 표시된 단어들을 정렬한 후 오류가 없는 단어로 일괄적으로 교정하는,

언어 교정 시스템.

청구항 8

제1항에 있어서,

상기 언어 교정부는,

교정 대상의 문장에 대해, 문장 단위의 문장 분리를 수행하고, 분리된 문장을 토큰화하는 전처리를 수행하는 전처리부;

상기 전처리부에 의해 전처리가 수행된 교정 대상의 문장에 대해 바이너리 분류기를 사용하여 오류 문장과 비오류 문장을 구분하는 오류 문장 검출부;

상기 오류 문장 검출부에 의해 오류 문장으로 구분되는 경우, 상기 교정 대상의 문장에 대해 철자 오류 교정을 수행하는 맞춤법 교정부;

상기 맞춤법 교정부에 의해 철자 오류 교정된 문장에 대해 상기 교정 모델을 사용하여 문법 교정을 위한 언어 교정을 수행하여 교정 문장을 생성하는 문법 교정부; 및

상기 문법 교정부에 의한 언어 교정시 교정된 부분을 표시하는 후처리를 수행하여 상기 교정 문장과 함께 출력하는 후처리부

를 포함하는, 언어 교정 시스템.

청구항 9

제8항에 있어서,

상기 오류 문장 검출부는 상기 교정 대상의 문장 구분시 파악되는 신뢰도 정보에 따라 상기 오류 문장과 상기 비오류 문장을 구분하는,

언어 교정 시스템.

청구항 10

제8항에 있어서,

상기 맞춤법 교정부는 철자 오류 교정시 철자 오류 출현 확률값을 신뢰도 정보로서 제공하고,

상기 문법 교정부는 상기 철자 오류 교정된 문장에 대한 언어 교정의 어텐션 가중치를 통한 확률값을 신뢰도 정보로서 제공하며,

상기 후처리부는 상기 맞춤법 교정부에서 제공되는 신뢰도 정보와 상기 문법 교정부에서 제공되는 신뢰도 정보를 조합하여 상기 교정 대상의 문장에 대한 언어 교정의 최종 신뢰도 정보로서 제공하는,

언어 교정 시스템.

청구항 11

제10항에 있어서,

상기 문법 교정부와 상기 후처리부 사이에,

상기 문법 교정부에서 생성되는 교정 문장에 대해 미리 설정된 추천 문장을 사용하여 언어 모델링을 수행하는 언어 모델링부

를 더 포함하고,

상기 언어 모델링부는 언어 모델링시 언어 모델의 퍼플릭시티(perplexity)와 상호 정보(Mutual Information, MI) 값의 조합으로 상기 교정 문장의 신뢰도 정보를 제공하며,

상기 후처리부는 상기 최종 신뢰도 제공시 상기 언어 모델링부에서 제공되는 신뢰도 정보도 함께 조합하는,

언어 교정 시스템.

청구항 12

제1항에 있어서,

사용자에 의해 등록된 소스 단어와 이에 대응되는 타겟 단어를 포함하는 사용자 사전을 더 포함하고 - 상기 소스 단어와 타겟 단어는 각각 적어도 하나의 단어임 -,

상기 교정 모델 학습부는 상기 복수의 데이터 집합 내에 상기 사용자 사전에 등록된 단어가 포함되어 있는 경우 해당 단어에 대해 미리 설정된 사용자 사전 마커로 대체하여 기계학습을 수행하고,

상기 언어 교정부는 교정 대상의 문장에 상기 사용자 사전에 포함된 단어가 있는 경우 상기 사용자 사전 마커로 대체하여 상기 교정 대상의 문장에 대한 언어 교정을 수행하고, 교정된 문장 내에 상기 사용자 사전 마커가 포함되어 있는 경우 상기 사용자 사전 마커를 상기 교정 대상의 문장에서 대응되는 단어에 대응하여 상기 사용자 사전에 등록되어 있는 단어로 대체하는,

언어 교정 시스템.

청구항 13

언어 교정 시스템이 기계 학습 기반으로 언어 교정 모델을 학습하는 방법으로서,

비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합에 대해, 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장 작업 및 기계학습용 병렬 데이터 구축 작업을 포함하는 학습 가공을 수행하는 단계; 및

상기 학습 가공이 수행된 복수의 데이터 집합에 대해 지도 학습 기반의 기계학습을 수행하여 대응되는 교정 모델을 생성하는 단계

를 포함하는 언어 교정 모델 학습 방법.

청구항 14

제13항에 있어서,

상기 기계학습 데이터 확장 작업은,

상기 비문 데이터에 포함된 문자를 타이핑하기 위한 키보드의 정위치를 기준으로 주변의 오타 문자로 형성되는 글자를 사용한 데이터 확장 작업

을 포함하고,

상기 기계학습용 병렬 데이터 구축 작업은,

교정이 불필요한 비문 문장과 이에 대응되는 정문 문장을 쌍으로 하는 병렬 코퍼스로 병렬 데이터를 구축하는 작업

을 포함하는, 언어 교정 모델 학습 방법.

청구항 15

제13항에 있어서,

상기 학습 가공을 수행하는 단계 전에,

상기 복수의 데이터 집합에 대해 언어 감지를 수행하여 단일어 문장으로의 필터링 작업, 데이터의 정제 및 정규화 작업을 포함하는 전처리를 수행하는 단계

를 더 포함하고,

상기 전처리를 수행하는 단계는,

상기 복수의 데이터 집합 내의 다량의 비문 데이터에 대해 번역 엔진을 통한 번역을 수행하는 단계;

상기 번역 엔진이 사용하는 사전에 등록되지 않은 단어에 대해 미리 설정된 마커를 사용하여 표시하는 단계;

상기 다량의 비문 데이터에 대한 번역 완료 후, 상기 미리 설정된 마커에 의해 표시된 단어들을 추출하는 단계; 및

추출된 단어들에 대해 오류가 없는 단어로 일괄적으로 교정하는 단계

를 포함하는, 언어 교정 모델 학습 방법.

청구항 16

제15항에 있어서,

상기 일괄적으로 교정하는 단계는,

상기 미리 설정된 마커에 의해 표시된 단어들을 추출하는 단계;

추출된 단어들의 빈도수를 파악하는 단계;

파악된 빈도수 기반으로 상기 미리 설정된 마커에 의해 표시된 단어들을 정렬하는 단계; 및

정렬된 단어들에 대해 오류가 없는 단어로 일괄적으로 교정하는 단계

를 포함하는, 언어 교정 모델 학습 방법.

청구항 17

제13항에 있어서,

상기 언어 교정 시스템이,

사용자에 의해 등록된 소스 단어와 이에 대응되는 타겟 단어를 포함하는 사용자 사전을 더 포함하고 - 상기 소스 단어와 타겟 단어는 각각 적어도 하나의 단어임 -,

상기 교정 모델을 생성하는 단계는,

상기 복수의 데이터 집합 내에 상기 사용자 사전에 등록된 단어가 포함되어 있는 경우 해당 단어에 대해 미리 설정된 사용자 사전 마커로 대체하여 기계학습을 수행하여 상기 교정 모델을 생성하는,

언어 교정 모델 학습 방법.

청구항 18

언어 교정 시스템이 기계 학습 기반으로 언어 교정하는 방법으로서,

언어 교정 대상의 문장에 대해 철자 오류 교정을 수행하는 단계; 및

철자 오류 교정된 문장에 대해 교정 모델을 사용하여 문법 교정을 수행하여 교정 문장을 생성하는 단계

를 포함하고,

상기 교정 모델은 비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합에 대해 지도 학습 기반의 기계학습을 수행하여 생성된 것인,

언어 교정 방법.

청구항 19

제18항에 있어서,

상기 철자 오류 교정을 수행하는 단계 전에,

상기 언어 교정 대상의 문장에 대해, 문장 단위의 문장 분리를 수행하고, 분리된 문장을 토큰화하는 전처리를 수행하는 단계; 및

상기 전처리가 수행된 언어 교정 대상의 문장에 대해 바이너리 분류기를 사용하여 오류 문장과 비오류 문장을 구분하는 단계

를 더 포함하고,

상기 오류 문장과 비오류 문장을 구분하는 단계에서, 상기 언어 교정 대상의 문장이 오류 문장으로 구분되는 경우 상기 철자 오류 교정을 수행하는 단계가 수행되는,

언어 교정 방법.

청구항 20

제19항에 있어서,

상기 오류 문장과 비오류 문장을 구분하는 단계에서,

상기 언어 교정 대상의 문장 구분시 파악되는 신뢰도 정보에 따라 상기 오류 문장과 상기 비오류 문장을 구분하는,

언어 교정 방법.

청구항 21

제18항에 있어서,

상기 교정 문장을 생성하는 단계 후에,

상기 교정 문장에 대해 미리 설정된 추천 문장을 사용하여 언어 모델링을 수행하는 단계; 및

상기 교정 문장 생성시 교정된 부분을 표시하는 후처리를 수행하여 상기 교정 문장과 함께 출력하는 단계

를 더 포함하는, 언어 교정 방법.

청구항 22

제18항에 있어서,

상기 언어 교정 시스템이,

사용자에 의해 등록된 소스 단어와 이에 대응되는 타겟 단어를 포함하는 사용자 사전을 더 포함하고 - 상기 소

스 단어와 타겟 단어는 각각 적어도 하나의 단어임 -,

상기 철자 오류 교정을 수행하는 단계 전에,

상기 언어 교정 대상의 문장 내에 상기 사용자 사전에 포함되어 있는 단어가 포함되어 있는지를 판단하는 단계; 및

상기 언어 교정 대상의 문장 내에 상기 사용자 사전에 포함되어 있는 단어가 포함되어 있는 경우, 상기 사용자 사전과 상기 언어 교정 대상의 문장에 공통적으로 포함되어 있는 단어를 미리 설정된 사용자 사전 마커로 대체하는 단계

를 더 포함하고,

상기 교정 문장을 생성하는 단계 후에,

생성된 교정 문장 내에 상기 사용자 사전 마커가 포함되어 있는지를 확인하는 단계;

상기 생성된 교정 문장 내에 상기 사용자 사전 마커가 포함되어 있는 경우, 포함된 사용자 사전 마커의 위치에 대응되는 상기 언어 교정 대상의 문장 내 단어에 대응되는 상기 사용자 사전의 단어로 대체하여 최종의 교정 문장을 생성하는 단계

를 더 포함하는, 언어 교정 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 언어 교정 시스템 및 그 방법과, 그 시스템에서의 언어 교정 모델 학습 방법에 관한 것이다.

배 경 기 술

[0002] 언어 교정은 다양한 형태의 언어로 작성된 문장, 예를 들면 인터넷에서 작성되거나 인터넷을 통해 배포되는 언어 문장, 즉 인터넷 데이터에 대해 맞춤법 오류나 문법 오류를 교정하는 것을 말한다. 이러한 언어 교정에는 맞춤법이나 문법에 어긋난 표현을 교정하는 것뿐만 아니라 문장을 보다 깔끔하고 읽기 쉽게 다듬는 교정 또한 포함될 수 있다.

[0003] 상기한 언어 교정은 언어 학습에 사용되거나, 또는 서적이나 신문 기사 등의 텍스트 출판물이 일정 수준을 유지할 수 있도록 하는 작업은 물론 다양한 형태로서 언어 교정이 필요한 영역에서 사용될 수 있다.

[0004] 특히, 최근 인터넷을 통해서 대용량의 언어 데이터가 유통되거나 사용되는데, 종래의 언어 교정은 통계 모델을 사용함으로써 단순한 형태의 맞춤법이나 문법을 위주로 하는 언어 교정이 수행되어 최근의 대용량의 언어 데이터에 대해 보다 효율적인 언어 교정의 필요성이 대두되고 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명이 해결하고자 하는 과제는 기계학습 기반의 교정 모델을 사용함으로써 효율적인 언어 교정 결과를 제공할 수 있는 언어 교정 시스템 및 그 방법과, 그 시스템에서의 언어 교정 모델 학습 방법을 제공하는 것이다.

과제의 해결 수단

[0006] 본 발명의 하나의 특징에 따른 언어 교정 시스템은,

[0007] 기계학습 기반의 언어 교정 시스템으로서, 비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합을 기계학습하여 교정 대상의 비문 데이터에 대응되는 정문 데이터를 검출하기 위한 교정 모델을 생성하는 교정 모델 학습부; 및 교정 대상의 문장에 대해 상기 교정 모델 학습부에 의해 생성된 교정 모델을 사용하여 대응되는 교정 문장을 생성하고, 생성되는 교정 문장과 함께 교정된 부분을 표시하여 출력하는 언어 교정부를 포함한다.

[0008] 여기서, 상기 교정 모델 학습부는, 상기 비문 데이터에 대해 언어 감지를 수행하여 단일어 문장으로의 필터링

작업, 데이터의 정제 및 정규화 작업을 수행하는 전처리부; 상기 전처리부에 필터링된 복수의 데이터 집합에 대해, 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장 작업 및 기계학습용 병렬 데이터 구축 작업을 수행하는 학습 가공부; 상기 학습 가공부에 의해 가공된 복수의 데이터 집합에 대해 지도 학습 기반의 기계학습을 수행하여 대응되는 상기 교정 모델을 생성하는 교정 학습부; 및 상기 학습 가공부에서의 지도 학습 데이터 레이블링 작업시 추가된 태그 부가 정보를 통해 오류 및 오류 카테고리 정보를 출력한 후 해당 태그 부가 정보를 제거하는 제1 후처리부를 포함한다.

[0009] 또한, 상기 학습 가공부에서의 기계학습 데이터 확장 작업은, 상기 비문 데이터에 포함된 문자를 타이핑하기 위한 키보드의 정위치를 기준으로 주변의 오타 문자로 형성되는 글자를 사용한 데이터 확장 작업을 포함한다.

[0010] 또한, 상기 학습 가공부에서의 기계학습용 병렬 데이터 구축 작업은, 교정이 불필요한 비문 문장과 이에 대응되는 정문 문장을 쌍으로 하는 병렬 코퍼스로 병렬 데이터를 구축하는 작업을 포함한다.

[0011] 또한, 상기 교정 학습부는 상기 지도 학습 기반의 기계학습시 학습 결과에 대한 오류 출현 확률값을 비문 데이터와 정문 데이터와의 어텐션(attention) 가중치 정보로서 제공한다.

[0012] 또한, 입력 문장에 대해 미리 설정된 언어로 번역을 수행하는 번역 엔진을 더 포함하고, 상기 전처리부는, 상기 복수의 데이터 집합 내의 다량의 비문 데이터에 대해 상기 번역 엔진을 통한 번역을 수행하면서 상기 번역 엔진이 사용하는 사전에 등록되지 않은 단어에 대해 미리 설정된 마커를 사용하여 표시하고, 상기 다량의 비문 데이터에 대한 번역 완료 후, 상기 미리 설정된 마커에 의해 표시된 단어들을 추출하여 오류가 없는 단어로 일괄적으로 교정하는 선(先)교정을 수행한다.

[0013] 또한, 상기 전처리부는, 상기 미리 설정된 마커에 의해 표시된 단어들을 추출하면서 빈도수를 파악하고, 파악된 빈도수 기반으로 상기 미리 설정된 마커에 의해 표시된 단어들을 정렬한 후 오류가 없는 단어로 일괄적으로 교정한다.

[0014] 또한, 상기 언어 교정부는, 교정 대상의 문장에 대해, 문장 단위의 문장 분리를 수행하고, 분리된 문장을 토큰화하는 전처리를 수행하는 전처리부; 상기 전처리부에 의해 전처리가 수행된 교정 대상의 문장에 대해 바이너리 분류기를 사용하여 오류 문장과 비오류 문장을 구분하는 오류 문장 검출부; 상기 오류 문장 검출부에 의해 오류 문장으로 구분되는 경우, 상기 교정 대상의 문장에 대해 철자 오류 교정을 수행하는 맞춤법 교정부; 상기 맞춤법 교정부에 의해 철자 오류 교정된 문장에 대해 상기 교정 모델을 사용하여 문법 교정을 위한 언어 교정을 수행하여 교정 문장을 생성하는 문법 교정부; 및 상기 문법 교정부에 의한 언어 교정시 교정된 부분을 표시하는 후처리를 수행하여 상기 교정 문장과 함께 출력하는 후처리부를 포함한다.

[0015] 또한, 상기 오류 문장 검출부는 상기 교정 대상의 문장 구분시 파악되는 신뢰도 정보에 따라 상기 오류 문장과 상기 비오류 문장을 구분한다.

[0016] 또한, 상기 맞춤법 교정부는 철자 오류 교정시 철자 오류 출현 확률값을 신뢰도 정보로서 제공하고, 상기 문법 교정부는 상기 철자 오류 교정된 문장에 대한 언어 교정의 어텐션 가중치를 통한 확률값을 신뢰도 정보로서 제공하며, 상기 후처리부는 상기 맞춤법 교정부에서 제공되는 신뢰도 정보와 상기 문법 교정부에서 제공되는 신뢰도 정보를 조합하여 상기 교정 대상의 문장에 대한 언어 교정의 최종 신뢰도 정보로서 제공한다.

[0017] 또한, 상기 문법 교정부와 상기 후처리부 사이에, 상기 문법 교정부에서 생성되는 교정 문장에 대해 미리 설정된 추천 문장을 사용하여 언어 모델링을 수행하는 언어 모델링부를 더 포함하고, 상기 언어 모델링부는 언어 모델링시 언어 모델의 퍼플릭시티(perplexity)와 상호 정보(Mutual Information, MI) 값의 조합으로 상기 교정 문장의 신뢰도 정보를 제공하며, 상기 후처리부는 상기 최종 신뢰도 제공시 상기 언어 모델링부에서 제공되는 신뢰도 정보도 함께 조합한다.

[0018] 또한, 사용자에 의해 등록된 소스 단어와 이에 대응되는 타겟 단어를 포함하는 사용자 사전을 더 포함하고 - 상기 소스 단어와 타겟 단어는 각각 적어도 하나의 단어임 -, 상기 교정 모델 학습부는 상기 복수의 데이터 집합 내에 상기 사용자 사전에 등록된 단어가 포함되어 있는 경우 해당 단어에 대해 미리 설정된 사용자 사전 마커로 대체하여 기계학습을 수행하고, 상기 언어 교정부는 교정 대상의 문장에 상기 사용자 사전에 포함된 단어가 있는 경우 상기 사용자 사전 마커로 대체하여 상기 교정 대상의 문장에 대한 언어 교정을 수행하고, 교정된 문장 내에 상기 사용자 사전 마커가 포함되어 있는 경우 상기 사용자 사전 마커를 상기 교정 대상의 문장에서 대응되는 단어에 대응하여 상기 사용자 사전에 등록되어 있는 단어로 대체한다.

[0019] 본 발명의 다른 특징에 따른 언어 교정 모델 학습 방법은,

[0020] 언어 교정 시스템이 기계 학습 기반으로 언어 교정 모델을 학습하는 방법으로서, 비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합에 대해, 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장 작업 및 기계학습용 병렬 데이터 구축 작업을 포함하는 학습 가공을 수행하는 단계; 및 상기 학습 가공이 수행된 복수의 데이터 집합에 대해 지도 학습 기반의 기계학습을 수행하여 대응되는 교정 모델을 생성하는 단계를 포함한다.

[0021] 여기서, 상기 기계학습 데이터 확장 작업은, 상기 비문 데이터에 포함된 문자를 타이핑하기 위한 키보드의 정위치를 기준으로 주변의 오타 문자로 형성되는 글자를 사용한 데이터 확장 작업을 포함하고, 상기 기계학습용 병렬 데이터 구축 작업은, 교정이 불필요한 비문 문장과 이에 대응되는 정문 문장을 쌍으로 하는 병렬 코퍼스로 병렬 데이터를 구축하는 작업을 포함한다.

[0022] 또한, 상기 학습 가공을 수행하는 단계 전에, 상기 복수의 데이터 집합에 대해 언어 감지를 수행하여 단일어 문장으로의 필터링 작업, 데이터의 정제 및 정규화 작업을 포함하는 전처리를 수행하는 단계를 더 포함하고, 상기 전처리를 수행하는 단계는, 상기 복수의 데이터 집합 내의 다량의 비문 데이터에 대해 번역 엔진을 통한 번역을 수행하는 단계; 상기 번역 엔진이 사용하는 사전에 등록되지 않은 단어에 대해 미리 설정된 마커를 사용하여 표시하는 단계; 상기 다량의 비문 데이터에 대한 번역 완료 후, 상기 미리 설정된 마커에 의해 표시된 단어들을 추출하는 단계; 및 추출된 단어들에 대해 오류가 없는 단어로 일괄적으로 교정하는 단계를 포함한다.

[0023] 또한, 상기 일괄적으로 교정하는 단계는, 상기 미리 설정된 마커에 의해 표시된 단어들을 추출하는 단계; 추출된 단어들의 빈도수를 파악하는 단계; 파악된 빈도수 기반으로 상기 미리 설정된 마커에 의해 표시된 단어들을 정렬하는 단계; 및 정렬된 단어들에 대해 오류가 없는 단어로 일괄적으로 교정하는 단계를 포함한다.

[0024] 또한, 상기 언어 교정 시스템이, 사용자에 의해 등록된 소스 단어와 이에 대응되는 타겟 단어를 포함하는 사용자 사전을 더 포함하고 - 상기 소스 단어와 타겟 단어는 각각 적어도 하나의 단어임 -, 상기 교정 모델을 생성하는 단계는, 상기 복수의 데이터 집합 내에 상기 사용자 사전에 등록된 단어가 포함되어 있는 경우 해당 단어에 대해 미리 설정된 사용자 사전 마커로 대체하여 기계학습을 수행하여 상기 교정 모델을 생성한다.

[0025] 본 발명의 또 다른 특징에 따른 언어 교정 방법은,

[0026] 언어 교정 시스템이 기계 학습 기반으로 언어 교정하는 방법으로서, 언어 교정 대상의 문장에 대해 철자 오류 교정을 수행하는 단계; 및 철자 오류 교정된 문장에 대해 교정 모델을 사용하여 문법 교정을 수행하여 교정 문장을 생성하는 단계를 포함하고, 상기 교정 모델은 비문(非文) 데이터와 상기 비문 데이터에 각각 대응되는 오류가 없는 정문(正文) 데이터로 구성된 복수의 데이터 집합에 대해 지도 학습 기반의 기계학습을 수행하여 생성된 것이다.

[0027] 여기서, 상기 철자 오류 교정을 수행하는 단계 전에, 상기 언어 교정 대상의 문장에 대해, 문장 단위의 문장 분리를 수행하고, 분리된 문장을 토큰화하는 전처리를 수행하는 단계; 및 상기 전처리가 수행된 언어 교정 대상의 문장에 대해 바이너리 분류기를 사용하여 오류 문장과 비오류 문장을 구분하는 단계를 더 포함하고, 상기 오류 문장과 비오류 문장을 구분하는 단계에서, 상기 언어 교정 대상의 문장이 오류 문장으로 구분되는 경우 상기 철자 오류 교정을 수행하는 단계가 수행된다.

[0028] 또한, 상기 오류 문장과 비오류 문장을 구분하는 단계에서, 상기 언어 교정 대상의 문장 구분시 파악되는 신뢰도 정보에 따라 상기 오류 문장과 상기 비오류 문장을 구분한다.

[0029] 또한, 상기 교정 문장을 생성하는 단계 후에, 상기 교정 문장에 대해 미리 설정된 추천 문장을 사용하여 언어 모델링을 수행하는 단계; 및 상기 교정 문장 생성시 교정된 부분을 표시하는 후처리를 수행하여 상기 교정 문장과 함께 출력하는 단계를 더 포함한다.

[0030] 또한, 상기 언어 교정 시스템이, 사용자에 의해 등록된 소스 단어와 이에 대응되는 타겟 단어를 포함하는 사용자 사전을 더 포함하고 - 상기 소스 단어와 타겟 단어는 각각 적어도 하나의 단어임 -, 상기 철자 오류 교정을 수행하는 단계 전에, 상기 언어 교정 대상의 문장 내에 상기 사용자 사전에 포함되어 있는 단어가 포함되어 있는지를 판단하는 단계; 및 상기 언어 교정 대상의 문장 내에 상기 사용자 사전에 포함되어 있는 단어가 포함되어 있는 경우, 상기 사용자 사전과 상기 언어 교정 대상의 문장에 공통적으로 포함되어 있는 단어를 미리 설정된 사용자 사전 마커로 대체하는 단계를 더 포함하고, 상기 교정 문장을 생성하는 단계 후에, 생성된 교정 문장 내에 상기 사용자 사전 마커가 포함되어 있는지를 확인하는 단계; 상기 생성된 교정 문장 내에 상기 사용자 사전 마커가 포함되어 있는 경우, 포함된 사용자 사전 마커의 위치에 대응되는 상기 언어 교정 대상의 문장 내 단

어에 대응되는 상기 사용자 사전의 단어로 대체하여 최종의 교정 문장을 생성하는 단계를 더 포함한다.

발명의 효과

[0031] 본 발명의 실시예에 따르면, 기계학습 기반의 교정 모델을 사용함으로써 효율적인 언어 교정 결과를 제공할 수 있다.

[0032] 또한, 언어 교육용 첨삭 지도에 활용하여 온라인 학습 시스템 개발이 가능하다.

[0033] 또한, 문장 단위의 검색에서 오타/문법 오류를 제거하여 검색 성능을 향상시킬 수 있다.

[0034] 또한, 각종 오피스 툴에 적용하여 문서 작성에 보조할 수 있다.

[0035] 또한, 사용자에 의해 미리 정의된 형태의 교정 정보를 변수 형태로 저장하고 이를 런타임에서 처리함으로써, 교정 모델에 별도로 추가하거나 변경하지 않고도 쉽게 언어 교정이 수행될 수 있다.

[0036] 또한, 교정이 어렵거나 의도적으로 잘 안되는 부분에 대해서도 사용자 사전에 등록하여 처리함으로써 언어 교정의 효율을 향상시킬 수 있다.

도면의 간단한 설명

[0037] 도 1은 본 발명의 실시예에 따른 언어 교정 시스템의 개략적인 구성도이다.

도 2는 도 1에 도시된 교정 모델 학습부의 구체적인 구성도이다.

도 3은 도 1에 도시된 언어 교정부의 구체적인 구성도이다.

도 4는 본 발명의 실시예에 따른 언어 교정 시스템에 의해 언어 교정이 수행된 결과 예를 도시한 도면이다.

도 5는 본 발명의 실시예에 따른 기계학습 기반 언어 교정 방법의 개략적인 흐름도이다.

도 6은 본 발명의 실시예에 따른 언어 교정 모델 학습 방법의 개략적인 흐름도이다.

도 7은 본 발명의 다른 실시예에 따른 교정 모델 학습부의 구체적인 구성도이다.

도 8은 본 발명의 다른 실시예에 따른 교정 모델 학습 문장의 선교정 방법의 흐름도이다.

도 9는 본 발명의 다른 실시예에 따른 교정 모델 학습 문장의 선교정 방법의 예를 도시한 도면이다.

도 10은 본 발명의 다른 실시예에 따른 언어 교정 시스템의 개략적인 구성도이다.

도 11은 도 10에 도시된 교정 모델 학습부의 구체적인 구성도이다.

도 12는 도 10에 도시된 언어 교정부의 구체적인 구성도이다.

도 13은 본 발명의 다른 실시예에 따른 언어 교정 모델 학습 방법의 흐름도이다.

도 14는 본 발명의 다른 실시예에 따른 언어 교정 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0038] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다

[0039] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0040] 또한, 명세서에 기재된 "…부", "…기", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0041] 이하, 도면을 참고하여 본 발명의 실시예에 따른 언어 교정 시스템에 대해 설명한다.

[0042] 도 1은 본 발명의 실시예에 따른 언어 교정 시스템의 개략적인 구성도이다.

[0043] 도 1에 도시된 바와 같이, 본 발명의 실시예에 따른 언어 교정 시스템(100)은 입력부(110), 교정 모델 학습부(120), 교정 모델 저장부(130), 언어 교정부(140) 및 출력부(150)를 포함한다. 여기서, 도 1에 도시된 언어 교정 시스템(100)은 본 발명의 하나의 실시예에 불과하므로 도 1을 통해 본 발명이 한정 해석되는 것은 아니며, 본 발명의 다양한 실시예들에 따라 도 1과 다르게 구성될 수도 있다.

[0044] 입력부(110)는 언어 교정의 학습을 위해 사용되는 데이터 또는 언어 교정 대상인 교정 대상 데이터를 입력받는다. 여기서, 언어 교정의 학습을 위해 사용되는 데이터로는 추후 설명될 지도 학습 기반의 기계학습을 위해 인터넷 대용량 데이터로서 교정 정보를 포함하는 비문(非文) 데이터와 오류가 없는 정문(正文) 데이터가 쌍으로서 입력된다.

[0045] 교정 모델 학습부(120)는 입력부(110)를 통해 입력되는 데이터 중에서 언어 교정의 학습을 위해 사용되는 데이터, 즉 비문 데이터와 정문 데이터의 쌍으로 구성된 대량의 학습 데이터를 사용하여 언어 교정을 위한 기계학습을 수행하여 언어 교정용 학습 모델인 교정 모델을 생성한다. 이 때, 교정 모델 학습부(120)에서 생성되는 교정 모델은 교정 모델 저장부(130)에 저장된다. 한편, 상기한 기계학습은 인공지능의 한 분야로서, 방대한 데이터를 분석해서 미래를 예측하는 기술이며, 컴퓨터가 스스로 학습 과정을 거치면서 입력되지 않은 정보를 습득하여 문제를 해결하는 기술이다. 기계학습을 위해 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), Transformer Networks 등의 신경망을 활용하는 딥러닝 기술이 사용될 수 있다. 이러한 기계학습 기술에 대해서는 이미 잘 알려져 있으므로 여기에서는 구체적인 설명을 생략한다.

[0046] 교정 모델 저장부(130)는 교정 모델 학습부(120)에 의한 기계학습을 통해 생성되는 교정 모델을 저장한다.

[0047] 언어 교정부(140)는 입력부(110)를 통해 입력되는 대용량의 언어 교정 데이터, 즉 맞춤법 오류나 문법 오류의 교정 대상인 교정 대상 데이터에 대해 교정 모델 저장부(130)에 저장된 교정 모델을 사용하여 교정 대상 데이터에 대한 맞춤법/문법 교정을 수행하고, 교정이 완료된 교정 데이터를 출력부(150)로 출력한다.

[0048] 선택적으로 언어 교정부(140)는 교정 대상 데이터에 대한 맞춤법/문법 교정이 완료되어 교정이 불필요한 경우라도 문장을 자연스러운 문장으로 교정해주는 언어 모델링 작업이 추가로 수행될 수 있다.

[0049] 출력부(150)는 언어 교정부(140)에서 언어 교정이 완료된 교정 데이터와 함께 교정 대상 데이터를 전달받아서 외부, 예를 들어 사용자에게 출력한다.

[0050] 또한, 출력부(150)는 교정 대상 데이터와 함께 이에 대응되는 교정 데이터를 함께 출력할 수 있다. 선택적으로, 출력부(150)는 교정 데이터에 대해 교정 대상 데이터에서 교정이 수행된 부분을 알 수 있도록 추가로 표시할 수 있다. 이 때, 교정이 수행된 부분에 대한 정보는 언어 교정부(140)에서 출력부(150)로 제공된다.

[0051] 한편, 상기한 교정 모델 학습부(120)와 언어 교정부(140)는 서로 통합되어 하나의 구성요소로서 구현될 수 있거나, 또는 서로 별개의 장치로서 구현될 수 있다. 예를 들어, 입력부(110), 교정 모델 학습부(120) 및 교정 모델 저장부(130)만을 포함하는 교정 모델 학습 장치와 입력부(110), 교정 모델 저장부(130), 언어 교정부(140) 및 출력부(150)만을 포함하는 언어 교정 장치와 같이 각각의 장치로서 구현될 수도 있다.

[0052] 이하, 전술한 교정 모델 학습부(120)에 대해 보다 구체적으로 설명한다.

[0053] 도 2는 도 1에 도시된 교정 모델 학습부(120)의 구체적인 구성도이다.

[0054] 도 2에 도시된 바와 같이, 교정 모델 학습부(120)는 전처리부(121), 학습 가공부(122), 교정 학습부(123), 후처리부(124) 및 교정 모델 출력부(125)를 포함한다.

[0055] 설명 전에, 본 발명의 실시예에서 수행되는 교정 모델의 기계학습은 지도 학습(supervised learning)을 사용하지만, 이에 한정되는 것은 아니다. 여기서 지도 학습은 입력과 출력 사이의 매핑을 학습하는 것이며, 입력과 출력 쌍이 데이터로 주어지는 경우에 적용된다. 본 발명의 실시예에 적용하면, 맞춤법 교정 및 문법 교정을 위한 소스 데이터인 비문 데이터가 입력이고, 이에 대응되어 교정된 문장에 해당되는 타겟 데이터인 정문 데이터가 출력에 해당된다. 이러한 지도 학습에 따른 기계학습 방법에 대해서는 이미 잘 알려져 있으므로 여기에서는 구체적인 설명을 생략한다.

[0056] 전처리부(121)는 입력부(110)를 통해 입력되는 언어 교정의 학습을 위해 사용되는 데이터, 즉 비문 데이터("소스 문장"이라고도 함)와 정문 데이터("타겟 문장"이라고도 함)의 쌍으로 구성된 학습 데이터 중에서 비문 및 정문 데이터에 대해 언어 감지(language identification) 기술을 적용하여 단일어 문장으로 필터링한다. 즉, 비문 혹은 정문 데이터는 기본적으로 동일한 언어 기반으로 학습을 수행할 수 있도록 언어 감지를 통해 단일어 문

장으로 필터링을 수행한다.

[0057] 선택적으로, 전처리부(121)는 언어 감지시 코드 스위칭 부분 필터링을 추가로 수행할 수 있다. 이것은, 예를 들어, “Korea는 굉장히 traditional한 thinking에 사로잡힌 것 같아요”와 같이 영어와 한국어가 섞여 있는 경우와 같이, 서로 다른 언어가 사용되더라도 코드 스위칭을 위한 언어 감지 기술을 통해 필터링되어 제거되지 않고 문장 내에 남아 있도록 하는 것이다.

[0058] 또한, 전처리부(121)는 비문 데이터에 대한 정제를 수행한다. 이러한 정제는 단일어 코퍼스(corpus) 또는 병렬 코퍼스에 적용될 수 있다.

[0059] 이외에도, 전처리부(121)는 소스/타켓 문장에서 중복 및 빈 정보 유무 체크, 글자/단어수 최대치/최소값 설정, 글자 및 단어 길이 공백 수 제한, 대문자 숫자 제한, 반복 단어 제한, 비 그래픽 문자(Non-graphic/Non-printable Character), 유니코드 처리 오류 체크, 외국어 비율 체크, 인코딩 유효성 확인 등의 작업을 더 수행할 수 있다. 이러한 작업에 대해서는 잘 알려져 있으므로, 여기에서는 구체적인 설명을 생략한다.

[0060] 또한, 전처리부(121)는 유니코드, 문장 부호, 대소문자, 지역별 철자가 다른 경우 등에 따라 데이터의 정규화를 추가로 수행할 수 있다. 이 때, 데이터의 정규화는 전술한 데이터의 정제와 통합될 수도 있다.

[0061] 학습 가공부(122)는 전처리부(121)에 의해 전처리가 수행된 데이터의 쌍, 즉 비문 데이터와 정문 데이터의 쌍을 사용하여 추후 교정 학습부(123)에 의해 수행되는 기계학습에 필요한 데이터를 준비하는 것으로, 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장(data augmentation) 작업, 기계학습용 병렬 데이터 구축 작업을 수행한다. 이러한 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장 작업 및 기계학습용 병렬 데이터 구축 작업은 순차적으로 실행될 필요가 없고, 또한 모든 작업이 아닌 일부만이 실행될 수 있다.

[0062] 먼저, 지도 학습 데이터 레이블링 작업은 다음과 같다.

[0063] 단어와 문자의 편집 거리를 이용하여 교정 문장에서의 교정 형태(삽입, 치환, 삭제)에 대한 정보를 부가 정보로서 추가한다.

[0064] 또한, 오류 범주 정보를 추가한다. 여기서, 오류 범주 정보로는 맞춤법 오류(누락, 첨가, 오선택, 순서 등의 오류), 문법 오류(품사, 일치 등의 오류), 언어 모델 오류(문장 구성, 대용어 참조, 관용 표현, 의미 표현, 모드 표현 등의 오류)가 해당된다.

[0065] 교류 오류의 범주 정보로는 다음의 [표 1]이 참조될 수 있다.

표 1

오류 대분류		오류 소분류	태그	설명
맞춤법 오류 (character 레벨)		누락(Omission)	SO	문장 띄어쓰기 부호, 철자 등이 누락된 경우 예: Discu(s)sion, h(e)ight, definit(e)
		첨가(Addition)	SA	불필요한 문장 부호, 철자 등이 추가된 경우 예: den(n)otation, forei(l)gn, ju(i)dgment
		오선택(False Choice)	SF	대소문자 오류 등 잘못된 글자의 사용 예) absen(s)e, abs(a)nce, indisp(a)nsable
		순서(Misplacement)	SM	잘못된 글자 순서를 포함하는 경우 예) ac(nk)owledge, acq(au)instance, l(ie)sure
문법 오류 (word/phrase/clause 레벨)	품사	전치사(Preposition)	GPre	장소, 원인, 목적, 방법, 출처, 시간 등의 개념으로 잘못 사용된 전치사의 용법
		접속사(Conjunction)	GCon	종속, 상관, 등위 접속사, 접속부사 등의 오류
		명사(Noun)	GN	셀 수 없는/있는 명사 등의 명사 사용 오류
		관사(Article)	GA	정관사/부정관사의 명사와의 부적절한 조합
		대명사(Pronoun)	GPro	부정대명사, 인칭대명사, 재귀대명사 등의 오류
		부사(Adverb)	GAdv	부사의 위치, 비교 구문 등에서 용법 오류
		형용사(Adjective)	GAdj	형용사의 위치, 비교 구문 등에서 용법 오류
		동사(Verb)	GV	동사의 시제, 조동사 등 동사의 용법 오류
	일치	격(Case)	GCas	소유격, 목적격 명사 등의 용법 오류
		성(Gender)	GG	여성, 남성 명사 등의 부적합한 사용
		수(Number)	GN	단수/복수 동사와 사용되는 복수/단수 명사 오류 등
		인칭(Personal)	GPer	남성/여성 명사의 부적합한 일치
	기타	관계사(Relative)	GR	관계 대명사 등의 사용 오류
		수사(Number)	GN	날짜/시간 표기, 통화, 수량, 서수, 기수 등
언어 모델 (sentence/paragraph 레벨)		문장 구성(Sentence form)	LSfo	문법 어순 오류, 수동/능동태, 주문과 종속문 등 문장 구조에서 나타나는 문법 오류
		대용어 참조 (Anaphoric Reference)	LA	문장 간의 대명사 등 부적절한 대용어 용법
		관용 표현(Idiom)	LI	Idiom 등으로 잘못 표현된 문장
		의미 표현(Semantic)	LSem	의미역 인식을 통해 행위자, 경험자, 대상격, 도구격 등 부적합한 의미 표현
		모드 표현(Mode)	LM	존칭/경어, 구어/문어체 등 화용적인 모드에 부적합한 표현

[0066]

[0067] 또한, 비문 및 정문 분류 정보를 바이너리 형태로 추가한다. 이러한 비문 및 정문 분류 정보를 통해 학습 데이터, 즉 비문 데이터와 정문 데이터의 쌍이 모두 교정이 필요없는 정문으로 분류되는 경우를 파악할 수 있다. 이것은 비문 데이터에 대해 교정이 필요없는 것으로 분류될 수 있으므로, 추후 이러한 학습 데이터의 사용을 통해 데이터 확장이 가능하고, 또한 추후 언어 교정시 교정이 필요 없음을 빨리 체크하여 응답할 수 있도록 할 수 있다. 여기서, 비문 데이터에 대해 바이너리 분류기를 통해 교정이 필요없는 정문과 교정이 필요한 비문의 분류를 수행하면서 비문 데이터가 비문과 정문에 해당될 확률값을 표시할 수 있다.

[0068] 또한, 전처리부(121)에서 수행된 코드 스위칭 부분에 대한 정보를 레이블링한다. 예를 들어, 한-영 코드 스위칭 부분의 레이블링을 수행한다.

[0069] 또한, 다양한 자연어 처리를 수행한 후 태그 정보를 추가한다. 여기서, 다양한 자연어 처리로는 문장 분리, 토큰 분리, 형태소 분석, 구문 분석, 개체명 인식, 의미역 인식, 상호참조, 패러프레이즈(paraphrase) 등이 포함될 수 있다.

[0070] 또한, [표 1]에서 필요한 세부 오류 범주 정보를 추가하여 기계학습할 수 있도록 언어의 자질 정보가 사용될 수 있다.

[0071] 다음, 기계학습 데이터 확장 작업은 다음과 같다. 여기서, 기계학습 데이터 확장 작업은 추후 교정 학습부(123)에서 학습시 사용될 기계학습 데이터의 양을 늘리기 위한 작업을 나타낸다.

[0072] 비문 데이터에 다양한 형태의 노이즈를 추가하여 기계학습 데이터 확장을 수행할 수 있다. 여기서, 노이즈 종류로는 단어/철자 누락, 치환, 첨가, 띄어쓰기 오류, 외국어 추가 등이 포함될 수 있다.

[0073] 또한, 빈도수가 높은 오타 오류 위주로 데이터 확장을 수행할 수 있다.

[0074] 또한, 키보드 주변 글자 위주의 오타로 데이터 확장을 수행할 수 있다. 즉, 비문 데이터의 특정 문자에 대해 해당 문자를 타이핑하기 위한 키보드의 정위치를 기준으로 주변의 오타 문자로 형성되는 글자의 오타를 사용하여 데이터 확장을 수행할 수 있다. 이러한 키보드 주변 글자 위주의 오타를 통한 데이터 확장으로 인해 작은 키보드 자판을 사용하는 스마트폰 등을 통해 입력되는 문장들에서의 언어 교정이 매우 효율적으로 수행될 수 있다.

[0075] 또한, VAE(Variational autoencoder), GAN(Generative Adversarial Networks) 등의 비지도 학습에서 사용되는 알고리즘을 응용하여 데이터 확장을 수행할 수 있다.

[0076] 다음, 기계학습용 병렬 데이터 구축 작업은 다음과 같다.

[0077] 상기한 바와 같이 확장된 데이터, 즉 대용량의 데이터 쌍을 노이즈가 포함된 교정 문장인 비문 문장과 교정이 불필요한 정문 문장을 쌍으로 하는 병렬 코퍼스로 구축하는 병렬 데이터 구축 작업을 수행한다.

[0078] 또한, 전처리부(121)에서 비문 및 정문 분류 정보의 바이너리 형태 추가에 의해, 교정이 필요없는 비문 데이터를 사용하여 교정이 불필요한 문장 쌍으로 하는 병렬 코퍼스로 구축하는 병렬 데이터 구축 작업을 수행한다. 이와 같이, 교정이 불필요한 문장 쌍으로 하는 병렬 코퍼스를 사용한 병렬 데이터 구축으로 인해 추후 언어 교정부(140)에서 교정 대상 데이터에 대해 교정이 필요없는 경우 이러한 교정 대상 데이터에 대해 교정을 위한 작업이 수행되지 않도록 처리될 수 있으므로, 전체적인 교정 작업의 처리가 빨리질 수 있다. 물론 이러한 교정이 필요없는 교정 대상 데이터에 대해서도 문장을 자연스럽게 하기 위한 언어 모델링은 수행될 수 있다.

[0079] 교정 학습부(123)는 학습 가공부(122)에 의해 가공된 데이터 쌍, 즉 비문 데이터와 정문 데이터 기반으로 구축된 병렬 데이터의 조합으로 전술한 바와 같은 지도 학습 기반의 기계학습을 적용하여 대응되는 교정 모델을 생성한다. 물론 본 발명은 지도 학습으로만 한정되는 것은 아니고, 비지도 학습 기반의 기계학습을 통해서도 교정 학습이 수행될 수도 있다. 이 경우, 이전의 전처리나 데이터 가공 등이 비지도 학습 기반의 기계학습에 적용될 수 있도록 하는 처리가 수반되어야 할 것이다. 여기서, 교정 학습부(123)는 지도 학습 기반의 기계학습시 기계학습 결과에 대한 오류 출현 확률값을 제공할 수 있다. 이 때, 오류 출현 확률값은 비문과 정문과의 어텐션(attention) 가중치 정보일 수 있다.

[0080] 선택적으로, 교정 학습부(123)는 대용량 인터넷 데이터 기반으로 사전 학습된 임베딩 벡터를 활용할 수 있다. 즉, 외부에서 방대하게 사전 학습된 데이터를 활용할 수 있다.

[0081] 후처리부(124)는 학습 가공부(122)에서의 지도학습 데이터 레이블링 작업시 추가된 태그 부가 정보를 통해 오류 및 오류 카테고리 정보를 출력한 후, 해당 태그 부가 정보를 제거한다.

[0082] 교정 모델 출력부(125)는 교정 학습부(123)에 의해 생성되는 교정 모델을 교정 모델 저장부(130)로 출력하여 저장시킨다.

[0083] 다음, 전술한 언어 교정부(140)에 대해 보다 구체적으로 설명한다.

[0084] 도 3은 도 1에 도시된 언어 교정부(140)의 구체적인 구성도이다.

[0085] 도 3에 도시된 바와 같이, 언어 교정부(140)는 전처리부(141), 오류 문장 검출부(142), 맞춤법 교정부(143), 문법 교정부(144), 언어 모델링부(145) 및 후처리부(146)를 포함한다.

[0086] 전처리부(141)는 입력부(110)를 통해 입력되는 언어 교정을 위한 교정 대상 데이터에 대해 문장 분리 작업을 수행한다. 이러한 문장 분리 작업은 교정 대상 데이터 내에 포함된 문장들의 끝 단위를 인식한 후 입력 단위를 문장 단위로 분리하는 작업이다.

[0087] 또한, 전처리부(141)는 분리된 문장을 다양하게 토큰화한다. 여기서, 토큰화는 문장을 원하는 단위로 자르는 것을 의미하며, 예를 들어, 글자 단위, 단어, 보조단어(subword), 형태소, 어절 등의 단위로 토큰화가 수행될 수 있다.

[0088] 또한, 전처리부(141)는 교정 모델 학습부(120)의 전처리부(121)에서 수행된 바와 같은 데이터 정규화 작업을 수행할 수 있다.

[0089] 다음, 오류 문장 검출부(142)는 바이너리 분류기를 사용하여 전처리부(141)에서 이미 태깅된 정보를 통해서 오류 문장과 비오류 문장을 구분한다. 이것은 기존 오류/비오류 문장쌍의 학습 데이터 외에 오류 문장 위치에 비

오류 문장을 추가해서 확장된 데이터를 기반으로 입력 문장과 기계학습된 오류 문장 또는 비오류 문장의 유사도 측정을 통해 분류하는 방식이다. 이때 오류 문장과 비오류 문장 구분시에 대응되는 신뢰도 수치를 표시한다.

[0090] 오류 문장 검출부(142)는 신뢰도 수치가 임계값 이상이면 오류 문장으로 검출하고, 만약 신뢰도 수치가 임계값보다 작으면 비오류 문장으로 검출한다.

[0091] 이러한 오류 문장 검출부(142)에서의 오류 문장 검출 결과에 따라, 오류 문장으로 검출되면 맞춤법 교정부(143)로 교정 대상 데이터가 전달되지만, 비오류 문장으로 검출되면 맞춤법 교정부(143)와 문법 교정부(144)를 거치지 않고 바로 언어 모델링부(145)로 전달된다.

[0092] 맞춤법 교정부(143)는 오류 문장 검출부(142)에서 전달되는 교정 대상 데이터 내의 교정 대상 문장에서 철자 오류를 검출하고 이를 교정한다. 여기서의 철자 교정은 띄어쓰기, 문장 부호(마침표, 물음표, 느낌표, 쉼표, 가운뎃점, 쌍점, 빗금, 큰 따옴표, 작은 따옴표, 소괄호, 중괄호, 대괄호, 겹낫표와 겹화살괄호, 홑낫표와 홑화살괄호, 줄표, 붙임표, 물결표, 드러냄표와 밑줄, 숨김표, 빠짐표, 줄임표) 등의 철자 오류에 대한 교정이 해당될 수 있다. 또한, 이러한 철자 교정에 대해서도 철자 교정을 위한 기계학습을 수행하여 대응되는 교정 모델을 생성하고, 이렇게 생성된 교정 모델을 사용하여 철자 교정을 수행할 수 있으나, 상기한 바와 같이, 철자 교정은 기계학습을 적용할 정도의 대상은 아니므로 기존의 철자 기준 표준어 사전 등을 사용하여 수행될 수 있다.

[0093] 선택적으로 맞춤법 교정부(143)는 교정 대상 데이터에 대한 철자 교정에 대해 사전 기반 철자 오류 출현 확률값을 신뢰도 정보로서 제공할 수 있다.

[0094] 문법 교정부(144)는 교정 모델 저장부(130)에 저장된 교정 모델을 사용하여 맞춤법 교정부(143)에서 철자 교정된 교정 대상 데이터에 대한 언어 교정, 특히 문법 교정을 수행한다. 즉, 문법 교정부(144)는 교정 대상 데이터에 대해 교정 모델을 적용함으로써 교정 대상 데이터에 대해 교정된 데이터를 결과물로서 획득할 수 있다. 이 때, 교정 모델에 의해 교정된 데이터와 함께 어텐션 가중치를 통한 확률값, 즉 신뢰도 정보가 제공될 수 있다.

[0095] 언어 모델링부(145)는 문법 교정부(144)에 의해 교정된 데이터, 또는 오류 문장 검출부(142)에서 전달되는 비오류 문장에 대해 교정이 불필요한 경우라도 문장을 문법 또한 의미/화용적인 범위에서 더 자연스러운 문장으로 교정하는 언어 모델링을 수행한다. 이러한 언어 모델링도 교정 모델과 같이 기계학습을 이용한 방법을 사용할 수도 있지만 본 발명에서는 적용하지 않고 다양한 형태의 추천 문장 등을 사용하여 대응되는 문장에 대해 언어 모델링을 수행하는 것으로만 설명한다.

[0096] 선택적으로, 언어 모델링부(145)는 언어 모델링을 수행하면서 언어 모델의 퍼플릭시티(perplexity(PPL)) 및 상호 정보(Mutual Information, MI) 값의 조합으로 교정 문장의 신뢰도 정보를 제공할 수 있다.

[0097] 후처리부(146)는 언어 모델링부(145)에 의해 언어 모델링이 수행된 교정된 데이터에 대해 교정 부분을 표시한다. 이러한 교정 부분의 표시는 다양한 색상으로 오류 정보의 시각화를 통해 수행될 수 있다.

[0098] 선택적으로, 후처리부(146)는 오류 문장 검출부(142)에서 바이너리 분류기를 사용하여 오류 문장과 비오류 문장으로 구분시 제공되는 확률값인 신뢰도 수치, 맞춤법 교정부(143)에서의 철자 교정시 제공되는 사전 기반 철자 오류 출현 확률값인 신뢰도 정보, 문법 교정부(144)에서 언어 교정시 제공되는 어텐션 가중치 정보, 언어 모델링부(145)에서 제공되는 언어 모델의 퍼플릭시티 값, 상호 정보(Mutual Information, MI) 등 각 구성요소에서 산출되는 신뢰도의 가중합을 휴리스틱(heuristic) 기반으로 조합하여 교정 대상 데이터에 대한 언어 교정의 최종 신뢰도 정보로서 제공할 수 있다.

[0099] 선택적으로, 후처리부(146)는 하나의 교정 대상 데이터에 대해 N-best 문장 처리를 수행할 수 있다. 즉, 하나의 교정 대상 데이터에 대해 복수의 교정된 데이터 후보군을 제공하면서 각 후보군별 신뢰도를 랭킹(ranking)으로서 제공하여 사용자에 의해 선택될 수 있도록 할 수 있다. 이러한 처리는 출력부(150)와 협력하여 수행될 수 있다.

[0100] 다음, 출력부(150)는 언어 교정부(140)에서 언어 교정이 완료된 교정 데이터와 함께 교정 대상 데이터를 전달받아서 외부로 출력한다. 이 때, 출력부(150)는 교정 대상 데이터와 이에 대응되는 교정된 데이터 및 교정 부분을 함께 표시할 수 있다. 예를 들어, 도 4에 도시된 바와 같이, 좌측에 교정 대상 데이터(Source), 중간에 교정된 데이터(Suggestion), 우측에 교정 부분을 함께 표시함으로써 교정 대상 데이터에 대해 교정된 데이터, 그리고 교정된 부분을 명확히 알 수 있도록 할 수 있다.

[0101] 이하, 본 발명의 실시예에 따른 기계학습 기반 언어 교정 방법에 대해 설명한다.

[0102] 도 5는 본 발명의 실시예에 따른 기계학습 기반 언어 교정 방법의 개략적인 흐름도이다. 도 5에 도시된 기계학습 기반 언어 교정 방법은 도 1 내지 도 4를 참조하여 설명한 언어 교정 시스템(100)에 의해 수행될 수 있다.

[0103] 도 5를 참조하면, 먼저, 언어 교정을 위한 교정 대상 문장이 입력되면(S100), 입력된 교정 대상 문장에 대해 문장 분리 작업, 문장의 토큰화 및 정규화 작업 등을 포함하는 전처리 작업을 수행한다(S110). 여기서, 입력된 교정 대상 문장에 대해 문장 분리 작업, 문장의 토큰화 및 정규화 작업 등을 포함하는 전처리 작업에 대해서는 도 3을 참조한다.

[0104] 다음, 전처리 작업이 수행된 교정 대상 문장에 대해 바이너리 분류기를 사용하여 오류 문장을 검출한다(S120). 도 3을 참조하여 설명한 바와 같이, 이 때 오류 문장 검출에 대한 신뢰도가 함께 제공될 수 있다.

[0105] 따라서, 상기 단계(S120)에서 제공되는 신뢰도가 미리 설정된 임계값 이상이면 오류가 검출된 것으로 언어 교정이 필요하고, 그렇지 않으면 오류가 검출되지 않은 비오류 문장으로 언어 교정이 필요하지 않음을 알 수 있다.

[0106] 따라서, 신뢰도가 미리 설정된 임계값 이상인지가 판단되고(S130), 만약 신뢰도가 미리 설정된 임계값 이상이면, 언어 교정을 위해 먼저 교정 대상 문장에 대한 철자 교정, 즉 맞춤법 교정이 수행된다(S140). 이러한 맞춤법 교정에 대해서도 구체적인 내용에 대해서는 도 3을 참조하여 설명한 부분을 참조한다.

[0107] 그 후 맞춤법 교정된 교정 대상 문장에 대해 지도 학습 기반의 기계학습을 통해 미리 생성되어 있는 생성 모델을 사용하여 언어 교정, 구체적으로는 문법 교정을 수행하여 교정 대상 문장에 대응되는 교정된 문장을 출력한다(S150). 이 때, 생성 모델은 교정 대상 문장에서 교정된 문장으로 교정된 부분에 대한 정보를 함께 제공한다. 또한, 교정 대상 문장의 교정에 대한 신뢰도 정보로서 어텐션 가중치가 함께 제공될 수 있다.

[0108] 계속해서, 교정된 문장에 대해 문법 또한 의미/화용적인 범위에서 더 자연스러운 문장으로 교정하는 언어 모델링이 수행된다(S160). 이러한 언어 모델링에 대해서도 도 3을 참조하여 설명한 부분을 참조한다.

[0109] 이와 같이, 언어 모델링된 문장에 대해 상기한 바와 같은 언어 교정에 대한 신뢰도 정보 제공, N-best 문장 처리 등의 후처리 작업을 수행한다(S170). 이러한 후처리 작업에 대해서는 구체적인 내용은 도 3을 참조하여 설명한 부분을 참조한다.

[0110] 그 후, 후처리 작업이 완료된 최종 교정 문장을 교정 대상 문장과 함께 출력하면서, 교정 부분을 함께 표시함으로써 사용자에게 교정 대상 문장에 대해 본 발명의 실시예에 따른 언어 교정된 교정 문장을 제공할 수 있다(S180).

[0111] 한편, 상기 단계(S130)에서, 신뢰도가 미리 설정된 임계값보다 작아서 언어 교정이 필요 없는 문장인 것으로 판단되는 경우에는 상기한 맞춤법 교정 단계(S140) 및 문법 교정 단계(S150)의 수행 없이 바로 언어 모델링 처리 단계(S160)를 수행하도록 한다.

[0112] 이하, 상기에서 사용되는 교정 모델을 생성하기 위해 기계학습을 수행하는 방법에 대해 설명한다.

[0113] 도 6은 본 발명의 실시예에 따른 언어 교정 모델 학습 방법의 개략적인 흐름도이다. 도 6에 도시된 언어 교정 모델 학습 방법은 도 1 내지 도 3을 참조하여 설명한 언어 교정 시스템(100)에 의해 수행될 수 있다.

[0114] 도 6을 참조하면, 먼저, 언어 교정 모델에 대한 지도 학습 기반의 기계학습을 위해 교정 학습 대상 데이터, 즉 비문 데이터와 정문 데이터의 쌍으로 구성된 대량의 학습 데이터가 입력되면(S200), 언어 감지 작업, 데이터 정제 작업, 정규화 작업 등의 전처리 작업을 수행한다(S210). 구체적인 전처리 작업에 대해서는 도 2를 참조하여 설명한 부분을 참조한다.

[0115] 다음, 전처리 작업이 완료된 교정 학습 대상 데이터에 대해 기계학습에 필요한 데이터로 기계학습 가공 작업을 수행한다(S220). 이러한 기계학습 가공 작업은 지도 학습 데이터 레이블링 작업, 기계학습 데이터 확장(data augmentation) 작업, 기계학습용 병렬 데이터 구축 작업 등을 포함하며, 구체적인 작업 내용에 대해서는 도 2를 참조하여 설명한 부분을 참조한다.

[0116] 그 후, 기계학습 가공 작업이 완료된 교정 학습 대상 데이터를 사용하여 지도 학습 기반의 기계학습을 수행한 후 대응되는 교정 모델을 생성한다(S230). 이 때, 교정 모델과 함께 기계학습 결과에 대한 오류 출현 확률값이 제공될 수 있다.

[0117] 다음, 기계학습 가공시 지도학습 데이터 레이블링 작업에 의해 추가된 태그 부가 정보를 통해 오류 및 오류 카

테고리 정보를 출력한 후, 해당 태그 부가 정보를 제거하는 후처리 작업을 수행한다(S240).

[0118] 마지막으로, 상기 단계(S230)에서 생성된 교정 모델을 교정 모델 저장부(130)에 저장하여, 추후 교정 대상의 문장에 대한 언어 교정시 사용될 수 있도록 한다(S250).

[0119] 한편, 상기한 지도 학습 기반의 교정 모델 학습시 전처리부(121)가 언어 감지 작업, 데이터 정제 작업, 정규화 작업 등의 전처리 작업을 수행하는 것으로만 설명하였으나, 본 발명은 이에 한정되지 않고 보다 정확한 기계학습 기반의 교정 모델 학습이 수행될 수 있도록 하는 다양한 형태의 전처리 작업이 추가로 수행될 수 있다.

[0120] 예를 들어, 교정 모델 학습에서 사용되는 비문 문장인 소스 문장의 오류(오탈자)를 교정 모델 학습 전에 사전에 일괄적으로 교정해서 실질적인 교정 모델 학습시 보다 정확한 소스 문장이 사용될 수 있도록 할 수 있다. 특히, 소스 문장에서 사전에 등록되어 있지 않아 식별이 불가능한 단어 등의 선(先)교정이 수행될 수 있도록 할 수 있다.

[0121] 도 7은 본 발명의 다른 실시예에 따른 교정 모델 학습부(220)의 구체적인 구성도이다.

[0122] 도 7에 도시된 바와 같이, 본 발명의 다른 실시예에 따른 교정 모델 학습부(220)는 전처리부(221), 학습 가공부(222), 교정 학습부(223), 후처리부(224), 교정 모델 출력부(225) 및 번역 엔진(226)을 포함한다. 여기서, 학습 가공부(222), 교정 학습부(223), 후처리부(224) 및 교정 모델 출력부(225)는 도 2를 참조하여 설명한 교정 모델 학습부(120)의 학습 가공부(122), 교정 학습부(123), 후처리부(124) 및 교정 모델 출력부(125)와 구성 및 기능이 동일하므로 도 2를 참조하여 설명한 부분을 참조한다.

[0123] 도 7에서, 번역 엔진(226)은 입력 문장에 대해 사용자에 의해 지정되는 언어로 번역을 수행하는 엔진으로, 예를 들어, RBMT(Rule Based Machine Translation) 엔진일 수 있으며, 본 발명에서는 이것으로만 한정되지는 않는다. 여기서, RBMT(Rule Based Machine Translation)는 수많은 언어 규칙과 언어 사전을 기반으로 번역하는 방식이다. 쉽게 설명해, RBMT는 언어학자가 영어 단어와 문법이 집대성된 교과서를 모두 입력한 번역기를 의미할 수 있다.

[0124] 전처리부(221)는 입력부(110)를 통해 입력되는 언어 교정의 학습을 위해 사용되는 대용량 데이터 내의 비문 데이터인 다량의 소스 데이터에 대해 번역 엔진(226)을 통해 번역을 수행하고, 번역 수행시 번역 엔진(226)에서 사용되는 사전에 등록된 단어가 아닌 경우 해당 단어에 대해 특정 마커, 예를 들어 “##"를 사용하여 표시한 후, 번역이 완료되면 특정 마커로 표시된 단어들을 추출하여 정확한 단어로 일괄적으로 교정한다. 상기에서, 교정 모델의 학습의 대상이 되는 언어와 번역을 수행하는 언어의 경우, 출발어로서 교정 대상의 언어와 같은 언어를 사용한다. 번역 엔진(226)의 출발어에 대한 전처리 과정에서 인식하는 단어 유닛이 사전 기능 및 토큰 분리 모듈을 통해 미등록어를 표시할 수 있도록 하여 오류률이 높은 미등록어에 대한 교정이 가능하도록 한다.

[0125] 선택적으로, 전처리부(221)는 특정 마커로 표시된 단어들을 추출한 후 빈도수를 파악하여 빈도수대로 정렬하고, 정렬된 단어들에 대해 정확한 단어로 교정한 후 일괄적으로 적용함으로써 다량의 소스 데이터에 대한 번역 엔진 기반의 선교정을 수행할 수 있다.

[0126] 이와 같이, 교정 모델 학습 전에 교정 학습에 사용될 다량의 소스 데이터들에 대해 선교정을 수행하여 보다 정확한 소스 데이터가 실제 교정 모델 학습에서 사용될 수 있도록 함으로써 보다 정확한 교정 모델 학습이 수행될 수 있고, 이로 인해 언어 교정의 효율성이 향상될 수 있다.

[0127] 이하, 본 발명의 다른 실시예에 따른 교정 모델 학습 문장의 선교정 방법에 대해 설명한다.

[0128] 도 8은 본 발명의 다른 실시예에 따른 교정 모델 학습 문장의 선교정 방법의 흐름도이다.

[0129] 도 8을 참조하면, 먼저 입력부(110)를 통해 언어 교정의 학습을 위해 사용되는 대용량 데이터 내의 비문 데이터인 다량의 소스 데이터가 입력되면(S300), 다량의 소스 데이터 내의 다량의 소스 문장에 대해 RBMT 엔진을 사용하여 번역을 수행한다(S310).

[0130] 번역 중에 단어가 사전에 등록된 단어인지가 판단되고(S320), 만약 사전에 등록된 단어가 아닌 경우에는 "##"와 같은 마커로 해당 단어의 앞에 미등록 단어로 표시한다(S330).

[0131] 도 9에 도시된 예를 참조하면, 영어 문장의 교정 모델 학습을 위해 “Sorry I don't anderstand."의 소스 문장이 입력되고(1), 이러한 소스 문장에 대해 한국어로의 RBMT 번역이 수행되는 중에, "anderstand"가 사전에 등록되지 않은 단어로 판단되어 이러한 미등록 단어 "anderstand" 앞에 마커 "##"가 표시되는 것을 알 수 있다(2).

[0132] 이와 같이, 다량의 소스 문장에 대해 RBMT 번역이 수행되어 사전에 등록되지 않은 단어들에 대해 마커가 표시되면서 번역이 완료되면(S340), 마커가 표시된 단어를 추출하고(S350), 추출된 단어들의 빈도수를 파악한 후(S360), 파악된 빈도수 기반으로 단어들을 정렬한다(S370). 도 9에 도시된 예를 참조하면, 마커 "##"로 표시된 단어들을 추출하고(3), 추출된 단어들의 빈도수를 파악하여 빈도수 기반으로 정렬한다(4). 예를 들어, 내림차순의 빈도수 기반으로 정렬될 수 있다.

[0133] 그 후, 빈도수 기반으로 정렬된 단어들에 대해 정확한 단어들을 사용하여 다량의 소스 문장에 대해 일괄적으로 교정을 수행(S380)함으로써 교정 모델 학습에 사용될 다량의 소스 문장에서 사전에 등록되지 않은 단어들에 대해 정확한 단어로의 선교정이 이루어질 수 있다.

[0134] 도 9에 도시된 예를 다시 참조하면, 최대 빈도수의 단어 순으로 "studing", "messaged", "Pratice" 등이 정렬되어 있고, 이들 단어에 대해 정확한 단어인 "studying", "sent a message", "practice" 등으로 일괄 교정이 수행될 수 있다(5).

[0135] 한편, 고유 명사와 같이 번역이나 교정시 그 의미가 원문의 의미와 다르게 적용되거나 또는 첫 문자를 대문자로 사용하여야 하는 등에 대해서 미리 정의된 형태의 교정 정보를 변수 형태로 저장하고 이를 런타임에서 처리할 수 있도록 하는 사용자 사전의 사용이 가능하다.

[0136] 이하, 사용자 사전을 만들어 사용자가 필요한 값(단어)를 등록하고, 설정된 값으로 결과물이 도출되도록 하는 내용에 대해 설명한다.

[0137] 도 10은 본 발명의 다른 실시예에 따른 언어 교정 시스템(300)의 개략적인 구성도이다.

[0138] 도 10에 도시된 바와 같이, 본 발명의 다른 실시예에 따른 언어 교정 시스템(300)은 입력부(310), 교정 모델 학습부(320), 교정 모델 저장부(330), 언어 교정부(340), 출력부(350) 및 사용자 사전(360)을 포함한다. 여기서, 입력부(310), 교정 모델 저장부(330) 및 출력부(350)는 도 1을 참조하여 설명한 입력부(110), 교정 모델 저장부(130) 및 출력부(150)와 동일하므로 이에 대한 설명은 생략하고, 그 구성이 상이한 교정 모델 학습부(320), 언어 교정부(340) 및 사용자 사전(360)에 대해서만 설명한다.

[0139] 먼저, 사용자 사전(360)은 사용자가 특정 단어에 대해 미리 정의한 값(단어)을 저장하고 있다. 예를 들어, 고유 명사, "labor day" - "Labor DAY", "memorial day" - "Memorial Day", "african amerian history month" - "African Amerian History Month" 등과 같이, 원래의 의미와는 달라 언어 교정시 의도적으로 잘 안될 수 있는 단어(들)에 대해 사용자가 사용자 사전을 생성하여 사용할 수 있다. 이하에서, "단어"는 설명의 편의를 위해 "단어" 또는 "단어들"을 의미하는 것으로 가정한다.

[0140] 따라서, 본 발명의 다른 실시예에서는 일부 단어에 대해 사용자에 의해 미리 사용자 사전(360)이 생성되어 있는 것으로 가정하여 설명한다.

[0141] 교정 모델 학습부(320)는 입력부(310)를 통해 입력되는 데이터 중에서 언어 교정의 학습을 위해 사용되는 데이터, 즉 비문 데이터와 정문 데이터의 쌍으로 구성된 대량의 학습 데이터를 사용하여 언어 교정을 위한 기계학습을 수행하여 언어 교정용 학습 모델인 교정 모델을 생성한다.

[0142] 특히, 본 발명의 다른 실시예에 따른 교정 모델 학습부(320)는 비문 데이터와 정문 데이터의 쌍으로 구성된 대량의 학습 데이터에서 사용자 사전(360)에 등록되어 있는 단어를 찾아서 사용자 사전 마커, 예를 들어 "UD_NOUN"으로 대체한 후, 기계학습을 수행하여 교정 모델을 생성한다. 여기서, 사용자 사전 마커 "UD_NOUN"은 그 전후에 사용자 사전 마커임을 인식할 수 있도록 다양한 형태의 특수 기호, 예를 들어 "<<", ">>, "_" 등이 더 추가될 수 있다. 이러한 기계학습을 통해 사용자 사전 마커의 위치가 학습되어, 구체적으로는 문맥 정보가 학습될 수 있다. 이 때, 하나의 학습 데이터, 즉 한 문장 내에 포함된 서로 다른 여러 개의 단어가 사용자 사전(360)에 등록되어 있는 경우, 서로 구별이 되는 사용자 사전 마커를 각각 사용하여 대체한 후 사용자 사전 마커의 위치가 다르게 기계학습을 수행할 수 있다. 예를 들면, 한 문장 내에 3개의 서로 다른 단어가 포함되어 있고, 이 단어들이 사용자 사전(360)에 등록되어 있으면, 이 단어들을 각각 "UD_NOUN#1", "UD_NOUN#2", "UD_NOUN#3"를 사용하여 대체할 수 있다.

[0143] 다음, 언어 교정부(340)는 입력부(310)를 통해 입력되는 대용량의 언어 교정 데이터, 즉 맞춤법 오류나 문법 오류의 교정 대상인 교정 대상 데이터에 대해 교정 모델 저장부(330)에 저장된 교정 모델을 사용하여 교정 대상 데이터에 대한 맞춤법/문법 교정을 수행하고, 교정이 완료된 교정 데이터를 출력부(350)로 출력한다.

[0144] 특히, 본 발명의 다른 실시예에 따른 언어 교정부(340)는 교정 대상 데이터내에 사용자 사전에 등록된 단어가

있는 경우, 이들을 사용자 사전 마커로 대체한 후 교정 모델을 사용하여 맞춤법/문법 교정을 수행하고, 그 후의 결과에 포함되어 있는 사용자 사전 마커에 해당하는 단어에 대해 사용자 사전에 등록되어 있는 결과 값(단어)으로 대체함으로써 언어 교정을 완료할 수 있다. 이 때, 하나의 교정 대상 데이터, 즉 한 문장 내에 포함된 서로 다른 여러 개의 단어가 사용자 사전(360)에 등록되어 있는 경우, 서로 구별이 되는 사용자 사전 마커를 각각 사용하여 대체한 후 맞춤법/문법 교정을 수행하고, 그 후에 서로 다른 사용자 사전 마커에 대응되는 단어들을 사용자 사전(360)에서 찾아서 대체하여 교정을 완료할 수 있다. 예를 들어, 하나의 교정 대상 문장 내에 3개의 서로 다른 단어가 포함되어 있고, 이 단어들이 사용자 사전(360)에 등록되어 있으면, 이 단어들을 각각 “UD_NOUN#1”, “UD_NOUN#2”, “UD_NOUN#3”로 대체한 후, 교정을 수행하고, 교정이 완료된 후에, “UD_NOUN#1”, “UD_NOUN#2”, “UD_NOUN#3”에 대응되는 단어들에 대해 사용자 사전(360)에 등록되어 있는 단어들로 대체함으로써 교정이 완료될 수 있다.

[0145] 상기한 바와 같은 본 발명의 다른 실시예에 따른 교정 모델 학습부(320) 및 언어 교정부(340)에 대해 구체적으로 설명한다.

[0146] 도 11은 도 10에 도시된 교정 모델 학습부(320)의 구체적인 구성도이다.

[0147] 도 11에 도시된 바와 같이, 교정 모델 학습부(320)는 전처리부(321), 학습 가공부(322), 교정 학습부(323), 후처리부(324) 및 교정 모델 출력부(325)를 포함한다. 여기서, 학습 가공부(322), 교정 학습부(323), 후처리부(324) 및 교정 모델 출력부(325)는 도 2를 참조하여 설명한 학습 가공부(122), 교정 학습부(123), 후처리부(124) 및 교정 모델 출력부(125)와 동일하므로 여기에서는 구체적인 설명을 생략하고, 구성이 상이한 전처리부(321)에 대해서만 설명한다.

[0148] 전처리부(321)는 도 2를 참조하여 설명한 전처리부(121)의 기능을 수행함은 물론, 이에 더하여, 입력부(110)를 통해 언어 교정의 학습을 위해 사용되는 데이터, 즉 비문 데이터(소스 문장을 뜻함)와 정문 데이터(타겟 문장을 뜻함)의 쌍으로 구성된 학습 데이터가 입력되면, 사용자 사전(360)에 등록되어 있는 단어가 학습 데이터에 포함되어 있는지를 확인하고, 만약 포함되어 있으면 포함된 단어를 사용자 사전 마커, 예를 들어 “<<UD_NOUN>>”으로 대체한다.

[0149] 따라서, 전처리부(321) 이후의 학습 가공부(322), 교정 학습부(323), 후처리부(324) 및 교정 모델 출력부(325)를 통해 기계학습이 수행되어 “<<UD_NOUN>>”으로 대체되어 사용자 사전 마커의 위치가 학습될 수 있다.

[0150] 도 12는 도 10에 도시된 언어 교정부(340)의 구체적인 구성도이다.

[0151] 도 12에 도시된 바와 같이, 언어 교정부(340)는 전처리부(341), 오류 문장 검출부(342), 맞춤법 교정부(343), 문법 수행부(344), 언어 모델링부(345) 및 후처리부(346)를 포함한다. 여기서, 오류 문장 검출부(342), 맞춤법 교정부(343), 문법 교정부(344) 및 언어 모델링부(346)는 도 3을 참조하여 설명한 오류 문장 검출부(142), 맞춤법 교정부(143), 문법 교정부(144) 및 언어 모델링부(145)와 동일하므로 여기에서는 구체적인 설명을 생략하고, 구성이 상이한 전처리부(341) 및 후처리부(346)에 대해서만 설명한다.

[0152] 전처리부(341)는 입력부(310)를 통해 입력되는 언어 교정을 위한 교정 대상 데이터 내에 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있는지를 확인하고, 만약 포함되어 있으면 포함된 단어를 사용자 사전 마커, 예를 들어 “<<UD_NOUN>>”으로 대체한다.

[0153] 후처리부(346)는 언어 모델링부(345)에 의해 언어 모델링이 수행된 교정된 데이터 내에 사용자 사전 마커, 예를 들어 “<<UD_NOUN>>”가 포함되어 있으면, 사용자 사전 마커에 대응되는 소스 문장, 즉 비문 데이터에서의 단어에 대해 사용자 사전(360)에 등록되어 있는 값(단어)으로 대체한다.

[0154] 이와 같이, 사용자 사전(360)에 미리 등록되어 있는 단어들에 대해서는 전처리부(341)에서 미리 사용자 사전 마커로 대체되어 있으므로, 사용자 사전 마커에 대한 문맥 정보가 학습되어 있는 교정 모델을 사용하여 언어 교정시, 즉 맞춤법 교정 및 문법 교정시 사용자 사전 마커에 대해서는 어떠한 교정 없이 후처리부(346)로 입력될 수 있으므로, 후처리부(346)에서 사용자 사전(360)을 사용하여 해당 단어들에 대한 대체가 가능해진다.

[0155] 따라서, 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있는 소스 문장에 대해 사용자 사전(360)을 기반으로 하는 교정이 성공적으로 수행될 수 있다.

[0156] 이하, 도면을 참조하여 본 발명의 다른 실시예에 따른 언어 교정 모델 학습 방법에 대해 설명한다. 이러한 언어 교정 모델 학습 방법은 전술한 도 10 내지 도 12를 참조하여 설명한 언어 교정 시스템(300)에 의해 수행될

수 있다.

[0157] 도 13은 본 발명의 다른 실시예에 따른 언어 교정 모델 학습 방법의 흐름도이다. 여기서, 도 13에 도시된 본 발명의 다른 실시예에 따른 언어 교정 모델 학습 방법은 도 10 내지 도 12를 참조하여 설명한 본 발명의 다른 실시예에 따른 언어 교정 시스템(300)에 의해 수행될 수 있다.

[0158] 설명 전에, 사용자가 특정 단어에 대해 미리 정의한 값(단어)을 저장하고 있는 사용자 사전(360)을 미리 구성하여 놓은 것으로 가정한다.

[0159] 도 13을 참조하면, 먼저, 언어 교정의 학습을 위해 사용되는 데이터, 즉 비문 데이터(소스 문장을 뜻함)와 정문 데이터(타겟 문장을 뜻함)의 쌍으로 구성된 학습 데이터가 입력되면(S400), 소스 문장 및 타겟 문장 내에 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있는지가 판단된다(S410).

[0160] 만약 소스 문장 및 타겟 문장 내에 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있는 것으로 판단되면, 사용자 사전(360)에 등록되어 있는 단어와 일치하는 단어를 사용자 사전 마커로 대체한다(S420). 예를 들어, 사용자 사전(360) 내에 <“memorial day” - “Memorial Day”>가 등록되어 있고, 언어 교정의 학습을 위해 입력되는 소스 문장이 “memorial day is observed on the last Monday”인 경우, 소스 문장 내에 있는 단어 “memorial day”가 사용자 사전(360)에 등록되어 있으므로, 이단어가 사용자 사전 마커, 예를 들어 “<<UD_NOUN>>”로 대체되어, 소스 문장이 “<<UD_NOUN>> is observed on the last Monday”와 같이 변경된다.

[0161] 그러나, 소스 문장 및 타겟 문장 내에 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있지 않으면 소스 문장 및 타겟 문장은 변경되지 않고 입력된 그대로 사용될 수 있다.

[0162] 그 후, 변경되었거나 또는 변경되지 않은 소스 문장과 타겟 문장인 언어 교정용 학습 데이터에 대해 기계학습을 수행하여 교정 모델을 생성한다(S430). 이러한 기계학습을 통해 사용자 사전 마커의 위치가 학습될 수 있다. 또한, 기계학습을 수행하는 구체적인 내용에 대해서는 도 1 내지 도 9를 참조하여 설명한 실시예를 참조한다.

[0163] 다음, 본 발명의 다른 실시예에 따른 언어 교정 방법에 대해 설명한다. 이러한 언어 교정 방법은 전술한 도 10 내지 도 12를 참조하여 설명한 언어 교정 시스템(300)에 의해 수행될 수 있다.

[0164] 도 14는 본 발명의 다른 실시예에 따른 언어 교정 방법의 흐름도이다. 여기서, 도 14에 도시된 본 발명의 다른 실시예에 따른 언어 교정 모델 학습 방법은 도 10 내지 도 12를 참조하여 설명한 본 발명의 다른 실시예에 따른 언어 교정 시스템(300)에 의해 수행될 수 있다.

[0165] 설명 전에, 사용자가 특정 단어에 대해 미리 정의한 값(단어)을 저장하고 있는 사용자 사전(360)을 미리 구성하여 놓은 것으로 가정한다.

[0166] 언어 교정 데이터, 즉 맞춤법 오류나 문법 오류의 교정 대상인 교정 대상 데이터가 입력되면(S500), 교정 대상 데이터 내에 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있는지를 확인한다(S510).

[0167] 만약 교정 대상 데이터 내에 사용자 사전(360)에 등록되어 있는 단어가 포함되어 있는 것으로 확인되면, 해당 단어를 사용자 사전 마커, 예를 들어 “<<UD_NOUN>>”으로 대체한다(S520). 전술한 도 13에서의 예를 참조하면, 사용자 사전(360) 내에 <“memorial day” - “Memorial Day”>가 등록되어 있고, “memorial day is observed on the last Monday”인 교정 대상 문장이 입력되면, 해당 문장 내에 “memorial day”가 사용자 사전(360)에 등록되어 있는 단어이므로 해당 단어가 사용자 사전 마커, 즉 “<<UD_NOUN>>”로 대체되어, 결과적으로 교정 대상 문장은 “<<UD_NOUN>> is observed on the last Monday”이 된다.

[0168] 그 후, 도 10 내지 도 13에서 설명한 바와 같은 언어 교정 학습을 통해 생성된 교정 모델을 사용하여 교정 대상 데이터에 대한 맞춤법/문법 교정을 수행하고(S530), 교정된 결과에 대해 언어 모델링을 수행한다(S540).

[0169] 그 후, 언어 모델링 결과의 문장 내에 사용자 사전 마커, 즉 “<<UD_NOUN>>”가 있는지를 확인하고(S550), 만약 사용자 사전 마커가 있으면 사용자 사전 마커에 대응되는 소스 문장의 단어에 대해 사용자 사전(360)에 등록되어 있는 단어로 대체한다(S560). 상기 예를 참조하면, 언어 모델링 결과 출력되는 문장 “<<UD_NOUN>> is observed on the last Monday” 내에 사용자 사전 마커 “<<UD_NOUN>>”가 포함되어 있으므로, 사용자 사전 마커 “<<UD_NOUN>>”에 해당하는 단어, 즉 “memorial day”에 대해 사용자 사전(360)에 등록되어 있는 단어, 즉 “Memorial Day”가 대체되어, 최종적으로 교정 후의 문장인 “Memorial Day is observed on the last Monday”가 완성된다.

[0170] 그 후, 교정이 완료된 교정 문장을 출력한다(S570).

[0171] 한편, 상기 단계(S550)에서 언어 모델링 결과 출력되는 문장 내에 사용자 사전 마커가 포함되어 있지 않으면 바로 교정 문장을 출력하는 단계(S570)를 수행한다.

[0172] 이와 같이, 본 발명의 실시예에 따르면, 사용자에 의해 미리 정의된 형태의 교정 정보를 변수 형태로 저장하고 이를 런타임에서 처리함으로써, 교정 모델에 별도로 추가하거나 변경하지 않고도 쉽게 언어 교정이 수행될 수 있다.

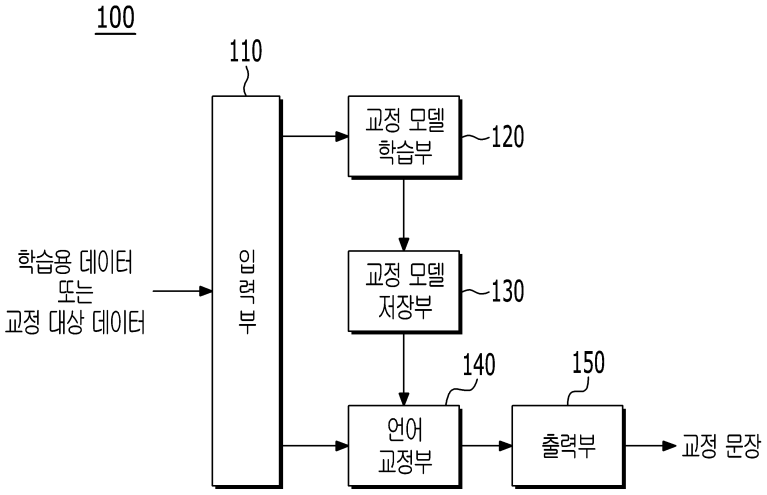
[0173] 따라서, 교정이 어렵거나 의도적으로 잘 안되는 부분에 대해서도 사용자 사전에 등록하여 처리함으로써 언어 교정의 효율을 향상시킬 수 있다.

[0174] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.

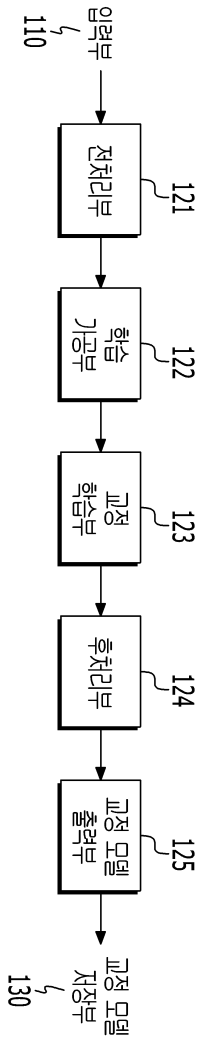
[0175] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

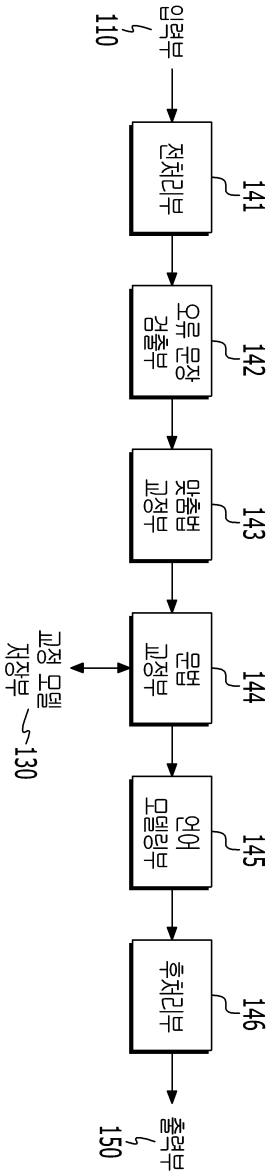
도면1



도면2


입력부
110
전처리부
121
학습 기능부
122
교정 학습부
123
후처리부
124
교정 모델 출력부
125
교정 모델 저장부
130

도면3


입력부
110
전처리부
141
오류 문장 검출부
142
맞춤법 교정부
143
문법 교정부
144
언어 모델링부
145
후처리부
146
출력부
150
교정 모델 저장부
130

도면4

[교정 대상 데이터]

When I will arrive. I will call you.
I've been here since three months.
My boyfriend has got a new work.
She doesn't listen me.
You speak English good.
The police is coming.
The house isn't enough big.
You should not to smoke.
Do you like a glass of wine?
There is seven girls in the class.
I didn't meet nobody.
My flight departs in 5:00 am.
I promise I call you next week.
Where is post office?
Please explain me how improve my English.
We studied during four hours.
Is ready my passport?
You cannot buy all what you like!
She is success.
My mother wanted that I be doctor.
The life is hard!
How many childrens you have?
My brother has 10 years.
I want eat now.
You are very nice, as your mother.
She said me that she liked you.
My husband engineer.
I came Australia to study English.
It is more hot now.
You can give me an information?
They cooked the dinner themself.
Me and Johnny live here.
I closed very quietly the door.
You like dance with me?

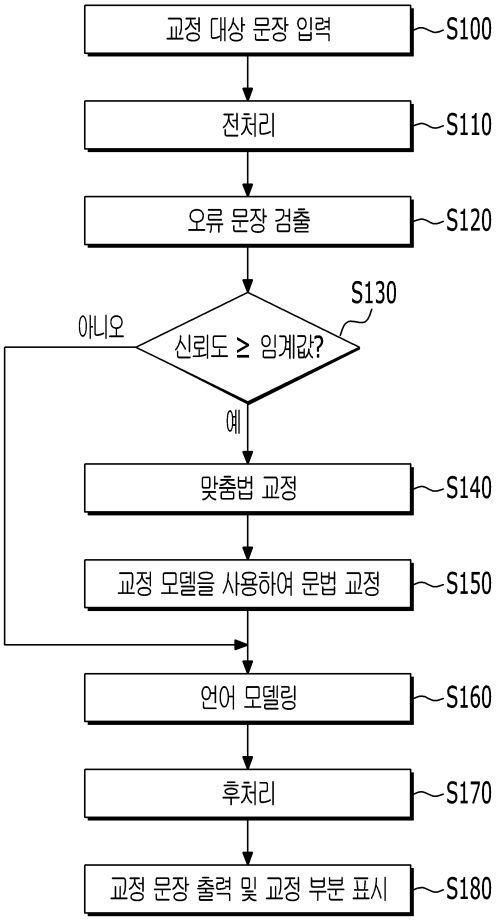
[교정된 데이터]

When I arrive. I will call you.
I've been here for three months.
My boyfriend has got a new job.
She doesn't listen to me.
You speak English well.
The police are coming.
The house isn't big enough.
You should not smoke.
Would you like a glass of wine?
There are seven girls in the class.
I didn't meet anybody.
My flight departs at 5:00 am.
I promise I will call you next week.
Where is the post office?
Please explain to me how to improve my English.
We studied for four hours.
Is my passport ready?
You cannot buy all that you like!
She is successful.
My mother wanted me to be a doctor.
Life is hard!
How many children do you have?
My brother is 10 years old.
I want to eat now.
You are very nice,like your mother.
She told me that she liked you.
My husband is engineer.
I came to Australia to study English.
It is hotter now.
Can you give me some information?
They cooked dinner by themselves.
Me and Johnny live there.
I closed the door very quietly.
You like to dance with me?

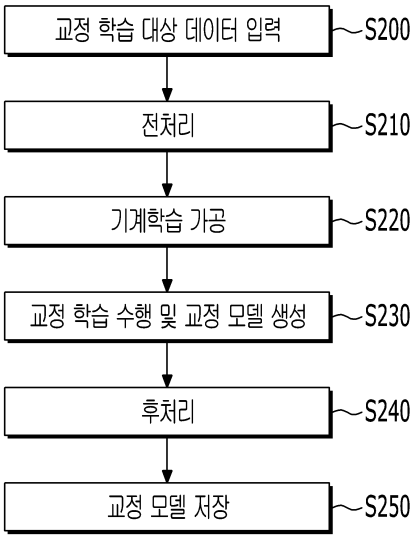
[교정된 부분]

When I ~~will~~ arrive. I will call you.
~~I've~~ I've been here ~~since~~ for three months.
My boyfriend has got a new ~~work~~ job.
She ~~doesn't~~ doesn't listen to me.
You speak English ~~good~~ well.
The police ~~is~~ are coming.
The house ~~isn't enough big~~ isn't big enough.
You should not ~~to~~ smoke.
~~Do~~ Would you like a glass of wine?
There ~~is~~ are seven girls in the class.
I ~~didn't~~ didn't meet ~~nobody~~ anybody.
My flight departs ~~in~~ at 5:00 am.
I promise I will call you next week.
Where is the post office?
Please explain to me how to improve my English.
We studied ~~during~~ for four hours.
Is ~~ready~~ my ~~passport?~~ passport ready?
You cannot buy all ~~what~~ that you like!
She is ~~success~~ successful.
My mother wanted ~~that I~~ me to be a doctor.
~~The life~~ Life is hard!
How many ~~childrens~~ children do you have?
My brother ~~has~~ is 10 years years old.
I want to eat now.
You are very nice, ~~as~~ like your mother.
She ~~said told~~ me that she liked you.
My husband is engineer.
I came to Australia to study English.
It is ~~more hot~~ hotter now.
~~You can~~ Can you give me ~~an~~ some information?
They cooked ~~the~~ dinner ~~themself~~ by themselves.
Me and Johnny live ~~here~~ there.
I closed the door very ~~quietly the door~~ quietly.
You like to dance with me?

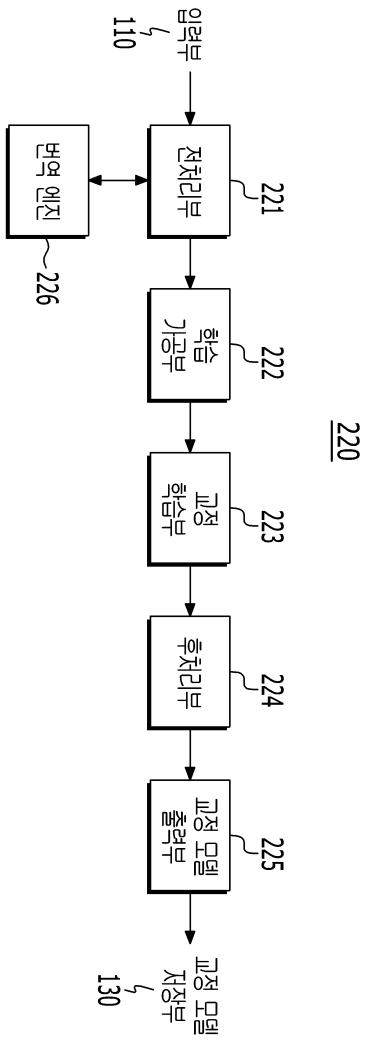
도면5


교정 대상 문장 입력
S100
전처리
S110
오류 문장 검출
S120
S130
신뢰도 ≥ 임계값?
아니오
예
맞춤법 교정
S140
교정 모델을 사용하여 문법 교정
S150
언어 모델링
S160
후처리
S170
교정 문장 출력 및 교정 부분 표시
S180

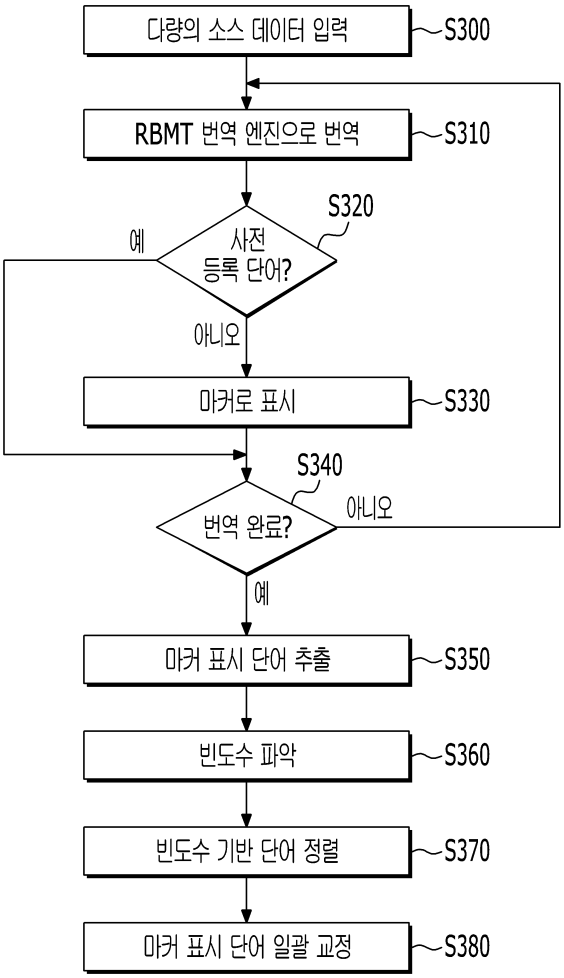
도면6


교정 학습 대상 데이터 입력
S200
전처리
S210
기계학습 가공
S220
교정 학습 수행 및 교정 모델 생성
S230
후처리
S240
교정 모델 저장
S250

도면7


입력부
110
번역 엔진
226
전처리부
221
학습 기공부
222
교정 학습부
223
후처리부
224
교정 모델 출력부
225
220
교정 모델 저장부
130

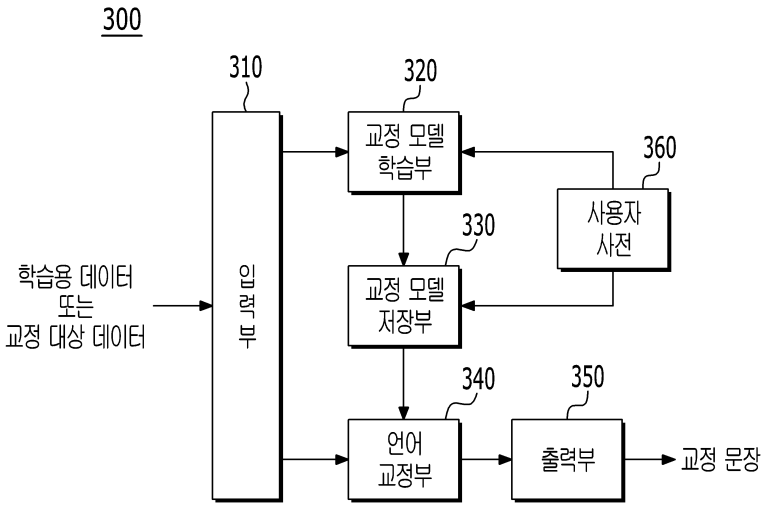
도면8


다량의 소스 데이터 입력
S300
RBMT 번역 엔진으로 번역
S310
S320
사전 등록 단어?
예
아니오
마커로 표시
S330
S340
번역 완료?
아니오
예
마커 표시 단어 추출
S350
빈도수 파악
S360
빈도수 기반 단어 정렬
S370
마커 표시 단어 일괄 교정
S380

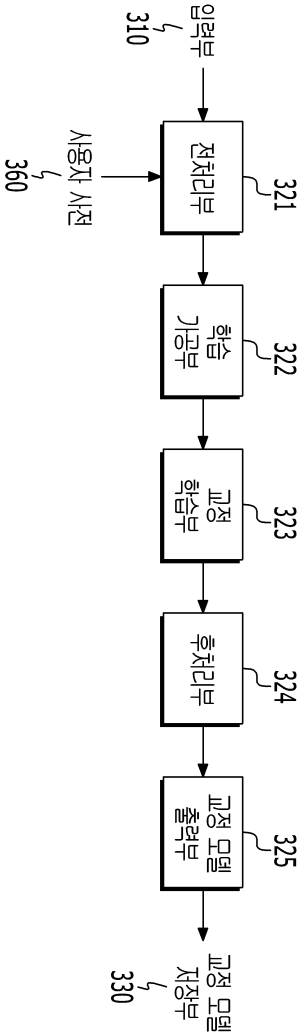
도면9

1. 소스 문장
Sorry I don't anderstand
2. 영한 RMBT 번역엔진으로 번역 (마커(##) 표시)
유감스러운 나는 ##anderstand 당신
3. 마커(##) 단어 추출
grep -P '##[a-Za-z]+' filename > nfw list.en
4. 빈도수 기반 정렬
sed -e 's/[^[:alpha:]]// /g' nfw list.en | tr '₩n' " " | tr -s " " | tr " " '₩n' | tr 'A-Z' 'a-z' | sort | uniq -c | sort -nr | nl
5. 결과물 교정 후 일괄 적용
번호 빈도수 마커(##) 단어 교정 후
1 1419 studing studying
2 607 messaged sent a message
3 395 Pratice practice

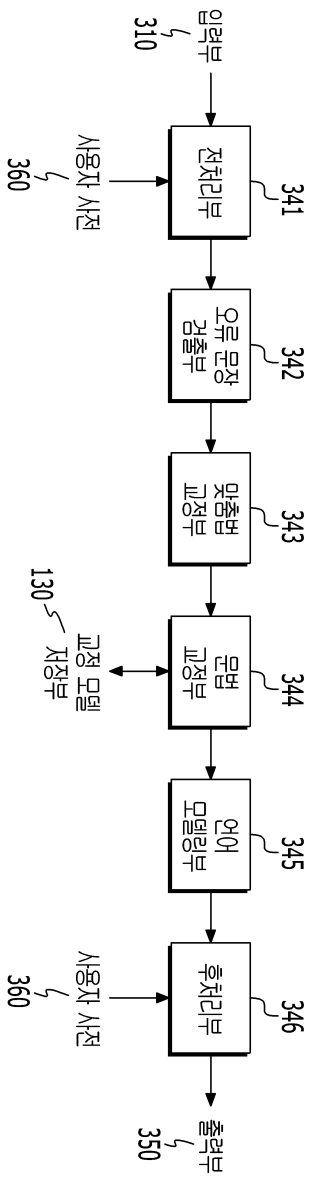
도면10


300
310
320
360
330
340
350
학습용 데이터
및
교정 대상 데이터
입력부
교정 모델
학습부
사용자
사전
교정 모델
저장부
언어
교정부
출력부
교정 문장

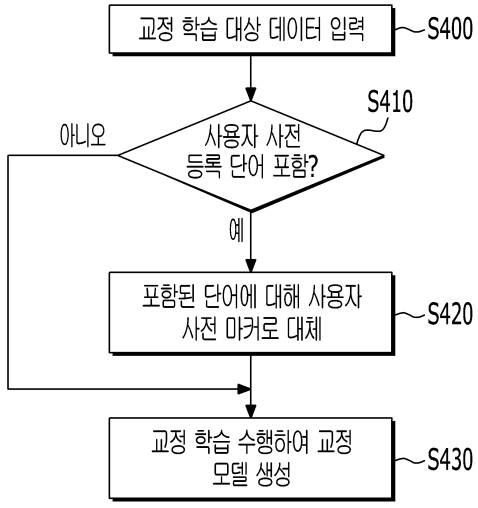
도면11


310
입력부
360
사용자 사전
전처리부
321
학습
기능부
322
교정
학습부
323
후처리부
324
교정 모델
출력부
325
330
교정 모델
저장부

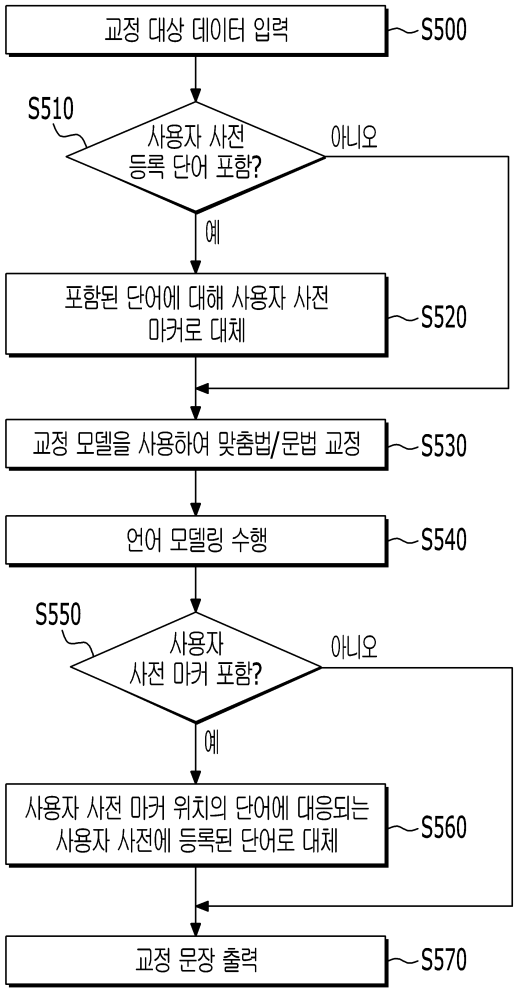
도면12


입력부
310
전처리부
341
사용자 사전
360
오류 문장 검출부
342
맞춤법 교정부
343
교정 모델 저장부
130
문법 교정부
344
언어 모델링부
345
후처리부
346
사용자 사전
360
출력부
350

도면13


교정 학습 대상 데이터 입력
S400
S410
사용자 사전
등록 단어 포함?
아니오
예
포함된 단어에 대해 사용자
사전 마커로 대체
S420
교정 학습 수행하여 교정
모델 생성
S430

도면14


교정 대상 데이터 입력
S500
S510
사용자 사전
등록 단어 포함?
아니오
예
포함된 단어에 대해 사용자 사전
마커로 대체
S520
교정 모델을 사용하여 맞춤법/문법 교정
S530
언어 모델링 수행
S540
S550
사용자
사전 마커 포함?
아니오
예
사용자 사전 마커 위치의 단어에 대응되는
사용자 사전에 등록된 단어로 대체
S560
교정 문장 출력
S570