(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2018 년 5 월 17 일 (17.05.2018)

WIPO | PCT

(10) 국제공개번호
WO 2018/088664 A1

(51) 국제특허분류:
G06F 17/27 (2006.01)

(21) 국제출원번호: PCT/KR2017/006916

(22) 국제출원일: 2017 년 6 월 29 일 (29.06.2017)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2016-0149597 2016 년 11 월 10 일 (10.11.2016) KR

(71) 출원인: 창원대학교 산학협력단 **(CHANGWON NATIONAL UNIVERSITY INDUSTRY UNIVERSITY COOPERATION FOUNDATION)** [KR/KR]; 51140 경상남도 창원시 의창구 창원대학로 20(사림동), Gyeongsangnam-do (KR).

(72) 발명자: **차정원 (CHA, Jeong Won)**; 51498 경상남도 창원시 성산구 동산로 115, 105동 1404호(상남동, 대동아파트), Gyeongsangnam-do (KR). **박태호 (PARK, Tae Ho)**; 35289 대전시 서구 도솔로245번길 46-10(내동), Daejeon (KR). **신창욱 (SHIN, Chang Uk)**; 50905 경상남도 김해시 가락로 303, 가동 306호(구산동, 유신장미타워아파트), Gyeongsangnam-do (KR). **박다솔 (PARK, Da Sol)**; 50980 경상남도 김해시 능동로 117, 414동 1101호(부곡동, 석봉마을4단지부영아파트), Gyeongsangnam-do (KR). **박성재 (PARK, Seong Jae)**; 51429 경상남도 창원시 의창구 창이대로478번길 19(용호동), Gyeongsangnam-do (KR).

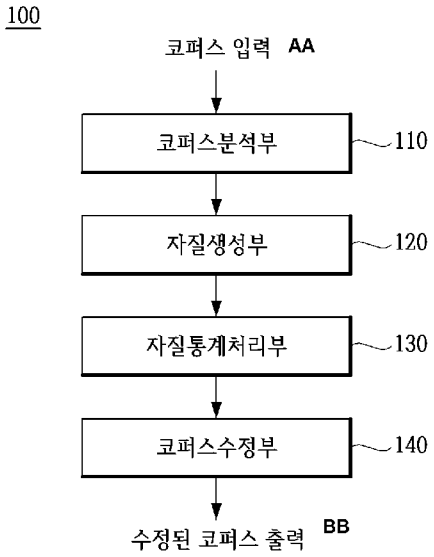
(74) 대리인: **김정수 (KIM, Jung Su)**; 05543 서울시 송파구 올림픽로 360, 5층(방이동), Seoul (KR).

(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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(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), 유

(54) **Title:** DEVICE FOR AUTOMATICALLY DETECTING MORPHEME PART OF SPEECH TAGGING CORPUS ERROR BY USING ROUGH SETS, AND METHOD THEREFOR

(54) **발명의 명칭:** 러프 셋을 이용한 형태소 품사 태깅 코퍼스 오류 자동 검출 장치 및 그 방법



110 ... Corpus analysis unit
120 ... Attribute generating unit
130 ... Attribute statistics processing unit
140 ... Corpus modifying unit
AA ... Corpus input
BB ... Modified corpus output

(57) **Abstract:** A device for detecting a morpheme tagging corpus error, of the present invention, comprises: an attribute generating unit (120) for generating attributes for word phrases included in an input corpus, by using a kernel to which a rough set theory is applied; and an attribute statistics processing unit (130) for generating part of speech tagging corpus error data through the calculation of attributes and frequency count for the same word phrases by counting attributes for the same word phrase among the word phrases, and thus the present invention can detect, quantify, and modify errors included in a corpus (learning data) required in learning for classifier generation and recognition for natural language processing.

(57) **요약서:** 본 발명의 형태소 태깅 코퍼스 오류 검출 장치는 입력된 코퍼스에 포함된 어절들에 대하여 러프 셋 이론을 적용한 커널을 이용하여 자질을 생성하는 자질생성부(120); 및 상기 어절들 중 동일 어절에 대한 자질을 카운트하여 동일 어절들에 대한 자질들과 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를 생성하는 자질통계부(130); 을 포함하여 구성되어, 자연어 처리를 위한 인식 및 분류기 생성을 위해 학습에 필요한 코퍼스(corpus, 말뭉치, 학습데이터)에 포함되는 오류를 검출하여 정량화할 수 있도록 하고 수정할 수 있도록 한다.

WO 2018/088664 A1

[다음 쪽 계속]

럽 (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: 러프 셋을 이용한 형태소 품사 태깅 코퍼스 오류 자동 검출 장치 및 그 방법

기술분야

[1] 본 발명은 자연어처리를 위한 코퍼스의 형태소 품사 오류 검출에 관한 것으로서, 더욱 상세하게는, 자연어 처리를 위한 인식 및 분류기 생성을 위해 학습에 필요한 코퍼스(corpus, 말뭉치, 학습데이터)에 포함되는 오류를 자동으로 검출하여 정량화할 수 있도록 하고 수정할 수 있도록 하는 러프 셋(Rough Set)을 이용한 형태소 품사 태깅 코퍼스 오류 검출 장치 및 그 방법에 관한 것이다.

배경기술

[2] 최근, 컴퓨터와 모바일 기기가 각 개인에게까지 널리 보급되고, 또한, 기계학습을 이용한 문제해결이 점점 더 많은 분야로 확대됨에 따라, 입력된 코퍼스에서 문맥이나 구문의 오류를 분석하고 수정하는 방법에 대하여 여러 가지 연구가 활발히 진행되고 있다.

[3] 이러한 입력된 코퍼스에서 문맥이나 구문의 오류를 분석하고 수정하는 방법에 관한 종래기술의 예로, 한국 등록특허공보 제10-1500617호는, 사용자가 입력한 한국어 문장에서 나타나는 여러 맞춤법 문법 오류 중에서 사전(事典) 검색을 통해 해결할 수 없는 문맥 철자오류(context-sensitive spelling error)를 검색하고, 이를 교정할 대치어를 제시하는 것에 의해 한국어 문서 교정 과정에서 가장 난도가 높은 문맥 철자오류를 교정함으로써, 한국어 문서 교정기의 성능을 높일 수 있도록 구성되는 한국어 어휘 의미망을 이용한 문맥 철자오류 교정 장치 및 방법을 개시한다.

[4] 또한, 한국 등록특허공보 제10-1491581호는, 철자오류 보정사전을 트라이(TRIE) 형태로 구성하는 것에 의해 메모리 사용량과 탐색 시간을 최소화하고, 등록되어 있는 문자열이 나타나면 문맥통계를 이용하여 해당 문자열을 보정 문자열로 교체할 것인지를 효율적으로 판단하는 것에 의해 작은 용량의 메모리와 단순 연산만을 이용하면서도 높은 철자오류 보정효과를 얻을 수 있도록 구성됨으로써, 휴대 단말기에서 입력된 문장의 철자 오류를 자동으로 인식하여 보정할 수 있는 철자 오류 보정 시스템 및 방법을 개시한다.

[5] 또한, 한국 등록특허공보 제10-1431339호에 따르면, 구문을 구성하는 각 단어가 코퍼스 내에서 출현할 출현확률을 구하고, 구문이 코퍼스 내에서 출현할 추정확률(Pe), 예상 출현빈도확률(Po) 및 실제 출현확률(Pa)을 구하여, 구문의 오류 여부를 판단하도록 구성됨으로써, 빈도수가 낮은 특수한 표현이나 반복적인 실수가 많이 행해지는 표현에 대해서도 오류검출을 정확히 할 수 있도록 구성되는 확률적 구문오류 검출방법 및 장치를 개시한다.

[6] 또한, 한국 등록특허공보 제10-1358614호에 따르면, 말뭉치를 분석하여

부분어절의 기분석 사전을 구축하는 것에 의해 간단하게 PWD(Partial Word morpheme madd Dictionary)와 형태소 위치 적합성을 구축하는 학습데이터 구축기 및 학습데이터 구축기에 의해 구축된 사전에 대하여 어절 전체를 둘 이상으로 나눈 뒤 검색하여 분석하는 형태소 분석기를 포함하여, 어절을 분석하는 속도 및 재현율을 높이고 태깅(Tagging)에서의 정확도를 높일 수 있도록 구성되는 말뭉치 기반의 한국어 형태소 분석장치 및 그 분석방법을 개시한다.

[7] 그러나 상술한 종래기술들의 경우 자연어 처리에 있어서, 문맥이나 구문 오류에 대한 검색 및 수정을 수행하는 기술만을 제공할 뿐, 자연어 처리의 기본이 되는 학습데이터로서의 코퍼스(corpus, 말뭉치)에 대한 오류를 검출하는 방법은 개시하지 못하고 있다.

[8] 또한, 종래 지도학습(supervised learning)을 대체하는 비지도학습(unsupervised learning)이나, 반지도학습(semi-supervised learning)에 대한 성공적인 연구결과에도 불구하고, 정보부착 코퍼스를 가공하여 활용할 수 있는 분야가 점점 증가하고 있음으로 인해 학습을 위한 정보부착 코퍼스의 중요성은 줄어들지 않고 있다.

[9] 상술한 바와 같은 코퍼스의 중요성의 증가에도 불구하고, 상기 대량의 코퍼스는 다수의 사람들의 수작업에 의해 제작되므로, 일관성 있는 코퍼스를 제작하기가 매우 어렵게 되고, 이에 따라, 제작된 코퍼스 오류 검정 또한 수작업으로 진행하게 되므로 시간과 비용이 크게 발생하는 문제점이 있었다.

[10] 따라서 자연어처리를 위한 코퍼스에 포함되는 오류를 검출하여 정량화하는 것에 의해 오류를 자동으로 수정할 수 있도록 하는 기술이 요구된다.

발명의 상세한 설명

기술적 과제

[11] 따라서 본 발명은 상술한 종래기술의 문제점을 해결하기 위한 것으로서, 자연어처리에서 인식 및 분류기 모델의 생성을 위해 다수의 사람에 의해 생성된 대용량의 코퍼스의 무결점을 보장할 수 있도록 하기 위하여, 자연어 처리를 위해 수작업으로 제작된 코퍼스의 오류를 자동으로 검출하고 정량화하며 수정할 수 있도록 하는 러프 셋을 이용한 형태소 품사 태깅 코퍼스 오류 검출 장치 및 그 방법을 제공하는 것을 목적으로 한다.

과제 해결 수단

[12] 상술한 목적을 달성하기 위한 본 발명의 러프 셋을 이용한 형태소 품사 태깅 코퍼스 오류 검출 장치는,

[13] 입력된 코퍼스에 포함된 어절들에 대하여 러프 셋 이론이 적용된 커널을 이용하여 자질을 생성하는 자질생성부(120); 및

[14] 상기 어절들 중 동일 어절에 대한 자질을 카운트하여 동일 어절들에 대한 자질들과 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를

생성하는 자질통계부(130);을 포함하여 구성될 수 있다.

[15] 상기 커널은,

[16] 입력된 코퍼스에서 분석 대상 어절에 대하여,

[17] 분석 대상 어절의 이전 어절의 형태소, 이전 어절의 품사, 현재 어절의 형태소, 다음 어절의 형태소, 다음 어절의 품사의 순으로 자질을 생성하도록 구성될 수 있다.

[18] 상기 형태소 품사 태킹 코퍼스 오류 검출 장치는,

[19] 상기 자질통계부(130)에서 생성된 어절별 자질의 통계자료를 이용하여, 동일한 어절에 대하여 가장 빈도수가 높은 자질로 생성된 자질을 수정하는 코퍼스수정부(140);를 더 포함하여 구성될 수 있다.

[20] 상기 형태소 품사 태킹 코퍼스 오류 검출 장치는,

[21] 품사 태깅이 수행된 학습데이터로서의 코퍼스를 입력 받아서 분석을 위한 데이터로 변환하는 코퍼스분석부(110);를 더 포함하여 구성될 수 있다.

[22] 상기 코퍼스분석부(110)는,

[23] 상기 입력된 코퍼스에 포함된 어절들 중 형태소와 수작업에 의한 형태소 입력값을 하나의 어절라인으로 순차적으로 형성하여 출력하도록 구성될 수 있다.

[24]

[25] 상술한 목적을 달성하기 위한 본 발명의 러프 셋을 이용한 형태소 품사 태깅 코퍼스 오류 검출 방법은,

[26] 코퍼스분석부(110)와 자질생성부(120), 자질통계부(130) 및 코퍼스수정부(140)를 포함하는 형태소 품사 태깅 코퍼스 오류 검출 장치에 의한 형태소 품사 태킹 코퍼스 오류 검출 방법에 있어서,

[27] 상기 자질생성부(120)가 입력된 코퍼스에 포함된 어절들에 대하여 러프 셋 이론이 적용되 커널을 이용하여 자질을 생성하는 자질생성과정(S120); 및

[28] 상기 자질통계부(130)가 상기 어절들 중 동일 어절에 대한 자질을 카운트하여 동일 어절들에 대한 자질들과 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를 생성하는 자질통계처리과정(S130);을 포함하여 구성될 수 있다.

[29] 상기 커널은,

[30] 입력된 코퍼스에서 분석 대상 어절에 대하여,

[31] 분석 대상 어절의 이전 어절의 형태소, 이전 어절의 품사, 현재 어절의 형태소, 다음 어절의 형태소, 다음 어절의 품사의 순으로 자질을 생성하도록 구성될 수 있다.

[32] 상기 형태소 품사 태킹 코퍼스 오류 검출 방법은,

[33] 상기 코퍼스수정부(140)가 상기 자질통계부(130)에서 생성된 어절별 자질의 통계자료를 이용하여, 동일한 어절에 대하여 가장 빈도수가 높은 자질로 생성된 자질을 수정하는 코퍼스수정과정(S140);을 더 포함하여 이루어 질 수 있다.

[34] 상기 러프 셋을 이용한 형태소 품사 태깅 코퍼스 오류 검출 방법은,

[35] 상기 코퍼스분석부(110)가 품사 태깅이 수행된 학습데이터로서의 코퍼스를

입력 받아서 분석을 위한 데이터로 변환한 후 상기 자질생성성부(120)로 출력하는 코퍼스분석과정(S100);을 더 포함할 수 있다.

[36] 상기 코퍼스분석과정(S100)은,

[37] 상기 코퍼스분석부(110)가, 상기 입력된 코퍼스에 포함된 어절들 중 형태소와 수작업에 의한 형태소 입력 값을 하나의 어절라인으로 순차적으로 형성하여 출력하는 코퍼스변환과정을 더 포함할 수 있다.

발명의 효과

[38] 상술한 구성을 가지는 본 발명은, 자연어처리 수행을 위해 다수의 작업자에 의해 수작업으로 형성된 코퍼스에 대하여 러프 셋을 개념을 이용하여 자동으로 어절들에 대한 자질을 생성한 후, 동일 어절에 대하여 빈도수가 높은 자질을 맞는 자질로 하여 오류를 검출하고 수정할 수 있도록 하는 효과를 제공한다.

도면의 간단한 설명

[39] 도 1은 본 발명의 실시예에 따르는 형태소 품사 태깅 코퍼스 오류 검출 장치(100)의 구성도.

[40] 도 2는 자질생성부(120)에 포함되는 러프 셋(Rough Set) 알고리즘이 적용된 커널의 예를 나타내는 도면.

[41] 도 3은 상기 형태소 품사 태깅 코퍼스 오류 검출 장기(100)가 소프트웨어로 구현된 후 컴퓨터에 설치되어 형성되는 품사 태깅 코퍼스 오류 검출 서버(1)의 기능 블록 구성도.

[42] 도 4는 본 발명의 형태소 품사 태깅 코퍼스 오류 검출 방법의 처리과정을 나타내는 순서도.

[43] 도 5는 오류 검사 대상 코퍼스의 예를 나타내는 도면.

[44] 도 6은 입력된 코퍼스가 오류 검사 분석을 위하여 변환된 분석 대상 데이터의 예를 나타내는 도면.

[45] 도 7은 상기 자질생성부(120)에 의해 생성된 자질을 포함하는 분석된 코퍼스의 예를 나타내는 도면.

발명의 실시를 위한 형태

[46] 하기에서 본 발명을 설명함에 있어서, 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.

[47] 본 발명의 개념에 따른 실시 예는 다양한 변경을 가할 수 있고 여러 가지 형태를 가질 수 있으므로, 특정 실시 예들을 도면에 예시하고 본 명세서 또는 출원서에 상세하게 설명하고자 한다. 그러나 이는 본 발명의 개념에 따른 실시 예를 특정한 개시 형태에 대해 한정하려는 것이 아니며, 본 발명은 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[48] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고

언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[49] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[50]

[51] 이하, 본 발명의 실시예를 나타내는 첨부 도면을 참조하여 본 발명을 더욱 상세히 설명한다.

[52] 도 1은 본 발명의 실시예에 따르는 형태소 품사 태깅 코퍼스 오류 검출 장치(100)(이하, '코퍼스 오류 검출 장치(100)'라 함)의 구성도이다.

[53] 도 1과 같이 상기 코퍼스 오류 검출 장치(100)는 코퍼스분석부(110), 자질생성부(120), 자질통계부(130) 및 코퍼스수정부(140)를 포함하여 구성될 수 있다.

[54] 상기 코퍼스분석부(110)는 품사 태깅이 수행된 학습데이터로서의 코퍼스를 입력 받아서 분석을 위한 데이터로 변환하도록 구성될 수 있다. 또한, 상기 코퍼스분석부(110)는 상기 입력된 코퍼스에 포함된 어절들 중 형태소와 수작업에 의한 형태소 입력 값을 하나의 어절라인으로 순차적으로 형성하여 출력하도록 구성될 수 있다.

[55] 상기 자질생성부(120)는 코퍼스의 어절을 분석한 후 자질을 생성할 수 있도록 하기 위한 러프 셋(Rough Set) 이론이 적용된 커널을 구비하여 상기 코퍼스분석부(110)에서 변환되어 입력된 코퍼스에 포함된 어절들에 대하여 자질을 생성하도록 구성될 수 있다.

[56] 도 2는 자질생성부(120)에 포함되는 러프 셋(Rough Set) 이론이 적용된 커널의 예를 나타내는 도면이다.

[57] 도 2와 같이, 상기 커널은, 입력된 코퍼스에서 분석 대상 어절에 대하여, 분석 대상 어절의 이전 어절의 형태소, 이전 어절의 품사, 현재 어절의 형태소, 다음 어절의 형태소, 다음 어절의 품사의 순으로 자질을 생성하도록 구성되는 것에 의해 어절별로 자질을 생성할 수 있도록 한다.

[58] 다시, 도 1을 참조하면, 상기 자질통계부(130)는 상기 어절들 중 동일 어절에

대한 자질을 카운트하여 동일 어절들에 대한 자질들과 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를 생성하도록 구성될 수 있다.

[59] 상기 코퍼스수정부(140)는, 상기 자질통계부(130)에서 생성된 어절별 자질의 통계자료를 이용하여, 동일한 어절에 대하여 가장 빈도수가 높은 자질로 생성된 자질을 수정하도록 구성된다. 이 경우, 상기 코퍼스의 형태소 품사 태깅은 수작업으로 진행하게 되므로, 가장 빈도수가 높은 자질이 정확한 자질로 평가되며, 빈도수가 낮을수록 오류 가능성이 높아지게 된다.

[60] 상술한 구성의 상기 코퍼스 오류 검출 장치(100)는 다수의 작업자에 의해 수작업으로 제작된 코퍼스를 입력받아서 코퍼스 오류를 검출하고 수정하여 출력하는 하드웨어 장치로 구성되거나, 컴퓨터로 읽혀져 수행되는 코드들을 기록한 기록매체의 형태로 제작될 수 있다.

[61] 도 3은 상기 형태소 품사 태깅 코퍼스 오류 검출 장기(100)가 소프트웨어로 구현된 후 컴퓨터에 설치되어 형성되는 품사 태깅 코퍼스 오류 검출 서버(1)의 기능 블록 구성도이다.

[62] 도 3과 같이, 상기 품사 태깅 코퍼스 오류 검출 서버(1)는 중앙처리장치로서의 제어부(10), 제어부(10)에 의해 실행되는 운영프로그램과 소프트웨어로 구현된 본 발명의 코퍼스 오류 검출장치(100)가 설치되는 저장부(60), 데이터의 입력 및 사용자 제어 명령을 입력할 수 있도록 구성되는 입력부(30), 내부 동작 과정을 표시하는 표시부(40) 및 외부와의 통신이 필요한 경우 외부와의 통신을 수행할 수 있도록 하는 통신부(50)를 포함하여 구성될 수 있다.

[63] 이와 달리, 상기 코퍼스 오류 검출 장치(100)는 FPGA 등이 적용되는 하드웨어 장치로 구현되어 제어부(10)의 일부로서 구성될 수도 있다.

[64] 도 4는 본 발명의 형태소 품사 태깅 코퍼스 오류 검출 방법(이하, '코퍼스 오류 검출 방법'이라 함)의 처리과정을 나타내는 순서도이다.

[65] 도 4와 같이, 상기 코퍼스 오류 검출 방법은, 코퍼스분석부(110)와 자질생성부(120), 자질통계부(130) 및 코퍼스수정부(140)를 포함하는 형태소 품사 태깅 코퍼스 오류 검출 장치에 의한 형태소 품사 태킹 코퍼스 오류 검출 방법에 있어서, 코퍼스분석과정(S100), 자질생성과정(S120), 자질통계처리과정(S130) 및 코퍼스수정과정(S140)을 포함하여 이루어진다.

[66] 상기 코퍼스분석과정(S110)은 품사 태깅이 수행된 학습데이터로서의 코퍼스를 입력 받아서 분석을 위한 데이터로 변환하는 처리과정을 수행한다.

[67] 도 5는 오류 검사 대상 코퍼스의 예를 나타내는 도면이다.

[68] 도 5와 같이, 입력되는 코퍼스는 어절 '그것은'에 대하여 다수의 작업자에 의해 수행된 형태소/(품사 태깅 코드)로 구성되는 '그것/NP+은/JX'로 되는 자질을 포함하여 구성된다. 상술한 코퍼스에서 본 발명은 '그것/NP+은/JX'의 자질에 오류가 포함되었는지를 검출하고 이를 수정하는 처리과정을 수행한다.

[69] 이러한 코퍼스 오류의 검출 또는 수정을 용이하게 하기 위하여, 상기 코퍼스분석과정(S110)에서 상기 코퍼스분석부(110)는 자질에 대하여 '형태소'와

'형태소/품사 태킹 코드'를 가지는 어절들의 라인으로 코퍼스를 변환하는 분석 데이터변환과정을 수행할 수 있다.

[70] 도 6은 입력된 코퍼스가 오류 검사 분석을 위하여 변환된 분석 대상 데이터의 예를 나타내는 도면이다.

[71] 도 6과 같이, 코퍼스분석과정(S110)의 분석 데이터변환과정에 의해 입력된 코퍼스들이 '그것(형태소) 그것(형태소)/NP(품사 태깅 코드)'를 가지는 어절 라인과 '은(형태소) 은(형태소)/JX(품사 태깅 코드)'를 가지는 어절 라인들을 가지도록 변환된 것을 확인할 수 있다.

[72] 다시 도 4을 참조하면, 상기 자질생성과정(S120)에서 상기 자질생성부(120)가 상기 코퍼스분석부(110)에서 변환되어 입력된 코퍼스에 포함된 어절들에 대하여 러프 셋 이론을 적용한 내부의 커널을 이용하여 분석 대상으로 변환된 코퍼스의 어절들에 대하여 각각 자질을 생성한다.

[73] 즉, 상기 자질생성부(120)는 변환된 코퍼스의 어절에 대하여. 이전 어절의 형태소, 이전 어절의 품사, 현재 어절의 형태소, 다음 어절의 형태소, 다음 어절의 품사의 순으로 추출하여 자질을 생성한다.

[74] 도 7은 상기 자질생성부(120)에 의해 생성된 자질을 포함하는 분석된 코퍼스의 예를 나타내는 도면이다.

[75] 도 7을 예로 들어 설명하면, '그것 그것/NP'의 어절에 대하여는 이전의 형태소와 이전의 품사가 없으므로 'X X'로 표시하고, 현재의 형태소는 자신이므로 '그것'을 표시하며, 다음 어절의 형태소는 다음 어절의 '은'을 표시하고, 다음의 어절의 품사는 은의 품사 'JX'를 추출한 후 표시하여, '그것 그것/NP'에 대하여 'X X 그것 은 JX'의 자질을 형성한다.

[76] 두 번째 어절인 '은 은/JX'에 대하여도, 같은 절차를 반복 수행하여 '그것 NP 은 사과 NNG'의 자질을 생성한다.

[77] 이러한 처리과정은 분석을 위해 변환된 코퍼스의 모든 어절들에 대하여 수행된다.

[78] 다시 도 4를 참조하면, 상기 자질통계처리과정(S130)에서, 상기 자질통계부(130)는 도 7과 같이 생성된 어절들에 대한 자질에 대하여 동일 어절에 대한 자질을 카운트하여 동일 어절들에 대한 서로 다른 자질들의 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를 생성한다. 일예로, 상기 자질통계처리과정(S130)은 '그것'. '은'. '사과다' 등의 어절들에 대한 서로 다른 자질들을 분류하고 각각 카운트한 후 각각의 서로 다른 자질들에 대하여 빈도수를 산출한다.

[79] 상기 코퍼스수정과정(S140)은 상기 코퍼스수정부(140)가 상기 자질통계부(130)에서 생성된 어절별 자질의 통계자료를 이용하여, 동일한 어절에 대하여 가장 빈도수가 높은 자질로 생성된 자질을 맞는 자질로 판단하여 이외의 자질들을 해당 자질로 수정하는 처리과정을 수행한다.

[80] 이와 같은 처리과정에 의해 수작업으로 생성된 코퍼스의 오류를 자동으로

검출하고 수정할 수 있게 된다.

[81]

[82] <실시예>

[83] 본 발명의 효율성을 검증하기 위하여, 다수의 연구자들이 손으로 작성한 품사부착 말뭉치인 코퍼스를 대상으로 실험하였다. 다수의 사람들에 의해서 생성된 코퍼스는 다양한 이유로 일관성에 문제가 발생한다. 이것은 지침이 부족해서 발생할 수도 있고 숙련도의 차이에 의해서 발생할 수도 있다. 언어정보 부착 코퍼스에서 일관성 오류(contradictionary)에 집중한다. 언어정보 부착 코퍼스에서 일관성 오류가 분류 오류보다 상대적으로 많고, 분류 오류는 말뭉치 내에서 오류와 정답을 비교할 수 없기 때문에 검출하기 어렵다.

[84] 우선 실험 데이터로서 생성된 코퍼스의 말뭉치와 수작업으로 수정한 정답에 대한 오류율이 표 1에 표시되어 있다.

[85] <표 1>은 초기 코퍼스에서 측정한 오류 수와 오류

[86]

말뭉치	정답 어절 수	오류 어절 수	오류율(%)
1	13,093	260	1.99
2	80,323	2,681	3.34
3	6,003	156	2.60

[87]

[88] 본 발명의 실시예의 경우 13,093 어절, 80.323 어절, 6,003 어절을 가지는 1, 2, 3 번의 코퍼스에 대하여 수작업으로 오류를 검출한 정답 코퍼스에 대한 오류율을 산출한 자료를 이용하였다.

[89]

[90] 상술한 입력 코퍼스를 이용하여 본원 발명의 자질생성부(120)가 커널을 이용하여 자질들을 생성한 후 생성된 자질들을 자질통계처리부(130)가 처리한 결과가 표 2에 표시된다.

[91] <표 2> 실험 대상 코퍼스에 대한 코퍼스 오류 검출 장치(100)에 의한 코퍼스 오류 검출 결과 표

[92]

말뭉치	오류율(%)	예측 오류 어절 수	예측 오류율(%)
1	1.19	271	2.07
2	3.34	2,419	3.01
3	2.60	74	1.23

[93] 표 2와 같이, 본 발명을 적용한 코퍼스 오류 검출 결과 초기의 코퍼스의 오류의 예측 값과 실제 코퍼스에 존재하는 오류를 비교하였을 때, 1번 코퍼스(1번 말뭉치)는 0.88%, 2번 코퍼스(2번 말뭉치)는 0.33%, 3번 코퍼스(3번 말뭉치)는

1.37% 차이가 발생하였다. 따라서 정답 코퍼스를 사용한 오류율과 거의 차이가 나지 않는 것을 알 수 있다.

[94]

[95] 상기에서 설명한 본 발명의 기술적 사상은 바람직한 실시예에서 구체적으로 기술되었으나, 상기한 실시예는 그 설명을 위한 것이며 그 제한을 위한 것이 아님을 주의하여야 한다. 또한, 본 발명의 기술적 분야의 통상의 지식을 가진자라면 본 발명의 기술적 사상의 범위 내에서 다양한 실시예가 가능함을 이해할 수 있을 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

산업상 이용가능성

[96] 본 발명은 컴퓨터를 이용한 언어처리 산업 분야에 적용될 수 있다.

청구범위

[청구항 1] 입력된 코퍼스에 포함된 어절들에 대하여 러프 셋 이론을 적용한 커널을 이용하여 자질을 생성하는 자질생성부(120); 및
상기 어절들 중 동일 어절에 대한 자질을 카운트하여 동일 어절들에 대한 자질들과 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를 생성하는 자질통계부(130);을 포함하여 구성되는 형태소 태깅 코퍼스 오류 검출 장치.

[청구항 2] 청구항 1에 있어서, 상기 커널은,
입력된 코퍼스에서 분석 대상 어절에 대하여,
분석 대상 어절의 이전 어절의 형태소, 이전 어절의 품사, 현재 어절의 형태소, 다음 어절의 형태소, 다음 어절의 품사의 순으로 자질을 생성하도록 구성되는 형태소 태깅 코퍼스 오류 검출 장치.

[청구항 3] 청구항 1에 있어서,
상기 자질통계부(130)에서 생성된 어절별 자질의 통계자료를 이용하여, 동일한 어절에 대하여 가장 빈도수가 높은 자질로 생성된 자질을 수정하는 코퍼스수정부(140);를 더 포함하여 구성되는 형태소 태깅 코퍼스 오류 검출 장치.

[청구항 4] 청구항 1에 있어서,
품사 태깅이 수행된 학습데이터로서의 코퍼스를 입력 받아서 분석을 위한 데이터로 변환하는 코퍼스분석부(110);를 더 포함하여 구성되는 형태소 태깅 코퍼스 오류 검출 장치.

[청구항 5] 청구항 4에 있어서, 상기 코퍼스분석부(110)는,
상기 입력된 코퍼스에 포함된 어절들 중 형태소와 수작업에 의한 형태소 입력 값을 하나의 어절라인으로 순차적으로 형성하여 출력하도록 구성되는 형태소 태깅 코퍼스 오류 검출 장치.

[청구항 6] 코퍼스분석부(110), 자질생성부(120), 자질통계부(130) 및
코퍼스수정부(140)를 포함하는 형태소 품사 태깅 코퍼스 오류 검출 장치에 의한 형태소 품사 태킹 코퍼스 오류 검출 방법에 있어서,
상기 자질생성부(120)가 입력된 코퍼스에 포함된 어절들에 대하여 러프 셋 이론을 적용한 커널을 이용하여 자질을 생성하는 자질생성과정(S120); 및
상기 자질통계부(130)가 상기 어절들 중 동일 어절에 대한 자질을 카운트하여 동일 어절들에 대한 자질들과 빈도수를 산출하는 것에 의해 품사 태깅 코퍼스 오류 데이터를 생성하는 자질통계처리과정(S130);을 포함하여 구성되는 형태소 태깅 코퍼스 오류 검출 방법.

[청구항 7] 청구항 6에 있어서, 상기 커널은,
입력된 코퍼스에서 분석 대상 어절에 대하여,

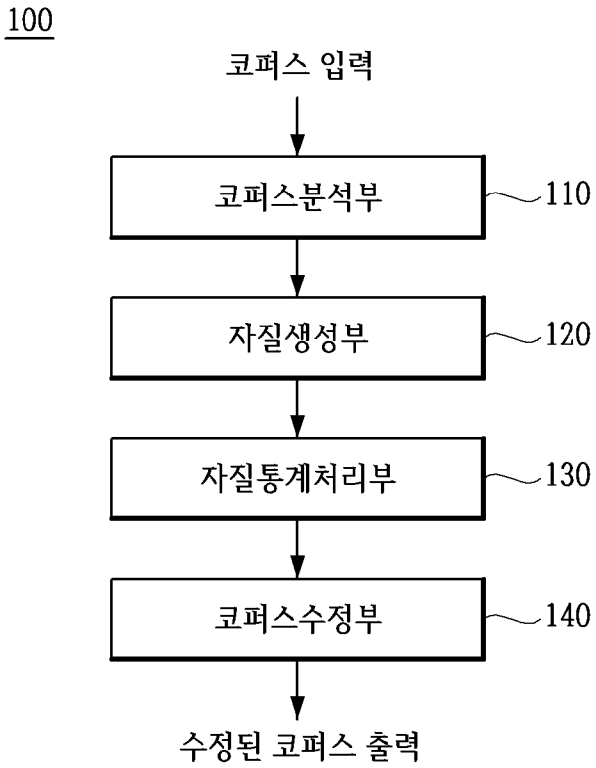
분석 대상 어절의 이전 어절의 형태소, 이전 어절의 품사, 현재 어절의 형태소, 다음 어절의 형태소, 다음 어절의 품사의 순으로 자질을 생성하도록 구성되는 형태소 태깅 코퍼스 오류 검출 방법.

[청구항 8] 청구항 1에 있어서,
상기 코퍼스수정부(140)가 상기 자질통계부(130)에서 생성된 어절별 자질의 통계자료를 이용하여, 동일한 어절에 대하여 가장 빈도수가 높은 자질로 생성된 자질을 수정하는 코퍼스수정과정(S140);을 더 포함하여 이루어지는 형태소 태깅 코퍼스 오류 검출 방법.

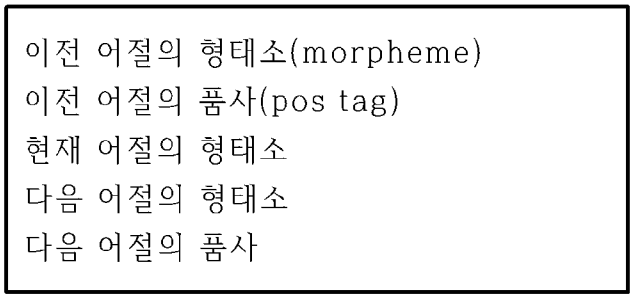
[청구항 9] 청구항 1에 있어서,
상기 코퍼스분석부(110)가 품사 태깅이 수행된 학습데이터로서의 코퍼스를 입력 받아서 분석을 위한 데이터로 변환한 후 상기 자질생성성부(120)로 출력하는 코퍼스분석과정(S100);을 더 포함하는 형태소 태깅 코퍼스 오류 검출 방법.

[청구항 10] 청구항 9에 있어서, 상기 코퍼스분석과정(S100)은,
상기 코퍼스분석부(110)가, 상기 입력된 코퍼스에 포함된 어절들 중 형태소와 수작업에 의한 형태소 입력 값을 하나의 어절라인으로 순차적으로 형성하여 출력하는 코퍼스변환과정을 더 포함하는 형태소 태깅 코퍼스 오류 검출 방법.

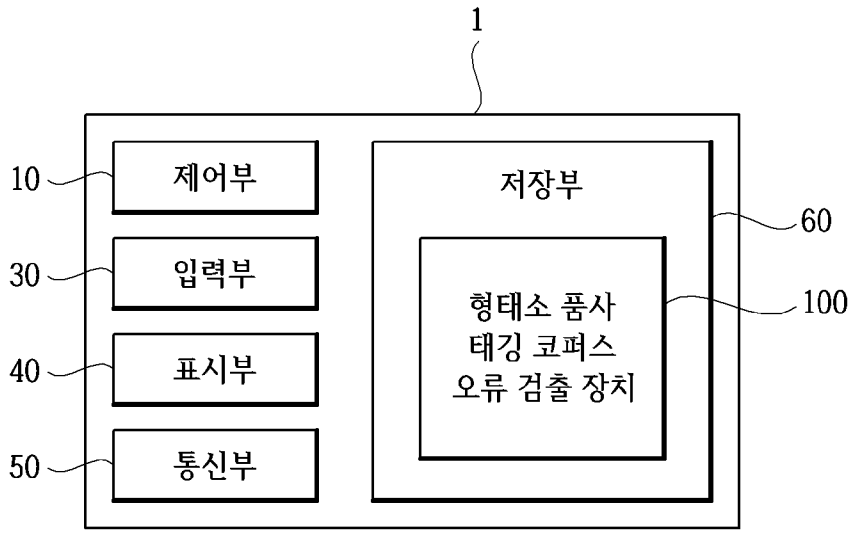
[도1]


100
코퍼스 입력
코퍼스분석부
110
자질생성부
120
자질통계처리부
130
코퍼스수정부
140
수정된 코퍼스 출력

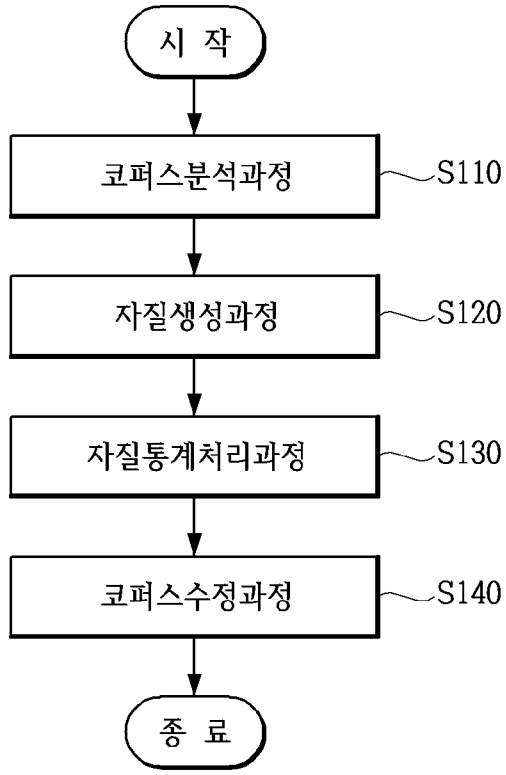
[도2]


이전 어절의 형태소(morpheme)
이전 어절의 품사(pos tag)
현재 어절의 형태소
다음 어절의 형태소
다음 어절의 품사

[도3]


1
10
제어부
30
입력부
40
표시부
50
통신부
저장부
60
형태소 품사
태깅 코퍼스
오류 검출 장치
100

[도4]



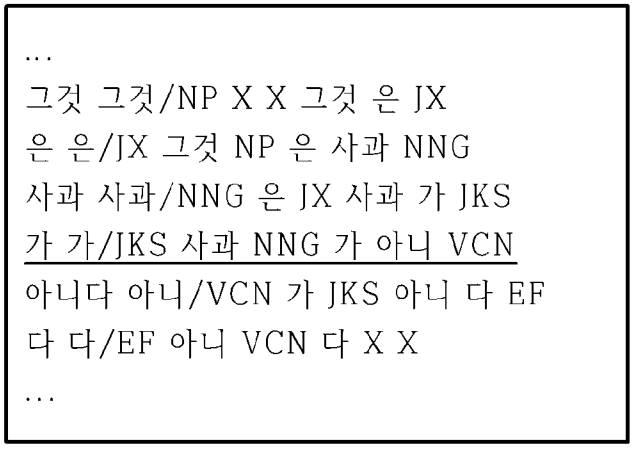
[도5]

```
…
그것은 그것/NP+은/JX
사과가 사과/NNG+가/JKS
아니다 아니/VCN+다/EF
…
```

[도6]

```
…
그것은 그것/NP
은 은/JX
사과 사과/NNG
가 가/JKS
아니다 아니/VCN
다 다/EF
…
```

[도7]


…
그것 그것/NP X X 그것 은 JX
은 은/JX 그것 NP 은 사과 NNG
사과 사과/NNG 은 JX 사과 가 JKS
가 가/JKS 사과 NNG 가 아니 VCN
아니다 아니/VCN 가 JKS 아니 다 EF
다 다/EF 아니 VCN 다 X X
…

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2017/006916

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(2006.01)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 17/27

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: corpus, syntactic word, rough set theory, kernel, feature, frequency, morpheme, tagging, error

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PARK, Tae-Ho et al., "Annotated Corpus Error Detection Using Rough Set", 2016 Korea Computer Congress 2000 (KCC 2016), pages 720-722, 30 June 2016. See pages 720-721 and tables 2-3.	1-2,4,6-7,9
A		3,5,8,10
A	CHUNG, Young-June et al., "Structure Optimization of Neural Networks Using Rough Set Theory", Proceedings of Korean Institute of Intelligent Systems(KIIS) Spring Conference, vol. 8, no. 1, pages 49-52, June 1998. See pages 49-52.	1-10
A	JP 2002-091961 A (COMMUNICATION RESEARCH LABORATORY) 29 March 2002 See paragraphs [0018]-[0027]; and figures 1-2.	1-10
A	KR 10-2012-0045906 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 09 May 2012 See paragraphs [0015]-[0022] and figure 1.	1-10
A	KR 10-2005-0039379 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 29 April 2005 See paragraphs [0019]-[0060] and figures 1-2.	1-10

☐ 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
"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	Date of mailing of the international search report
29 SEPTEMBER 2017 (29.09.2017)	**29 SEPTEMBER 2017 (29.09.2017)**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office Government Complex-Daejeon, 189 Seonsa-ro, Daejeon 302-701, Republic of Korea Facsimile No. +82-42-481-8578	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2017/006916

Patent document cited in search report	Publication date	Patent family member	Publication date
JP 2002-091961 A	29/03/2002	JP 3396734 B2	14/04/2003
KR 10-2012-0045906 A	09/05/2012	NONE	
KR 10-2005-0039379 A	29/04/2005	KR 10-0496873 B1	22/06/2005

국 제 조 사 보 고 서

국제출원번호
PCT/KR2017/006916

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06F 17/27(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)
G06F 17/27

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: 코퍼스, 어절, 러프 셋 이론, 커널, 자질, 빈도수, 형태소, 태깅, 오류

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
X	박태호 등. 'Rough Set을 이용한 형태소 품사 태깅 코퍼스 오류 정량화'. 2016 한국컴퓨터종합학술대회 (KCC 2016), 페이지 720-722. 2016.06.30. 페이지 720-721 및 표 2-3 참조.	1-2,4,6-7,9
A		3,5,8,10
A	정영준 등. '러프셋 이론을 이용한 신경망의 구조 최적화'. 한국지능시스템학회 학술발표 논문집, 제8권, 제1호, 페이지 49-52. 1998.06. 페이지 49-52 참조.	1-10
A	JP 2002-091961 A (COMMUNICATION RESEARCH LABORATORY) 2002.03.29 단락 [0018]-[0027]; 및 도면 1-2 참조.	1-10
A	KR 10-2012-0045906 A (한국전자통신연구원) 2012.05.09 단락 [0015]-[0022] 및 도면 1 참조.	1-10
A	KR 10-2005-0039379 A (한국전자통신연구원) 2005.04.29 단락 [0019]-[0060] 및 도면 1-2 참조.	1-10

☐ 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

"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	국제조사보고서 발송일
2017년 09월 29일 (29.09.2017)	2017년 09월 29일 (29.09.2017)

ISA/KR의 명칭 및 우편주소
대한민국 특허청
(35208) 대전광역시 서구 청사로 189,
4동 (둔산동, 정부대전청사)
팩스 번호 +82-42-481-8578

심사관
노지명
전화번호 +82-42-481-8528

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2017/006916

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
JP 2002-091961 A	2002/03/29	JP 3396734 B2	2003/04/14
KR 10-2012-0045906 A	2012/05/09	없음	
KR 10-2005-0039379 A	2005/04/29	KR 10-0496873 B1	2005/06/22