

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2022년 7월 7일 (07.07.2022)



(10) 국제공개번호
WO 2022/146126 A1

(51) 국제특허분류:

G06F 40/232 (2020.01) G16H 15/00 (2018.01)
G06F 16/215 (2019.01) G06N 20/00 (2019.01)
G06F 16/2458 (2019.01)

(21) 국제출원번호: PCT/KR2022/000061

(22) 국제출원일: 2022년 1월 4일 (04.01.2022)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2021-0000424 2021년 1월 4일 (04.01.2021) KR

(71) 출원인: 고려대학교 산학협력단 (KOREA UNIVERSITY RESEARCH AND BUSINESS FOUNDATION)
[KR/KR]; 02841 서울특별시 성북구 안암로 145, 고려대학교 (안암동5가), Seoul (KR).

(72) 발명자: 주형준 (JOO, Hyung Joon); 06289 서울특별시 강남구 선릉로 120, 10-1405, Seoul (KR). 한성원 (HAN, Sung Won); 06001 서울특별시 강남구 압구정로 151, 102-1105, Seoul (KR). 손장욱 (SOHN, Jang Wook); 06177 서울특별시 강남구 영동대로72길 10, 301-1403, Seoul (KR). 김태형 (KIM, Tae-Hyeong); 13583 경기도 성남시 분당구 분당로263번길 24, 119-501, Gyeonggi-do

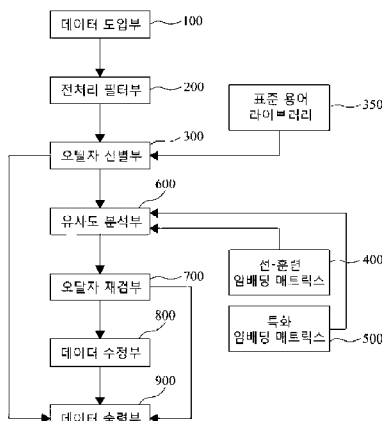
(KR). 김종호 (KIM, Jong-Ho); 02751 서울특별시 성북구 화랑로 76, 102-3504, Seoul (KR). 박수완 (PARK, Soo Wan); 15468 경기도 안산시 단원구 광덕2로 241, 806-1403, Gyeonggi-do (KR). 장준호 (CHANG, Junho); 08720 서울특별시 관악구 봉천로19길 18, 303호, Seoul (KR). 이정문 (LEE, Jeong Moon); 15884 경기도 군포시 송부로49번길 15, 105-2402, Gyeonggi-do (KR). 강민지 (KANG, Minji); 02095 서울특별시 중랑구 동일로 752, 103-2302, Seoul (KR).

(74) 대리인: 특허법인 태백 (TAEBAEK INTELLECTUAL PROPERTY LAW FIRM); 08506 서울특별시 금천구 가산디지털1로151 이노플렉스1차 601호, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: DEVICE AND METHOD FOR CORRECTING MEDICAL TERM TYPOS

(54) 발명의 명칭: 의학용어 오탃자 수정 장치 및 방법



100 ... Data introduction unit
200 ... Preprocessing filter unit
300 ... Typo sorting unit
350 ... Standard terms library
400 ... Pre-training embedding matrix
500 ... Specialized embedding matrix
600 ... Similarity analysis unit
700 ... Typo re-examination unit
800 ... Data correction unit
900 ... Data output unit

(57) Abstract: A device and a method for correcting medical term typos are disclosed. The device comprises: a similarity analysis unit for analyzing the similarity between sorted typos and a correction candidate group generated by a pre-training embedding matrix and a specialized embedding matrix; and a data correction unit for correcting data determined to be typos according to the similarity analysis result. According to the present invention, atypical text data in a legally effective document written in the medical field and the health care industry is refined by using two or more word embedding matrices, and thus the limitations of different words being used by each medical institution can be overcome. In addition, currently generated typos and even typos that can be generated in the future can be corrected consistently.

(57) 요약서: 본 발명은 의학용어 오탃자 수정 장치 및 방법을 공개한다. 이 장치는 선별된 오탃자와 전-훈련 임베딩 매트릭스 및 특화 임베딩 매트릭스에서 생성된 교정 후보군의 유사도를 분석하는 유사도 분석부; 및 상기 유사도 분석 결과에 따라 오탃자로 판단된 데이터를 수정하는 데이터 수정부;를 포함하는 것을 특징으로 한다. 본 발명에 의할 경우, 의학 분야 및 의료업계에서 작성되는 법적 효력이 있는 문서에 있는 비정형 텍스트 데이터를 두 가지 이상의 단어 임베딩 매트릭스를 적용하여 정제함으로써, 의료기관별로 사용되는 단어가 상이한 한계를 극복할 수 있게 된다. 또한, 현재 생성되는 오탃자뿐 아니라 앞으로 발생할 수 있는 오탃자에 대하여도 일관성 있게 수정할 수 있게 된다.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 의학용어 오탃자 수정 장치 및 방법

기술분야

- [1] 본 발명은 오탃자 수정 장치 및 방법에 관한 것으로서, 보다 상세하게는 의료 데이터에 존재하는 오탃자를 자동으로 탐색하여 수정하기 위해 2개의 임베딩 매트릭스를 활용하여 교정 후보군의 유사도를 분석하고 오탃자를 수정할 수 있는 의학용어 오탃자 수정 장치 및 방법에 관한 것이다.

배경기술

- [2] 일반적으로 오탃자 교정 기술이란 사용자가 실수한 검색어에 대해 확률에 따라 원래 의도한 검색어를 제시하여 주는 기술을 의미하는데, 오탃자는 크게 두 가지로 나눌 수 있다.
- [3] 하나는 타이핑(typos) 오탃자로서, 키보드 상의 배치된 키의 위치에 기인하는 오탃자이다. 또 하나는 인식(cognitive) 오탃자로서, 단어의 정확한 철자에 대한 오해에 기인하는 오탃자이다.
- [4] 초기의 오탃자 교정 기술은 사전을 기반으로 하였으며, 사전에서 발견하지 못한 단어를 오탃자로 간주하여 사전 내의 유사한 단어들 중에서 정타 후보를 선택하여 교정해주는 방법이 일반적이다.
- [5] 하지만, 검색 질의어(query)의 오탃자 교정은 기존의 사전에 의존하는 방법으로는 정확한 오탃자 교정 효과를 얻을 수 없다.
- [6] 왜냐하면, 검색 질의어는 매우 다양한 형태를 띄고 있으며 인터넷 신조어 등을 포함하고 있기 때문에, 이들 모두를 수작업으로 DB화하는 것은 매우 어려운 실정이다.
- [7] 한편, 의료 분야의 빅데이터를 구축하기 위해서는 데이터의 정제가 필수적인 성공 요소 중 하나이다.
- [8] 현재 병원 등의 의료기관 내 대다수의 데이터는 코드화되지 않은 비정형 텍스트 혹은 이미지로 구성되어 있다.
- [9] 특히, 병원 등의 의료기관의 데이터(예를 들어, 진료 기록 정보 등)는 많은 경우에 길고 복잡한 학명 및 의학 용어를 포함함으로써, 철자의 오류율이 높고 일반적인 통상의 데이터 정제 과정을 통해 수정하기 어려운 한계가 있었다.
- [10] 또한, 의학 논문 등 의학 관련 문서에도 복잡한 학명 및 의학 용어를 많이 포함하고 있으며, 의학 관련 학문이 발전하면서 새로운 의학 용어도 점점 빠르게 늘고 있는 실정이다.
- [11] 한편, 기존의 오탃자 자동 수정 기술들은 일반 영역에서 오탃자를 수정하는 기술들이며, 앞서 언급한 복잡한 학명 및 의학 용어에 사용하기에는 한계가 많다고 할 수 있다.
- [12] 더구나 기존 기술들은 주로 기 구축된 단어 목록 데이터베이스를 토대로

유사도를 분석하는 방법을 통해 오타자를 수정하는데, 이들은 단어 목록 데이터베이스 자체에 오류가 있는 경우 수정 대상 데이터의 오타자를 정확하게 수정할 수 없을 뿐 아니라, 새롭게 사용되는 신조어는 수정할 수 없다는 한계가 있었다.

- [13] 예를 들어, 국내공개특허 2020-0057824 A1은 단어 임베딩 학습을 수행하여 오타자를 교정하는 단어 교정 시스템에 관한 기술이고, 국내등록특허 10-1638953 B1은 오타자 DB에 저장된 정보와 비교를 통해 오타자를 추출하도록 구성된 온라인상 텍스트 자동 분석 및 수정 시스템에 관한 기술이며, 일본공개특허 JP 2011-65384 A1은 입력 텍스트를 형태소 해석하여 단어 열 데이터를 출력하고 입력 텍스트의 단어 소정 문자 길이의 단어에 대해서 근사 대조하여 근사 사전 대조 단어 열 데이터를 출력하며 단어 열 데이터와 근사 사전 대조 단어 열 데이터를 이용하여 오타자를 수정하는 방법에 관한 기술이 제시되었다.
- [14] 하지만, 상기 종래 기술들은 입력 텍스트의 오자 및 탈자 수정을 위한 기술이라는 공통점이 있으나, 의료 데이터를 대상으로 하는 것이 아니라 일반 문서 작성기, 모바일 메시징 어플리케이션 등과 같은 일반적인 영역에서의 오타자 교정 기술에 관한 것이라는 한계가 있었다.
- [15] 이에 따라, 병원, 약국, 제약 회사 등 실제 의료업계의 모든 의학 분야별로 완벽한 단어 목록 데이터베이스를 구축하는 방법은 현실적으로 불가능하며, 학명 및 의학 용어의 양 자체도 급속도로 늘어나고 풍부해지는 것을 고려할 때, 이러한 현실적 필요를 고려한 새로운 차원의 오타자 수정 기술이 필요한 실정이다.

발명의 상세한 설명

기술적 과제

- [16] 본 발명의 목적은 병원 등의 의료기관 및 의료관련 문서 및 문자형 데이터베이스에 존재하는 오타자를 자연어 처리 기법을 활용한 유사도 분석을 통해 자동으로 탐색하고, 두 가지 이상의 단어 임베딩 매트릭스를 병합하여 단어 목록을 구성함으로써 적절하게 수정할 수 있는 의학용어 오타자 수정 장치를 제공하는 데 있다.
- [17] 본 발명의 다른 목적은 상기 목적을 달성하기 위한 의학용어 오타자 수정 방법을 제공하는 데 있다.
- [18] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있으며, 본 발명의 실시예에 의해 보다 분명하게 알게 될 것이다. 또한, 본 발명의 목적 및 장점들은 특허청구범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제 해결 수단

- [19] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치는 선별된 오탃자와 전-훈련 임베딩 매트릭스 및 특화 임베딩 매트릭스에서 생성된 교정 후보군의 유사도를 분석하는 유사도 분석부; 및 상기 유사도 분석 결과에 따라 오탃자로 판단된 데이터를 수정하는 데이터 수정부;를 포함하는 것을 특징으로 한다.
- [20] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치는 분석할 데이터를 메모리에 저장하는 데이터 도입부; 상기 저장된 데이터를 분석 단위로 분할하고, 전처리하여 필터링하는 전처리 필터부; 및 상기 필터링된 데이터가 오탃자인지 선별하는 오탃자 선별부;를 더 포함하고, 상기 전-훈련 임베딩 매트릭스는 상기 선별된 데이터 중에서 교정 단어 1차 후보군을 생성하고, 상기 특화 임베딩 매트릭스는 의료 기관에서 보관하고 있는 문서를 이용하여 임베딩 매트릭스를 구성하고, 교정 단어 2차 후보군 생성하며, 상기 유사도 분석부는 선별된 오탃자와 상기 생성된 교정 단어 1차 및 2차 후보군의 유사도를 분석하여 순위를 정하고 최종 후보군을 제시하는 것을 특징으로 한다.
- [21] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치는 상기 제시된 최종 후보군 데이터의 오탃자를 재검사하는 오탃자 재검부; 및 상기 오탃자 선별부 및 상기 오탃자 재검부에서 오탃자로 선별되지 않은 데이터를 그대로 출력하거나, 상기 수정된 데이터를 출력하는 상기 데이터 출력부;를 더 포함하고, 상기 데이터 수정부는 상기 재검사 결과 오탃자로 판단된 데이터를 수정하는 것을 특징으로 한다.
- [22] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치의 상기 저장된 데이터는 단말기 또는 데이터베이스에 입력되어 있는 데이터인 것을 특징으로 한다.
- [23] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치의 상기 전-훈련 임베딩 매트릭스는 사전에 훈련된 단어 라이브러리인 것을 특징으로 한다.
- [24] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치의 상기 전-훈련 임베딩 매트릭스는 BioWordVec, Sub-word pre-trained model을 포함하는 것을 특징으로 한다.
- [25] 상기 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 장치의 상기 특화 임베딩 매트릭스는 fastText 기법을 이용하여 임베딩 매트릭스를 구성하는 것을 특징으로 한다.
- [26] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 방법은 유사도 분석부가 선별된 오탃자와 전-훈련 임베딩 매트릭스 및 특화 임베딩 매트릭스에서 생성된 교정 후보군의 유사도를 분석하는 단계; 및 데이터 수정부가 상기 유사도 분석 결과에 따라 오탃자로 판단된 데이터를 수정하는 단계;를 포함하는 것을 특징으로 한다.
- [27] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탃자 수정 방법은 (a)

데이터 도입부가 분석할 데이터를 메모리에 저장하는 단계; (b) 전처리 필터부가 상기 저장된 데이터를 분석 단위로 분할하고, 전처리하여 필터링하는 단계; (c) 오탈자 선별부가 상기 필터링된 데이터가 오탈자인지 선별하는 단계; (d) 상기 전-훈련 임베딩 매트릭스가 상기 선별된 데이터 중에서 교정 단어 1차 후보군을 생성하는 단계; (e) 상기 특화 임베딩 매트릭스가 의료 기관에서 보관하고 있는 문서를 이용하여 임베딩 매트릭스를 구성하고, 교정 단어 2차 후보군을 생성하는 단계; (f) 상기 유사도 분석부가 선별된 오탈자와 상기 생성된 교정 단어 1차 및 2차 후보군의 유사도를 분석하여 순위를 정하고 최종 후보군을 제시하는 단계; (g) 오탈자 재검부가 상기 제시된 최종 후보군 데이터의 오탈자를 재검사하는 단계; 및 (h) 상기 데이터 수정부가 상기 재검사 결과 오탈자로 판단된 데이터를 수정하는 단계; 를 포함하는 것을 특징으로 한다.

- [28] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탈자 수정 방법은 상기 (h) 단계 이후에, 데이터 출력부가 상기 오탈자 선별부 및 상기 오탈자 재검부에서 오탈자로 선별되지 않은 데이터를 그대로 출력하거나, 상기 수정된 데이터를 출력하는 단계; 를 더 포함하는 것을 특징으로 한다.
- [29] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탈자 수정 방법의 상기 전처리는 형식 정리 및 불용어 전처리를 포함하는 것을 특징으로 한다.
- [30] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탈자 수정 방법은 상기 (c) 단계에서, 상기 오탈자인지 선별은 입력된 데이터의 단어가 표준 용어 라이브러리에 존재하는지 여부를 소정의 규칙 기반으로 판단하는 것을 특징으로 한다.
- [31] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탈자 수정 방법의 상기 표준 용어 라이브러리는 SNOMED Clinical Terms 임상의학 문서 및 보고서에 사용되는 의학 용어 집합을 사용하는 것을 특징으로 한다.
- [32] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탈자 수정 방법의 상기 소정의 규칙은 오탈자로 분류되는 경우 후속의 알고리즘을 따르고, 오탈자로 분류되지 않는 경우 상기 저장된 데이터를 상기 데이터 출력부를 통해 그대로 출력하는 것을 특징으로 한다.
- [33] 상기 다른 목적을 달성하기 위한 본 발명의 의학용어 오탈자 수정 방법의 상기 유사도는 편집 거리, 빈도 및 Character level N-gram을 적용하는 것을 특징으로 한다.
- [34] 기타 실시예의 구체적인 사항은 "발명을 실시하기 위한 구체적인 내용" 및 첨부 "도면"에 포함되어 있다.
- [35] 본 발명의 이점 및/또는 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 각종 실시예를 참조하면 명확해질 것이다.
- [36] 그러나, 본 발명은 이하에서 개시되는 각 실시예의 구성만으로 한정되는 것이 아니라 서로 다른 다양한 형태로도 구현될 수도 있으며, 단지 본 명세서에서 개시한 각각의 실시예는 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는

기술 분야에서 통상의 지식을 가진 자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구범위의 각 청구항의 범주에 의해 정의될 뿐임을 알아야 한다.

발명의 효과

- [37] 본 발명에 의할 경우, 의학 분야 및 의료업계에서 작성되는 법적 효력이 있는 문서에 있는 비정형 텍스트 데이터를 두 가지 이상의 단어 임베딩 매트릭스를 적용하여 정제함으로써, 의료기관별로 사용되는 단어가 상이한 한계를 극복할 수 있게 된다.
- [38] 또한, 현재 생성되는 오탈자뿐 아니라 앞으로 발생할 수 있는 오탈자에 대하여도 일관성 있게 수정할 수 있게 된다.

도면의 간단한 설명

- [39] 도 1은 본 발명의 일 실시예에 따른 의학용어 오탈자 수정 장치의 블록도이다.
- [40] 도 2는 본 발명의 일 실시예에 따른 의학용어 오탈자 수정 방법의 동작을 설명하기 위한 순서도이다.
- [41] 도 3은 소정의 종합병원에서 사용하는 세균배양 검사 보고서에 대해 본 발명의 일 실시예에 따라 적용한 결과 오탈자 단어 수에 대한 표이다.
- [42] 도 4는 소정의 종합병원에서 사용하는 세균배양 검사 보고서에 대해 본 발명의 일 실시예에 따라 적용한 결과 데이터가 수정된 예시에 대한 표이다.

발명의 실시를 위한 최선의 형태

- [43] 이하, 첨부한 도면을 참고로 하여 본 발명의 바람직한 실시예에 대하여 상세히 설명하면 다음과 같다.
- [44] 본 발명을 상세하게 설명하기 전에, 본 명세서에서 사용된 용어나 단어는 통상적이거나 사전적인 의미로 무조건 한정하여 해석되어서는 아니되며, 본 발명의 발명자가 자신의 발명을 가장 최선의 방법으로 설명하기 위해서 각종 용어의 개념을 적절하게 정의하여 사용할 수 있다.
- [45] 더 나아가 이들 용어나 단어는 본 발명의 기술적 사상에 부합하는 의미와 개념으로 해석되어야 함을 알아야 한다.
- [46] 즉, 본 명세서에서 사용된 용어는 본 발명의 바람직한 실시예를 설명하기 위해서 사용되는 것일 뿐이고, 본 발명의 내용을 구체적으로 한정하려는 의도로 사용된 것이 아니다.
- [47] 이들 용어는 본 발명의 여러 가지 가능성을 고려하여 정의된 용어임을 알아야 한다.
- [48] 또한, 본 명세서에 있어서, 단수의 표현은 문맥상 명확하게 다른 의미로 지시하지 않는 이상, 복수의 표현을 포함할 수 있다.
- [49] 또한, 유사하게 복수로 표현되어 있다고 하더라도 단수의 의미를 포함할 수 있음을 알아야 한다.
- [50] 본 명세서의 전체에 걸쳐서 어떤 구성 요소가 다른 구성 요소를 "포함"한다고

기재하는 경우에는, 특별히 반대되는 의미의 기재가 없는 한 임의의 다른 구성 요소를 제외하는 것이 아니라 임의의 다른 구성 요소를 더 포함할 수도 있다는 것을 의미할 수 있다.

- [51] 더 나아가서, 어떤 구성 요소가 다른 구성 요소의 "내부에 존재하거나, 연결되어 설치된다"고 기재한 경우에는, 이 구성 요소가 다른 구성 요소와 직접적으로 연결되어 있거나 접촉하여 설치되어 있을 수 있다.
- [52] 또한, 일정한 거리를 두고 이격되어 설치되어 있을 수도 있으며, 일정한 거리를 두고 이격되어 설치되어 있는 경우에 대해서는 해당 구성 요소를 다른 구성 요소에 고정 내지 연결시키기 위한 제 3의 구성 요소 또는 수단이 존재할 수 있다.
- [53] 한편, 상기 제 3의 구성 요소 또는 수단에 대한 설명은 생략될 수도 있음을 알아야 한다.
- [54] 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결"되어 있다거나, 또는 "직접 접속"되어 있다고 기재되는 경우에는, 제 3의 구성 요소 또는 수단이 존재하지 않는 것으로 이해하여야 한다.
- [55] 마찬가지로, 각 구성 요소 간의 관계를 설명하는 다른 표현들, 즉 "~ 사이에"와 "바로 ~ 사이에", 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 취지를 가지고 있는 것으로 해석되어야 한다.
- [56] 또한, 본 명세서에 있어서 "일면", "타면", "일측", "타측", "제 1", "제 2" 등의 용어는, 하나의 구성 요소에 대해서 이 하나의 구성 요소가 다른 구성 요소로부터 명확하게 구별될 수 있도록 하기 위해서 사용된다.
- [57] 하지만, 이와 같은 용어에 의해서 해당 구성 요소의 의미가 제한적으로 사용되는 것은 아님을 알아야 한다.
- [58] 또한, 본 명세서에서 "상", "하", "좌", "우" 등의 위치와 관련된 용어는, 사용된다면, 해당 구성 요소에 대해서 해당 도면에서의 상대적인 위치를 나타내고 있는 것으로 이해하여야 한다.
- [59] 또한, 이들의 위치에 대해서 절대적인 위치를 특정하지 않는 이상은, 이들 위치 관련 용어가 절대적인 위치를 언급하고 있는 것으로 이해하여서는 아니 된다.
- [60] 더욱이, 본 발명의 명세서에서는, "...부", "...기", "모듈", "장치" 등의 용어는, 사용된다면, 하나 이상의 기능이나 동작을 처리할 수 있는 단위를 의미한다.
- [61] 이는 하드웨어 또는 소프트웨어, 또는 하드웨어와 소프트웨어의 결합으로 구현될 수 있음을 알아야 한다.
- [62] 본 명세서에 첨부된 도면에서 본 발명을 구성하는 각 구성 요소의 크기, 위치, 결합 관계 등은 본 발명의 사상을 충분히 명확하게 전달할 수 있도록 하기 위해서 또는 설명의 편의를 위해서 일부 과장 또는 축소되거나 생략되어 기술되어 있을 수 있고, 따라서 그 비례나 축척은 엄밀하지 않을 수 있다.
- [63] 또한, 이하에서, 본 발명을 설명함에 있어서, 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 구성, 예를 들어, 종래 기술을 포함하는 공지 기술에 대한

상세한 설명은 생략될 수도 있다.

- [64] 도 1은 본 발명의 일 실시예에 따른 의학용어 오탈자 수정 장치의 블록도로서, 데이터 도입부(100), 전처리 필터부(200), 오탈자 선별부(300), 전-훈련 임베딩 매트릭스(400), 특화 임베딩 매트릭스(500), 유사도 분석부(600), 오탈자 재검부(700), 데이터 수정부(800) 및 데이터 출력부(900)를 구비한다.
- [65] 도 2는 본 발명의 일 실시예에 따른 의학용어 오탈자 수정 방법의 동작을 설명하기 위한 순서도이다.
- [66] 도 1 및 도 2를 참조하여 본 발명의 일 실시예에 따른 의학용어 오탈자 수정 방법의 동작을 개략적으로 설명하면 다음과 같다.
- [67] 데이터 도입부(100)가 분석할 데이터를 메모리에 저장한다(S100).
- [68] 전처리 필터부(200)가 메모리에 저장된 데이터를 분석 단위로 분할하고, 전처리하여 필터링한다(S200).
- [69] 오탈자 선별부(300)가 전처리 필터부(200)에서 필터링된 데이터가 오탈자인지 선별한다(S300).
- [70] 전-훈련 임베딩 매트릭스(400)가 오탈자 선별부(300)에서 선별된 데이터 중에서 교정 단어 1차 후보군을 생성한다(S400).
- [71] 특화 임베딩 매트릭스(500)가 의료 기관에서 보관하고 있는 문서를 이용하여 임베딩 매트릭스를 구성하고, 교정 단어 2차 후보군을 생성한다(S500).
- [72] 유사도 분석부(600)가 선별된 오탈자와 전-훈련 임베딩 매트릭스(400)에서 생성된 교정 단어 1차 후보군과 특화 임베딩 매트릭스(500)에서 생성된 교정 단어 2차 후보군의 유사도를 분석하여 순위를 정하고 최종 후보군을 제시한다(S600).
- [73] 오탈자 재검부(700)가 유사도 분석부(600)에서 제시된 최종 후보군 데이터의 오탈자를 재검사한다(S700).
- [74] 데이터 수정부(800)가 오탈자 재검부(700)에서 재검사 결과 오탈자로 판단된 데이터를 수정한다(S800).
- [75] 데이터 출력부가 오탈자 선별부(300) 및 오탈자 재검부(700)에서 오탈자로 선별되지 않은 데이터를 그대로 출력하거나, 데이터 수정부(800)에서 수정된 데이터를 출력한다.
- [76] 도 3은 소정의 종합병원에서 사용하는 세균배양 검사 보고서에 대해 본 발명의 일 실시예에 따라 적용한 결과 오탈자 단어 수에 대한 표이다.
- [77] 도 4는 소정의 종합병원에서 사용하는 세균배양 검사 보고서에 대해 본 발명의 일 실시예에 따라 적용한 결과 데이터가 수정된 예시에 대한 표이다.
- [78] 도 1 내지 도 4를 참조하여 본 발명의 일 실시예에 따른 의학용어 오탈자 수정 장치의 유기적인 동작을 상세하게 설명하면 다음과 같다.
- [79] 데이터 도입부(100)는 단말기나 데이터베이스에 입력되어 있는 분석할 데이터를 메모리에 새로 저장한다.
- [80] 전처리 필터부(200)는 데이터를 분석 단위로 분할하고, 이후 분석이 용이하게

- 데이터 형식 정리, 불용어 전처리 등을 통해 데이터를 필터링한다.
- [81] 표준 용어 라이브러리(350)는 오탈자 선별부(300)에서 활용할 단어 라이브러리로서, 본 실시 예에서는 SNOMED Clinical Terms라는 임상의학 문서 및 보고서에 사용되는 체계적으로 구조화된 의학 용어 집합을 사용하여 20만건의 철자 사전을 구축한다.
- [82] 이때, 본 발명은 SNOMED에 국한되지 않을 뿐만 아니라, 후술하는 전-훈련 임베딩 매트릭스(400) 및 특화 임베딩 매트릭스(500)도 이러한 용도로 활용할 수도 있다.
- [83] 오탈자 선별부(300)는 입력된 데이터가 오탈자인지 선별하는 구성요소로서, 본 실시 예에서는 입력된 데이터의 단어가 표준 용어 라이브러리(350)에 존재하는지 여부를 소정의 규칙 기반으로 판단한다.
- [84] 판단 결과, 단어가 표준 용어 라이브러리(350)에 존재하지 않아 오탈자로 분류되는 경우에는 오탈자 가능성이 높은 단어로서 후속의 알고리즘을 따르게 되고, 오탈자로 분류되지 않는 경우에는 그대로 입력된 데이터를 출력한다.
- [85] 전-훈련 임베딩 매트릭스(400)는 사전에 훈련된 단어 라이브러리로서, 유사도 분석부(600)에서 사용할 교정 단어 1차 후보군을 생성한다.
- [86] 본 실시 예에서는 BioWordVec, Sub-word pre-trained model 등을 이용하여 구성한다.
- [87] BioWordVec은 PubMed와 MIMIC-III 임상 데이터베이스(Clinical Database)의 문서를 기반으로 200 차원의 사전 훈련된 임베딩 매트릭스를 제공한다.
- [88] 즉, MeSH(Medical Subject Headings)에서 heading(heading) 시퀀스를 생성하고 PubMed 문장 시퀀스와 결합하여 단어 임베딩 학습에 사용한다.
- [89] 특화 임베딩 매트릭스(500)는 의료기관에서 보관하고 있는 실제 문서를 이용하여 임베딩 매트릭스를 구성한다.
- [90] 앞서 전-훈련 임베딩 매트릭스(400)와 함께 유사도 분석부(600)에서 사용할 교정 단어 2차 후보군 생성에 활용한다.
- [91] 본 실시 예에서는 fastText 기법을 이용하여 임베딩 매트릭스를 구성한다.
- [92] 여기에서, fastText 기법은 철자 수준의 n-gram 벡터 합으로 표현된 매트릭스이다.
- [93] 철자 수준의 임베딩 벡터는 형태소 변화가 다양하거나 자주 사용되지 않는 희귀한 단어에도 적용할 수 있다는 강점을 가진다.
- [94] 예를 들어, 'fastText'라는 단어를 3-gram 으로 분해하는 경우, 'fas/ast/stT/tTe/Tex/ext'로 나누어진다.
- [95] 그리고, 이 6개의 단어들이 학습을 통해 임베딩되며 각 벡터의 합으로 전체 단어를 표현하게 된다.
- [96] 여기서, 교정 단어 1차 후보군 및 교정 단어 2차 후보군은 각각 오탈자로 분류된 단어에서 오탈자를 수정한 교정 단어의 후보를 의미한다.
- [97] 유사도 분석부(600)는 선별된 오탈자와 전-훈련 임베딩 매트릭스(400) 및 특화

- 임베딩 매트릭스(500)에서 생성된 교정 후보군(교정 단어 1차 후보군 및 교정 단어 2차 후보군)의 유사도를 분석하여 순위를 매기고 최종 후보군을 제시한다.
- [98] 즉, 전-훈련 임베딩 매트릭스(400)와 특화 임베딩 매트릭스(500)에서 한 번이라도 생성된 교정 후보 단어라면 전체 교정 후보군에 포함된다.
- [99] 유사도 분석부(600)는 전-훈련 임베딩 매트릭스(400)와 특화 임베딩 매트릭스(500)를 통하여 생성된 교정 후보군과 오탃자의 유사도를 각 임베딩 매트릭스에서 계산하고, 두 개의 교정 후보군에 모두 포함된 교정 후보 단어의 경우에는 두 개의 유사도 값을 기하 평균하여 최종 유사도를 산출한다.
- [100] 본 실시 예에서는 편집 거리(Levenshtein distance), 빈도, Character level N-gram을 적용한다.
- [101] 구체적으로, 본 실시 예에서는 편집 거리(Levenshtein distance)와 빈도를 고려하여 Character level N-gram 기반의 코사인 유사도를 산출한 후 교정 후보군의 순위를 매기어 최종 교정 단어를 제안한다.
- [102] 전-훈련 임베딩 매트릭스(400)와 특화 임베딩 매트릭스(500)는 fastText 기법으로 학습되었으며 빈도수에 따라 내림차순으로 어휘가 정렬되어 있다.
- [103] 전-훈련 임베딩 매트릭스(400)와 특화 임베딩 매트릭스(500)는 다음의 두 가지의 가정을 충분히 만족할 수 있다.
- [104] 첫번째, 임상 데이터베이스에서 올바른 철자의 단어는 철자 오류의 단어보다 상대적으로 빈번하게 나타날 수 있다.
- [105] 두번째, 학습에 사용되는 말뭉치가 클수록 올바른 철자의 단어의 빈도수는 크게 증가한다.
- [106] 편집 거리(Levenshtein distance)는 한 단어를 다른 단어로 변환하는데 필요한 작업 수다.
- [107] 오탃자와 교정 후보군의 편집 거리의 범위를 설정한 후 조건에 맞는 교정 후보군에 대하여 최종적으로 코사인 유사도를 산출한다.
- [108] 교정 후보군의 산출된 유사도를 기반으로 순위를 정렬한 후 최종 후보군을 제시한다.
- [109] 즉, 유사도 분석부(600)는 유사도가 높은 순으로 교정 후보군을 정렬한 후에, 빈도를 기준으로 선택적으로 순위를 매기게 되며, 오탃자와 가장 유사한 단어의 범위를 제한적으로 탐색하는 과정을 통하여 주어진 조건에 맞는 최종 후보군을 제시한다.
- [110] 오탃자 재검부(700)는 유사도 분석부(600)에서 제시된 최종 후보군 데이터의 오탃자를 다시 검사한다.
- [111] 데이터 출력부(900)는 오탃자 선별부(300) 및 오탃자 재검부(700)에서 올바른 단어(오탃자가 아닌 단어)로 판단되는 단어를 그대로 혹은 데이터 수정부(800)를 거쳐 수정하여 출력한다.
- [112] 만일, 최종 후보군이 표준 용어 라이브러리(350)에 없는 단어인 경우 유사도가 높은 단어를 출력한다.

- [113] 도 3에서 보는 바와 같이, 소정의 종합병원에서 사용하는 세균배양 검사 보고서를 기반으로 만든 180,000 여건의 샘플 데이터를 이용하여 본 발명을 적용한 결과, 총 18개 종류의 철자 오류가 발견되었고, 전체 오타자 단어의 개수는 918건이었음을 알 수 있다.
- [114] 또한, 도 4에서 보는 바와 같이, 도 3에서 기반으로 한 세균배양 검사 보고서에 대하여 본 발명을 적용한 결과, 전처리 필터 과정에서 필터링되어 수정된 예시가 Example 1 내지 4와 같이 수정되었음을 알 수 있다.
- [115] 이와 같이, 본 발명은 병원 등의 의료기관 및 의료관련 문서 및 문자형 데이터베이스에 존재하는 오타자를 자연어 처리 기법을 활용한 유사도 분석을 통해 자동으로 탐색하고, 두 가지 이상의 단어 임베딩 매트릭스를 병합하여 단어 목록을 구성함으로써 적절하게 수정할 수 있는 의학용어 오타자 수정 장치 및 방법을 제공한다.
- [116] 이를 통하여, 본 발명은 의학 분야 및 의료업계에서 작성되는 법적 효력이 있는 문서에 있는 비정형 텍스트 데이터를 두 가지 이상의 단어 임베딩 매트릭스를 적용하여 정제함으로써, 의료기관별로 사용되는 단어가 상이한 한계를 극복할 수 있게 된다.
- [117] 또한, 현재 생성되는 오타자뿐 아니라 앞으로 발생할 수 있는 오타자에 대하여도 일관성 있게 수정할 수 있게 된다.
- [118] 이상, 일부 예를 들어서 본 발명의 바람직한 여러 가지 실시예에 대해서 설명하였지만, 본 "발명을 실시하기 위한 구체적인 내용" 항목에 기재된 여러 가지 다양한 실시예에 관한 설명은 예시적인 것에 불과한 것이며, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 이상의 설명으로부터 본 발명을 다양하게 변형하여 실시하거나 본 발명과 균등한 실시를 행할 수 있다는 점을 잘 이해하고 있을 것이다.
- [119] 또한, 본 발명은 다른 다양한 형태로 구현될 수 있기 때문에 본 발명은 상술한 설명에 의해서 한정되는 것이 아니며, 이상의 설명은 본 발명의 개시 내용이 완전해지도록 하기 위한 것으로 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이며, 본 발명은 청구범위의 각 청구항에 의해서 정의될 뿐임을 알아야 한다.

청구범위

- [청구항 1] 선별된 오탈자와 전-훈련 임베딩 매트릭스 및 특화 임베딩 매트릭스에서 생성된 교정 후보군의 유사도를 분석하는 유사도 분석부; 및
상기 유사도 분석 결과에 따라 오탈자로 판단된 데이터를 수정하는
데이터 수정부;
를 포함하는 것을 특징으로 하는,
의학용어 오탈자 수정 장치.
- [청구항 2] 제 1 항에 있어서,
분석할 데이터를 메모리에 저장하는 데이터 도입부;
상기 저장된 데이터를 분석 단위로 분할하고, 전처리하여 필터링하는
전처리 필터부; 및
상기 필터링된 데이터가 오탈자인지 선별하는 오탈자 선별부;
를 더 포함하고,
상기 전-훈련 임베딩 매트릭스는 상기 선별된 데이터 중에서 교정 단어
1차 후보군을 생성하고,
상기 특화 임베딩 매트릭스는 의료 기관에서 보관하고 있는 문서를
이용하여 임베딩 매트릭스를 구성하고, 교정 단어 2차 후보군 생성하며,
상기 유사도 분석부는 선별된 오탈자와 상기 생성된 교정 단어 1차 및 2차
후보군의 유사도를 분석하여 순위를 정하고 최종 후보군을 제시하는
것을 특징으로 하는,
의학용어 오탈자 수정 장치.
- [청구항 3] 제 2 항에 있어서,
상기 제시된 최종 후보군 데이터의 오탈자를 재검사하는 오탈자 재검부;
및
상기 오탈자 선별부 및 상기 오탈자 재검부에서 오탈자로 선별되지 않은
데이터를 그대로 출력하거나, 상기 수정된 데이터를 출력하는 상기
데이터 출력부;
를 더 포함하고,
상기 데이터 수정부는 상기 재검사 결과 오탈자로 판단된 데이터를
수정하는 것을 특징으로 하는,
의학용어 오탈자 수정 장치.
- [청구항 4] 제 2 항에 있어서,
상기 저장된 데이터는
단말기 또는 데이터베이스에 입력되어 있는 데이터인 것을 특징으로
하는,
의학용어 오탈자 수정 장치.
- [청구항 5] 제 1 항에 있어서,

- 상기 전-훈련 임베딩 매트릭스는
사전에 훈련된 단어 라이브러리인 것을 특징으로 하는,
의학용어 오탃자 수정 장치.
- [청구항 6] 제 1 항에 있어서,
상기 전-훈련 임베딩 매트릭스는
BioWordVec, Sub-word pre-trained model을 포함하는 것을 특징으로 하는,
의학용어 오탃자 수정 장치.
- [청구항 7] 제 1 항에 있어서,
상기 특화 임베딩 매트릭스는
fastText 기법을 이용하여 임베딩 매트릭스를 구성하는 것을 특징으로
하는,
의학용어 오탃자 수정 장치.
- [청구항 8] 유사도 분석부가 선별된 오탃자와 전-훈련 임베딩 매트릭스 및 특화
임베딩 매트릭스에서 생성된 교정 후보군의 유사도를 분석하는 단계; 및
데이터 수정부가 상기 유사도 분석 결과에 따라 오탃자로 판단된
데이터를 수정하는 단계;
를 포함하는 것을 특징으로 하는,
의학용어 오탃자 수정 방법.
- [청구항 9] 제 8 항에 있어서,
(a) 데이터 도입부가 분석할 데이터를 메모리에 저장하는 단계;
(b) 전처리 필터부가 상기 저장된 데이터를 분석 단위로 분할하고,
전처리하여 필터링하는 단계;
(c) 오탃자 선별부가 상기 필터링된 데이터가 오탃자인지 선별하는 단계;
(d) 상기 전-훈련 임베딩 매트릭스가 상기 선별된 데이터 중에서 교정
단어 1차 후보군을 생성하는 단계;
(e) 상기 특화 임베딩 매트릭스가 의료 기관에서 보관하고 있는 문서를
이용하여 임베딩 매트릭스를 구성하고, 교정 단어 2차 후보군을 생성하는
단계;
(f) 상기 유사도 분석부가 선별된 오탃자와 상기 생성된 교정 단어 1차 및
2차 후보군의 유사도를 분석하여 순위를 정하고 최종 후보군을 제시하는
단계;
(g) 오탃자 재검부가 상기 제시된 최종 후보군 데이터의 오탃자를
재검사하는 단계; 및
(h) 상기 데이터 수정부가 상기 재검사 결과 오탃자로 판단된 데이터를
수정하는 단계;
를 포함하는 것을 특징으로 하는,
의학용어 오탃자 수정 방법.
- [청구항 10] 제 9 항에 있어서,

상기 (h) 단계 이후에,
 데이터 출력부가 상기 오탈자 선별부 및 상기 오탈자 재검부에서
 오탈자로 선별되지 않은 데이터를 그대로 출력하거나, 상기 수정된
 데이터를 출력하는 단계;
 를 더 포함하는 것을 특징으로 하는,
 의학용어 오탈자 수정 방법.

[청구항 11] 제 9 항에 있어서,
 상기 전처리는
 형식 정리 및 불용어 전처리를 포함하는 것을 특징으로 하는,
 의학용어 오탈자 수정 방법.

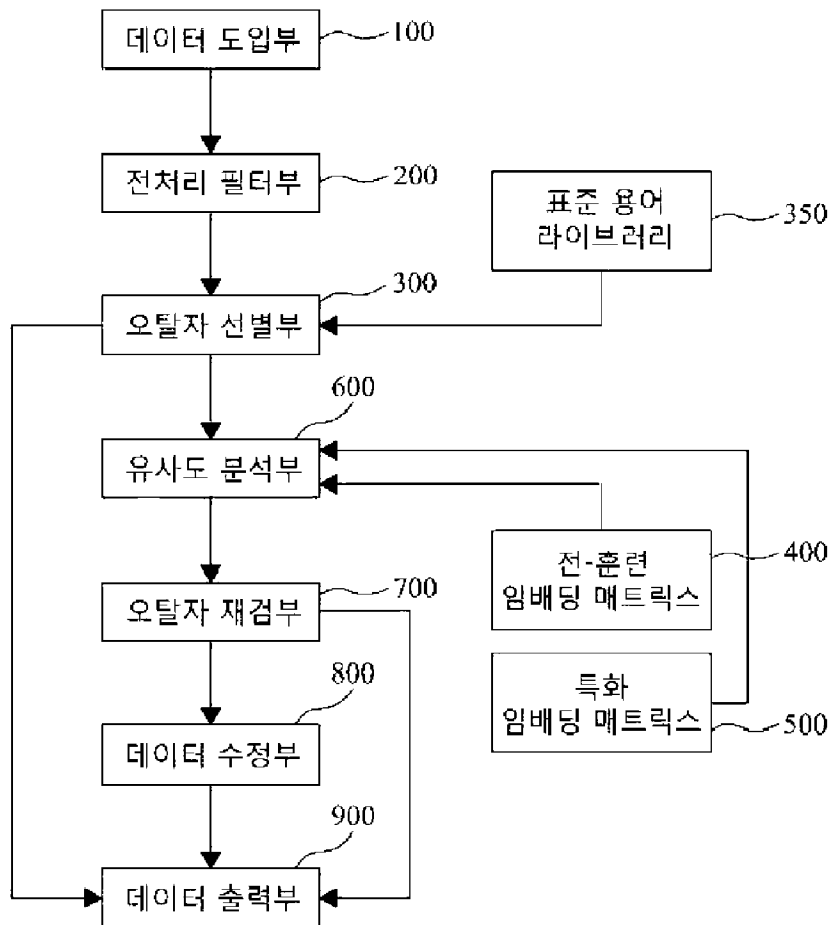
[청구항 12] 제 10 항에 있어서,
 상기 (c) 단계에서,
 상기 오탈자인지 선별은
 입력된 데이터의 단어가 표준 용어 라이브러리에 존재하는지 여부를
 소정의 규칙 기반으로 판단하는 것을 특징으로 하는,
 의학용어 오탈자 수정 방법.

[청구항 13] 제 12 항에 있어서,
 상기 표준 용어 라이브러리는
 SNOMED Clinical Terms 임상의학 문서 및 보고서에 사용되는 의학 용어
 집합을 사용하는 것을 특징으로 하는,
 의학용어 오탈자 수정 방법.

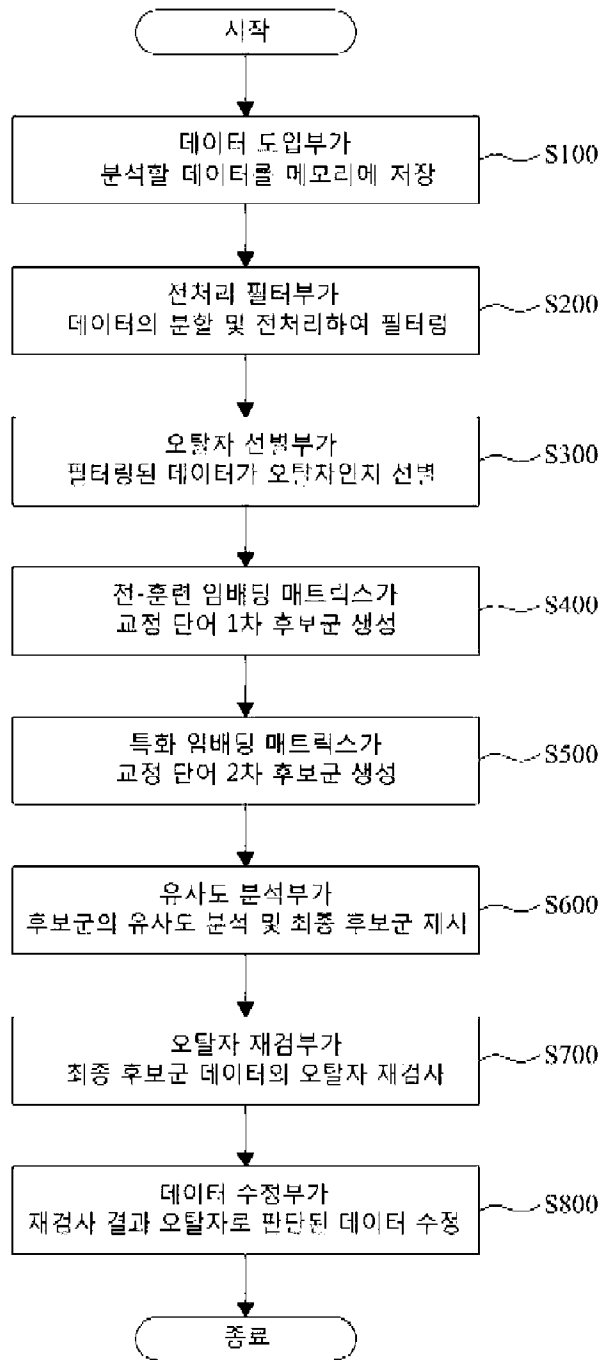
[청구항 14] 제 12 항에 있어서,
 상기 소정의 규칙은
 오탈자로 분류되는 경우 후속의 알고리즘을 따르고,
 오탈자로 분류되지 않는 경우 상기 저장된 데이터를 상기 데이터
 출력부를 통해 그대로 출력하는 것을 특징으로 하는,
 의학용어 오탈자 수정 방법.

[청구항 15] 제 8 항에 있어서,
 상기 유사도는
 편집 거리, 빈도 및 Character level N-gram을 적용하는 것을 특징으로
 하는,
 의학용어 오탈자 수정 방법.

[도1]



[도2]



[도3]

Misspelling	Miscount	Misspelling	Miscount
adecarboxylate	19	orytihabitans	1
chromogens	2	papatyphi	7
ferentum	1	parpinfluenzae	18
flavbacterium	2	pseudodiphthericum	6
koneensis	1	shingobacterium	1
nonenterococci	1	sstreptococcus	21
nonenterococcus	2	stacherbrandfii	1
ochrobacterium	1	stapylcoccus	827
		urealyticm	5

[도4]

	Misspelling	Candidates
Example 1	<i>staphylococcus</i> -> <i>stapylococcus</i>	staphylococcus
Example 2	<i>streptococcus</i> -> <i>sstreptococcus</i>	streptococcus
Example 3	<i>adecarboxylata</i> -> <i>adecarboxylata</i>	adecarboxylata
Example 4	<i>aaratyphi</i> -> <i>papatyphi</i>	paratyphi

INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2022/000061

A. CLASSIFICATION OF SUBJECT MATTER G06F 40/232(2020.01)i; G06F 16/215(2019.01)i; G06F 16/2458(2019.01)i; G16H 15/00(2018.01)i; G06N 20/00(2019.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F 40/232(2020.01); G06F 17/30(2006.01); G06F 17/40(2006.01) Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models: IPC as above Japanese utility models and applications for utility models: IPC as above Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS (KIPO internal) & keywords: 오타자(error), 임베딩 매트릭스(embedding matrix), 교정(correction), 후보군(candidate), 유사도(similarity)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LU, Chris J. et al. Spell checker for consumer language (CSpell). Journal of the American Medical Informatics Association. vol. 26, no. 3, pp. 211-218, 21 January 2019. [Retrieved on 13 March 2022]. Retrieved from <URL: https://academic.oup.com/jamia/article/26/3/211/5298352>.	1-3,8-12,14-15
Y	See pages 212 and 215-216; table 3; and figures 1 and 3.	4-7,13
Y	HIRST, Graeme et al. Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering. vol. 11, no. 1, pp. 87-111, 28 February 2005. [Retrieved on 13 March 2022]. Retrieved from <URL: http://journals.cambridge.org/abstract_S1351324904003560>. See page 97.	4
Y	KHATTAK, Faiza Khan et al. A survey of word embeddings for clinical text. Journal of Biomedical Informatics. vol. 100, pp. 1-18, 28 October 2019. [Retrieved on 13 March 2022]. Retrieved from <URL: https://www.sciencedirect.com/science/article/pii/S2590177X19300563>. See pages 6 and 13.	5-7,13
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "D" document cited by the applicant in the international application "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 24 March 2022		Date of mailing of the international search report 25 March 2022
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex-Daejeon Building 4, 189 Cheongsaro, Seo-gu, Daejeon 35208 Facsimile No. +82-42-481-8578		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2022/000061**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2007-0098252 A (NHN CORPORATION) 05 October 2007 (2007-10-05) See paragraphs [0037]-[0074]; and figures 3-8.	1-15
A	KR 10-2013-0067660 A (KOREA INSTITUTE OF ORIENTAL MEDICINE) 25 June 2013 (2013-06-25) See paragraphs [0023]-[0040]; and figure 1.	1-15

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2022/000061

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
KR	10-2007-0098252	A	05 October 2007	KR	10-0806936	B1	22 February 2008
KR	10-2013-0067660	A	25 June 2013	KR	10-1293832	B1	07 August 2013

A. 발명이 속하는 기술분류(국제특허분류(IPC)) G06F 40/232(2020.01)i; G06F 16/215(2019.01)i; G06F 16/2458(2019.01)i; G16H 15/00(2018.01)i; G06N 20/00(2019.01)i		
B. 조사된 분야 조사된 최소문헌(국제특허분류를 기재) G06F 40/232(2020.01); G06F 17/30(2006.01); G06F 17/40(2006.01) 조사된 기술분야에 속하는 최소문헌 이외의 문헌 한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC 국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우)) eKOMPASS(특허청 내부 검색시스템) & 키워드: 오타자(error), 임베딩 매트릭스(embedding matrix), 교정(correction), 후보군(candidate), 유사도(similarity)		
C. 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	CHRIS J. LU et al., Spell checker for consumer language (CSpell), Journal of the American Medical Informatics Association, 제26권, 제3호, pp. 211-218, 2019.01.21 [검색일: 2022.03.13], 출처: <URL: https://academic.oup.com/jamia/article/26/3/211/5298352 > 페이지 212, 215-216; 표 3; 및 도면 1, 3	1-3,8-12,14-15
Y		4-7,13
Y	GRAEME HIRST et al., Correcting real-word spelling errors by restoring lexical cohesion, Natural Language Engineering, 제11권, 제1호, pp. 87-111, 2005.02.28 [검색일: 2022.03.13], 출처: <URL: http://journals.cambridge.org/abstract_S1351324904003560 > 페이지 97	4
Y	Faiza Khan Khattak et al., A survey of word embeddings for clinical text, Journal of Biomedical Informatics, 제100권, pp. 1-18, 2019.10.28 [검색일: 2022.03.13], 출처: <URL: https://www.sciencedirect.com/science/article/pii/S2590177X19300563 > 페이지 6, 13	5-7,13
☒ 추가 문헌이 C(계속)에 기재되어 있습니다. ☒ 대응특허에 관한 별지를 참조하십시오.		
* 인용된 문헌의 특별 카테고리: “A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌 “D” 본 국제출원에서 출원인이 인용한 문헌 “E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌 “L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌 “O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌 “P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌 “T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌 “X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다. “Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다. “&” 동일한 대응특허문헌에 속하는 문헌		
국제조사의 실제 완료일 2022년03월24일(24.03.2022)		국제조사보고서 발송일 2022년03월25일(25.03.2022)
ISA/KR의 명칭 및 우편주소 대한민국 특허청 (35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사) 팩스 번호 +82-42-481-8578		심사관 양정록 전화번호 +82-42-481-5709

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	KR 10-2007-0098252 A (엔에이치엔(주)) 2007.10.05 단락 [0037]-[0074]; 및 도면 3-8	1-15
A	KR 10-2013-0067660 A (한국 한위학 연구원) 2013.06.25 단락 [0023]-[0040]; 및 도면 1	1-15

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2007-0098252 A	2007/10/05	KR 10-0806936 B1	2008/02/22
KR 10-2013-0067660 A	2013/06/25	KR 10-1293832 B1	2013/08/07