

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2011년 4월 28일 (28.04.2011)



(10) 국제공개번호
WO 2011/049313 A2

- (51) 국제특허분류:
G06Q 50/00 (2006.01) G06F 17/21 (2006.01)
- (21) 국제출원번호: PCT/KR2010/006943
- (22) 국제출원일: 2010년 10월 11일 (11.10.2010)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
10-2009-0100962 2009년 10월 23일 (23.10.2009) KR
- (71) 출원인 (US 을(를) 제외한 모든 지정국에 대하여): 포항공과대학교 산학협력단 (POSTECH ACADEMY-INDUSTRY FOUNDATION) [KR/KR]; 경상북도 포항시 남구 효자동 산 31, 790-784 Gyeongsangbuk-do (KR).
- (72) 발명자; 겸
- (75) 발명자/출원인 (US 에 한하여): 노형중 (NOH, Hyung Jong) [KR/KR]; 인천광역시 부평구 일신동 주공아파트 112동 901호, 403-764 Incheon (KR). 이종훈 (LEE, Jong Hoon) [KR/KR]; 대구광역시 달성군 다사읍 매곡리 태왕드림하이츠 101동 1901호, 711-765 Daegu

(KR). 이성진 (LEE, Sung Jin) [KR/KR]; 경상북도 포항시 남구 효자동 포항공과대학교 대학원아파트 1동 1403호, 790-784 Gyeongsangbuk-do (KR). 이근배 (LEE, Gary Geunbae) [KR/KR]; 경상북도 포항시 남구 지곡동 교수아파트 9-1003, 790-751 Gyeongsangbuk-do (KR).

(74) 대리인: 이철희 (LEE, Cheol Hee); 서울특별시 강남구 삼성동 156-13 우경빌딩 2층 가디언국제특허법률사무소, 135-090 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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, 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.

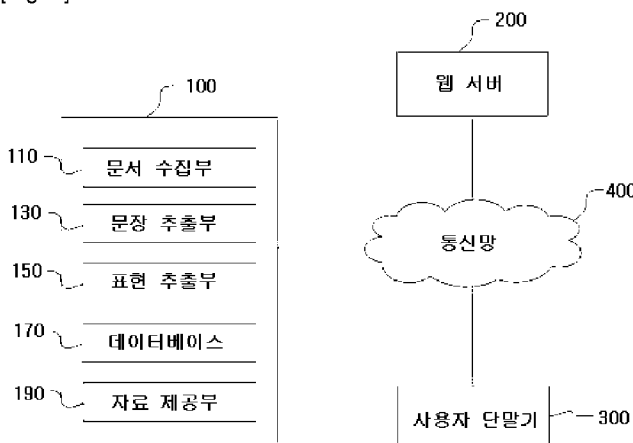
(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM,

[다음 쪽 계속]

(54) Title: APPARATUS AND METHOD FOR PROCESSING DOCUMENTS TO EXTRACT EXPRESSIONS AND DESCRIPTIONS

(54) 발명의 명칭: 표현 및 설명 추출을 위한 문서 처리 장치 및 방법

[Fig. 1]



- 110 ... Document collection unit
130 ... Sentence extraction unit
150 ... Expression extraction unit
170 ... Database
190 ... Material-providing unit
200 ... Web server
300 ... User terminal
400 ... Communication network

(57) Abstract: The present invention relates to an apparatus and method for processing documents to extract expressions and descriptions. The apparatus comprises: a document collection unit which collects documents from websites, and divides each of the collected documents into a script portion and a description portion to generate a script document and a description document; and an expression extraction unit which extracts expression description sentences on the basis of the description document, and extracts expressions described by the expression description sentences in the script document. According to the present invention, study material consisting of a pair of an expression to be studied and a description portion for describing the expression to be studied can be automatically obtained.

(57) 요약서: 본 발명은 표현 및 설명 추출을 위한 문서 처리 장치 및 방법에 관한 것으로서, 이 장치는, 웹으로부터 문서를 수집하고 이를 스크립트 부분과 설명 부분으로 분리하여 스크립트 문서 및 설명 문서를 생성하는 문서 수집부, 그리고 설명 문서에 기초하여 표현 설명 문장을 추출하고 표현 설명 문장이 스크립트 문서 내에서 설명하고 있는 표현을 추출하는 표현 추출부를 포함한다. 본 발명에 의하면, 학습 표현과 이에 대한 설명 부분이 하나의 쌍으로 이루어진 학습 자료를 자동으로 구축할 수 있다.



KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, MD, 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, ML, MR, NE, SN, TD, TG).

공개:

- 국제조사보고서 없이 공개하며 보고서 접수 후 이를 별도 공개함 (규칙 48.2(g))

명세서

발명의 명칭: 표현 및 설명 추출을 위한 문서 처리 장치 및 방법 기술분야

- [1] 본 발명은 문서 처리 장치 및 방법에 관한 것으로, 보다 상세하게는 웹에서 문서를 수집하고 수집된 문서로부터 표현 및 설명을 추출하는 문서 처리 장치 및 방법에 관한 것이다.

배경기술

- [2] 영어 교육의 중요성이 계속하여 강조되면서 이를 위한 다양한 교육 방법이 개발되고 있다. 그러나 현재 존재하는 대부분의 영어 교육 방법은 문제 풀이 및 암기 위주로 되어 있고, 읽기나 쓰기 부분에 특화되어 있다. 이와 같은 영어 교육 방법의 한계를 극복하기 위한 방안의 하나로 영어를 이용한 대화 시스템이 개발되어 이용되고 있다. 이 시스템은 영어 학습자와 시스템 사이의 대화를 통해 실생활에 필요한 다양한 표현을 자연스럽게 습득하도록 유도할 수 있다.
- [3] 그러나 대화 도중에 시스템에 의해 제시된 단어나 숙어 등의 영어 표현이 익숙하지 않는 경우 학습자는 이에 대한 설명을 확인하고 싶어할 수 있다. 이와 같은 경우 학습자는 주로 영어 사전 등을 통해 해당 영어 표현에 대한 설명을 확인할 수 있다. 하지만 영어 사전 등에서 영어 표현에 관한 설명을 제시하고 있기는 하나 영어 초심자를 이해시키기 위한 것으로 원론적인 설명이 주를 이루고 영어 표현을 한글로 설명하기 때문에 정확한 의미를 전달하는데 한계가 있다.

발명의 상세한 설명

기술적 과제

- [4] 본 발명이 해결하고자 하는 과제는 학습 표현과 이를 설명하는 설명 부분이 하나의 쌍으로 이루어진 학습 자료를 웹으로부터 자동으로 구축할 수 있는 문서 처리 장치 및 방법을 제공하는 것이다.

과제 해결 수단

- [5] 이러한 기술적 과제를 해결하기 위한 본 발명의 실시예에 따른 문서 처리 장치는, 웹으로부터 문서를 수집하고 상기 문서를 스크립트 부분과 설명 부분으로 분리하여 스크립트 문서 및 설명 문서를 생성하는 문서 수집부, 그리고 상기 설명 문서에 기초하여 표현 설명 문장을 추출하고 상기 표현 설명 문장이 상기 스크립트 문서 내에서 설명하고 있는 표현을 추출하는 표현 추출부를 포함한다.
- [6] 상기 스크립트 문서에서 문장을 분리하여 복수의 스크립트 문장을 추출하고 상기 설명 문서에서 문장을 분리하여 복수의 설명 문장을 추출하는 문장 추출부를 더 포함할 수 있다.
- [7] 상기 문장 추출부는 각 설명 문장이 상기 복수의 스크립트 문장 중 어느

- 스크립트 문장에 대한 설명인지 분류할 수 있다.
- [8] 상기 문장 추출부는 조건부 랜덤 필드(conditional random field) 분류기를 사용하여 상기 분류를 수행할 수 있다.
 - [9] 상기 표현 추출부는 레이블드 순차 패턴(labeled sequential pattern) 기법을 이용하여 상기 복수의 설명 문장으로부터 상기 표현 설명 문장을 추출할 수 있다.
 - [10] 상기 표현 추출부는 상기 표현 설명 문장과 상기 스크립트 문장 내의 표현의 유사도를 측정하여 상기 표현 설명 문장이 상기 스크립트 문장 내에서 설명하고 있는 표현을 추출할 수 있다.
 - [11] 상기 유사도는 코사인 유사도(cosine similarity) 기법이나 편집 거리(edit distance) 기법을 사용하여 측정될 수 있다.
 - [12] 상기 문서 수집부는 상기 스크립트 부분 및 상기 설명 부분의 시작 또는 끝을 나타내는 식별자 또는 문장에 기반하여 상기 스크립트 문서 및 상기 설명 문서를 생성할 수 있다.
 - [13] 상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 저장하는 데이터베이스를 더 포함할 수 있다.
 - [14] 사용자로부터의 요청에 따라 상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 상기 사용자에게 제공하는 자료 제공부를 더 포함할 수 있다.
 - [15] 상기 문서는 영어로 작성될 수 있다.
 - [16] 본 발명의 다른 태양에 따른 어학 교육 시스템은 상기한 장치 중 어느 하나를 포함한다.
 - [17] 본 발명의 다른 태양에 따른 문서 처리 방법은, 웹으로부터 문서를 수집하고 상기 문서를 스크립트 부분과 설명 부분으로 분리하여 스크립트 문서 및 설명 문서를 생성하는 단계, 그리고 상기 설명 문서에 기초하여 표현 설명 문장을 추출하고 상기 표현 설명 문장이 상기 스크립트 문서 내에서 설명하고 있는 표현을 추출하는 단계를 포함한다.
 - [18] 상기 스크립트 문서에서 문장을 분리하여 복수의 스크립트 문장을 추출하고 상기 설명 문서에서 문장을 분리하여 복수의 설명 문장을 추출하는 단계를 더 포함할 수 있다.
 - [19] 상기 문장 추출 단계는 각 설명 문장이 상기 복수의 스크립트 문장 중 어느 스크립트 문장에 대한 설명인지 분류하는 단계를 포함할 수 있다.
 - [20] 상기 분류 단계는 조건부 랜덤 필드(conditional random field) 분류기를 사용하는 단계를 포함할 수 있다.
 - [21] 상기 표현 추출 단계는 레이블드 순차 패턴(labeled sequential pattern) 기법을 이용하여 상기 복수의 설명 문장으로부터 상기 표현 설명 문장을 추출하는 단계를 포함할 수 있다.
 - [22] 상기 표현 추출 단계는 상기 표현 설명 문장과 상기 스크립트 문장 내의 표현의 유사도를 측정하여 상기 표현 설명 문장이 상기 스크립트 문장 내에서 설명하고

있는 표현을 추출하는 단계를 포함할 수 있다.

[23] 상기 유사도는 코사인 유사도(cosine similarity) 기법이나 편집 거리(edit distance) 기법을 사용하여 측정될 수 있다.

[24] 상기 문서 생성 단계는 상기 스크립트 부분 및 상기 설명 부분의 시작 또는 끝을 나타내는 식별자 또는 문장에 기반하여 상기 스크립트 문서 및 상기 설명 문서를 생성할 수 있다.

[25] 상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 저장하는 단계를 더 포함할 수 있다.

[26] 사용자로부터의 요청에 따라 상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 상기 사용자에게 제공하는 단계를 더 포함할 수 있다.

[27] 본 발명의 다른 태양에 따른 어학 교육 방법은 상기한 방법 중 어느 하나를 포함한다.

[28] 본 발명의 다른 태양에 따른 컴퓨터로 읽을 수 있는 매체는 상기한 방법 중 어느 하나를 컴퓨터에 실행시키기 위한 프로그램을 기록한다.

발명의 효과

[29] 이와 같이 본 발명에 의하면, 학습 표현과 이에 대한 설명 부분이 하나의 쌍으로 이루어진 학습 자료를 자동으로 구축할 수 있다. 또한 실제 문장 내에서 사용되는 표현을 학습 대상으로 함으로써 사전적인 의미만을 제시하는 사전 검색 결과에 비하여 표현의 쓰임새를 비롯한 자세한 설명을 제공할 수 있고 설명 또한 학습 표현과 동일한 언어로 되어 있으므로 번역으로 인한 의미 훼손 없이 표현의 의미를 정확히 제공할 수 있다.

도면의 간단한 설명

[30] 도 1은 본 발명의 실시예에 따른 문서 처리 장치를 설명하기 위한 블록도이다.

[31] 도 2는 본 발명의 실시예에 따른 문서 처리 장치가 수집한 문서를 예시한 도면이다.

[32] 도 3은 본 발명의 실시예에 따른 문서 처리 방법을 설명하기 위한 흐름도이다.

발명의 실시를 위한 최선의 형태

[33] 그러면 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다.

[34] 먼저, 도 1 및 도 2를 참고하여 본 발명의 실시예에 따른 문서 처리 장치에 대하여 상세하게 설명한다.

[35] 도 1은 본 발명의 실시예에 따른 문서 처리 장치를 설명하기 위한 블록도이고, 도 2는 본 발명의 실시예에 따른 문서 처리 장치가 수집한 문서를 예시한 도면이다.

[36] 도 1을 참고하면, 본 발명의 실시예에 따른 문서 처리 장치(100)는 통신망(400)을 통하여 웹 서버(200) 및 사용자 단말기(300)에 연결되어 있다.

문서 처리 장치(100)는 웹 서버(200)에서 제공하는 웹 페이지를 통해 수집한 문서로부터 학습 표현과 이에 대한 설명을 자동으로 추출하여 저장한다. 또한 문서 처리 장치(100)는 사용자 단말기(300)로부터 질의를 받고 질의에 대응하는 학습 표현과 이에 대한 설명으로 이루어진 학습 자료를 사용자 단말기(300)에 제공한다.

- [37] 웹 서버(200)는 통신망(400)을 통하여 사용자 단말기(300) 및 문서 처리 장치(100)로부터의 요청에 따라 웹 페이지를 제공한다. 웹 서버(200)는 어학 교육 사이트를 통하여 웹 페이지를 제공할 수 있으며, 문서 처리 장치(100)는 웹 페이지에 포함되어 있는 다양한 문서 중에서 소정 문서를 추출할 수 있다. 예를 들면 웹 사이트 "www.eslpod.com"에서는 영어 스크립트(script)와 이에 대한 설명(description)을 포함하는 문서를 제공한다. 이하 문서 처리 장치(100)가 이 사이트에서 추출한 소정 문서를 가지고 작업을 수행하는 것으로 설명한다. 물론 이 사이트에 한정되지 않으며, 문서 처리 장치(100)는 작업에 필요한 문서를 다른 사이트에서 추출할 수도 있다.
- [38] 문서의 한 예로서, 도 2를 참고하면, "www.eslpod.com" 사이트에서 제공하는 문서(DC)는 스크립트 부분(SP)과 설명 부분(DP)를 포함하여 이루어진다. 문서(DC)마다 스크립트 부분(SP)에는 다양한 이야기가 전개되며 설명 부분(DP)에는 해당 스크립트 부분(SP)에 있는 표현에 대한 의미나 용어가 설명된다. 여기서 문서(DC)는 영어로 작성되어 있는 것으로 설명하나 이에 한정되지 않으며 한국어, 일어, 독어, 중국어 등 다른 언어라도 무관하다. 그리고 표현이란 적어도 하나의 의미를 전달하는 단위로서, 형태소, 단어나 숙어, 이들의 조합으로 이루어진다. 앞으로 설명의 편의를 위하여 표현도 영어로 작성된 영어 표현에 대하여 설명하나 이에 한정되지 않는다. 또한 스크립트 부분(SP)과 설명 부분(DP) 모두 동일한 언어로 작성되어 있는 것이 바람직하나 다르더라도 무관하다.
- [39] 사용자 단말기(300)는 사용자가 웹 서비스를 받기 위해 사용하는 통신 단말 장치로서, 통신망(400)을 통해 웹 서버(200), 문서 처리 장치(100) 등에 접속하여 정보를 주고받는다. 사용자 단말기(300)는 데스크톱 컴퓨터뿐만 아니라 노트북 컴퓨터, 워크스테이션, 팜톱(palmtop) 컴퓨터, UMPC(ultra mobile personal computer), 태블릿 PC, 개인 휴대 정보 단말기(personal digital assistant, PDA), 웹 패드, 휴대전화 등과 같이 메모리 수단을 구비하고 마이크로 프로세서를 탑재하여 연산 능력을 갖춘 단말기로 이루어질 수 있다.
- [40] 통신망(400)은 구내 정보 통신망(local area network, LAN), 도시권 통신망(metropolitan area network, MAN), 광역 통신망(wide area network, WAN), 인터넷 등을 포함하는 데이터 통신망뿐만 아니라 전화망 등을 포함할 수 있고, 유선과 무선은 가리지 않으며, 어떠한 통신 방식을 사용하더라도 상관없다.
- [41] 그러면 본 발명의 실시예에 따른 문서 처리 장치(100)에 대하여 좀 더 상세하게 설명한다. 도 1을 참고하면 문서 처리 장치(100)는 문서 수집부(110), 문장

추출부(130), 표현 추출부(150), 데이터베이스(170), 그리고 자료 제공부(190)를 포함한다.

- [42] 문서 수집부(110)는 웹 서버(200)에서 제공하는 웹 사이트에서 문서를 수집한다. 앞서 설명한 것처럼, 문서는 "www.eslpod.com"와 같이 미리 정해져 있는 웹 사이트에서 추출할 수 있으며 스크립트와 이에 대한 설명을 포함한다. 여기서 스크립트는 특정 주제에 대한 짧은 이야기 글을 의미하며, 서술문이나 대화문을 가리지 않는다.
- [43] 또한 문서 수집부(110)는 스크립트 부분과 설명 부분의 시작 또는 끝을 나타내는 식별자나 문장에 기반하여 수집된 문서에서 스크립트 부분과 설명 부분을 추출하여 스크립트 문서와 설명 문서를 각각 생성한다. 예를 들면 도 2에 도시한 것처럼, <start of story>라는 식별자가 스크립트 부분(SP)의 시작을 나타내고, <end of story>라는 식별자가 스크립트 부분(SP)의 끝과 설명 부분(DP)의 시작을 나타낼 수 있다. 이 외에도 식별자는 예를 들면 '*'나 '※' 등의 특정 기호나 'start'나 'end' 등의 단어와 같이 다양한 형태로 존재할 수 있다. 또한 "Let's go!"나 "Thank you"와 같은 문장 형식으로도 스크립트 부분과 설명 부분을 구별하도록 할 수 있다. 문서는 이와 같이 특정 형식을 가지고 있어서 스크립트 부분과 설명 부분을 쉽게 추출할 수 있도록 하는 것이 바람직하나 이에 한정되지 않는다.
- [44] 문장 추출부(130)는 마침표, 물음표, 느낌표 등의 문장의 끝을 나타내는 기호 등의 식별자를 기초로 스크립트 문서에서 각각의 스크립트 문장을 분리하고 설명 문서에서 각각의 설명 문장을 분리한다. 그런 후 문장 추출부(130)는 각 설명 문장이 어떤 스크립트 문장에 대한 설명인지 분류한다. 설명 문장의 분류를 위하여 예를 들면 조건부 랜덤 필드(conditional random field, CRF) 분류기를 사용할 수 있으나 이에 한정되지 않는다. 조건부 랜덤 필드 분류기는 자연 언어 텍스트와 같은 순차 데이터의 레이블링(labeling)이나 파싱(parsing)에 사용되는 일종의 식별 확률 모델(discriminative probabilistic model) 분류기이다.
- [45] 이때 분류 결과는 임의의 i 번째 스크립트 문장에 대하여 각 설명 문장이 i 번째 스크립트 문장에 대한 설명인지 아닌지 이진 분류로 나타난다. 좀 더 상세하게 설명하면, 하나의 스크립트 문서에 포함되어 있는 스크립트 문장의 수효를 M 이라 하면, i 번째 스크립트 문장에 대하여 설명 문서의 각 설명 문장이 분류기의 입력이 되고 설명 문서의 모든 설명 문장에 대하여 이진 분류 결과가 출력된다. 이와 같은 분류 과정을 모든 스크립트 문장($1 \leq i \leq M$)에 대하여 수행하면 각 스크립트 문장에 대응하는 설명 문장을 추출할 수 있으며, 따라서 설명 문서의 각 문장이 스크립트 문서의 어떤 문장에 대한 설명인지 분류할 수 있다. 앞으로 어떤 스크립트 문장에 대응하는 설명 문장들을 해당 스크립트 문장에 대한 문장 설명이라 한다. 물론 스크립트 문장에 대응하는 설명 문장이 존재하지 않을 수도 있으며, 설명 문장 중에서 스크립트 문장에 대응하지 않는 설명 문장도 존재할 수 있다.

- [46] 표현 추출부(150)는 문장 설명으로부터 표현 설명 문장을 추출하고 추출된 표현 설명 문장이 해당 스크립트 문장 내에서 어떤 표현을 설명하고 있는지 알아내고 추출해 낸다.
- [47] 우선, 표현 추출부(150)는 추출된 전체 문장 설명의 모든 문장에 대하여 각 문장이 영어 표현의 설명 형태를 가지고 있는지 그렇지 않은지에 대한 이진 분류를 수행한다. 이를 위하여 예를 들면 레이블드 순차 패턴(labeled sequential pattern, LSP) 기법을 이용할 수 있으나 이에 한정되지 않는다. LSP 기법은 어떤 문장이 주어졌을 때 이를 패턴화하여 표현하기 위한 방법으로서, 예를 들어 "Where can I find a job?"이라는 문장을 "Where can PRN VB DT NN?"으로 치환하는 것처럼 일반적인 단어를 좀 더 큰 범위의 클래스로 치환하는 방법을 말한다. 여기서 PRN은 대명사, VB는 동사, DT는 관사, 그리고 NN은 명사를 나타낸다.
- [48] 영어 표현의 설명 형태를 나타낼 수 있는 패턴은 스크립트 부분과 설명 부분을 포함하는 문서로부터 훈련될 수 있다. 예를 들면 영어 표현을 설명할 때 많이 나타나는 표현인 "... means to ..." 또는 "... is a ..." 등이 패턴으로 훈련된다. 이 패턴에 따라 전체 문장 설명의 각 문장들에 대하여 이진 분류를 수행하여 영어 표현을 나타내는 표현 설명 문장들을 추출해 낼 수 있다.
- [49] 다음으로, 표현 추출부(150)는 추출된 표현 설명 문장이 설명하고 있는 해당 스크립트 문장 내의 영어 표현을 추출한다. 스크립트 문장 내의 표현과 표현 설명 문장의 유사도가 클수록 표현 설명 문장이 해당 영어 표현에 대한 설명이 될 확률이 높다. 따라서 표현 추출부(150)는 스크립트 문장 내의 표현과 표현 설명 문장의 유사도에 기반하여 표현 설명 문장이 설명하고 있는 영어 표현을 추출한다. 유사도는 예를 들면 텍스트 마이닝(text mining)에서 문서들을 비교하는 데 사용되며 텍스트 사이의 유사도를 측정할 수 있는 코사인 유사도(cosine similarity) 기법이나 문자열 사이의 거리를 측정할 수 있는 편집 거리(edit distance) 기법을 사용할 수 있으나 이에 한정되지 않는다.
- [50] 표현 추출부(150)는 이러한 방법을 통하여 전체 표현 설명 문장에 대하여 해당 영어 표현을 추출하고, 추출된 영어 표현과 이를 설명하는 표현 설명 문장을 데이터베이스(170)에 저장한다. 이때 표현 추출부(150)는 영어 표현과 이를 설명하는 표현 설명 문장을 쌍(이하 '영어 표현과 이를 설명하는 표현 설명 문장 쌍'을 '표현-설명 쌍'이라 함)으로 하여 데이터베이스(170)에 저장할 수 있다.
- [51] 데이터베이스(170)는 표현 추출부(150)에서 제공받은 표현-설명 쌍을 저장하고 자료 제공부(190)의 요청에 따라 표현-설명 쌍을 제공한다.
- [52] 자료 제공부(190)는 데이터베이스(170)를 용이하게 검색할 수 있도록 표현-설명 쌍에 대한 인덱싱 작업을 수행한다. 또한 자료 제공부(190)는 사용자 단말기(300)로부터 질의어를 입력 받으면 분석을 통하여 질의어에 가장 적합한 표현을 추출하는 등의 전처리 과정을 거친 후 해당 표현이 포함된 표현-설명 쌍을 데이터베이스(170)에서 검색하고 검색된 표현-설명 쌍을 사용자

단말기(300)에 제공한다.

- [53] 이와 같이, 본 발명의 실시예에 따른 문서 처리 장치(100)에 의하면 영어 표현과 이에 대한 설명 부분이 하나의 쌍으로 이루어진 학습 자료를 웹으로부터 수집하여 자동으로 구축할 수 있다. 이러한 자료는 단어 및 구문 사전의 자동 구축에 사용될 수 있으며 외국어 교육을 위한 소프트웨어 전반에서 교육 자료로서 활용될 수 있다.
- [54] 또한 본 발명의 실시예에 따른 문서 처리 장치(100)는 표현 및 설명 데이터베이스를 통하여 실제 사용자가 이해하지 못하거나 친숙하지 않은 영어 표현에 대해 설명을 요구할 때 적절한 영어 표현 및 설명 문장 쌍을 검색하여 제시할 수 있다. 결국, 웹 문서를 이용하여 영어 표현 및 설명을 제시함으로써 사전적인 의미만을 포함하는 한글로 된 설명의 영어 사전 검색 결과에 비해 실제 스크립트 내에서의 표현 쓰임새를 비롯하여 자세한 설명을 제공할 수 있고 설명 문장 또한 영어로 되어 있으므로 원래의 의미를 번역으로 인한 훼손 없이 제공할 수 있다.
- [55] 본 발명의 실시예에 따른 문서 처리 장치(100)는 어학 교육을 위한 표현-설명 쌍을 수집하는 데 사용될 수 있어서 어학 교육을 위한 다양한 시스템(도시하지 않음)에 포함될 수도 있다. 이러한 시스템의 예로서, 온라인 사전 시스템이나 대화형 교육 시스템을 들 수 있으나 이에 한정되지 않는다.
- [56] 한편 본 발명의 실시예에 따른 문서 처리 장치(100)는 데이터베이스(170) 및 자료 제공부(190) 중 적어도 어느 하나를 포함하지 않게 구현될 수도 있으며, 이 경우 데이터베이스(170) 및/또는 자료 제공부(190)는 영어 대화 시스템 등의 외국어 교육 시스템에 포함될 수도 있다. 그리고 문서 처리 장치(100)는 독립적인 서버 형태로 구현될 수 있으나, 사용자 단말기(300)와 일체로 통합된 형태로 구현될 수도 있다. 즉, 문서 처리 장치(100)의 기능 전부가 사용자 단말기(300)에 구현되어 사용자 단말기(300)에서 웹 문서를 수집하고 이로부터 표현-설명 쌍을 추출할 수 있으며, 사용자 요청에 따라 표현-설명 쌍을 검색하여 사용자에게 제공할 수 있다. 물론 문서 처리 장치(100)의 기능 일부가 사용자 단말기(300)에 구현될 수도 있다.
- [57] 그러면, 도 3을 참고하여 본 발명의 실시예에 따른 문서 처리 방법에 대해 설명한다.
- [58] 도 3은 본 발명의 실시예에 따른 문서 처리 방법을 설명하기 위한 흐름도이다.
- [59] 먼저, 문서 처리 장치(100)는 웹 서버(200)에서 제공하는 웹 사이트에서 스크립트와 이에 대한 설명을 포함하는 문서를 수집한다(S510). 그런 후 문서 처리 장치(100)는 스크립트 부분과 설명 부분의 시작과 끝을 나타내는 식별자나 문장에 기반하여 수집된 문서에서 스크립트 부분과 설명 부분을 추출하여 스크립트 문서와 설명 문서를 각각 생성한다(S520).
- [60] 문서 처리 장치(100)는 마침표, 물음표, 느낌표 등의 문장의 끝을 나타내는 기호 등의 식별자를 기초로 스크립트 문서에서 각각의 스크립트 문장을 분리하고

설명 문서에서 각각의 설명 문장을 분리한다(S530). 그런 후 문서 처리 장치(100)는 CRF 분류기 등을 이용하여 각 설명 문장이 어떤 스크립트 문장에 대한 설명인지 분류한다(S540).

- [61] 이후 문서 처리 장치(100)는 LSP 기법 등을 이용하여 추출된 전체 문장 설명의 모든 문장에 대하여 각 문장이 영어 표현의 설명 형태를 가지고 있는지 그렇지 않은지에 대한 이진 분류를 수행하고 영어 표현의 설명 형태를 가지고 있는 표현 설명 문장을 추출한다(S550). 그리고 문서 처리 장치(100)는 추출된 표현 설명 문장들이 해당 스크립트 문장 내에서 설명하고 있는 영어 표현을 추출한다(S560). 이때 문서 처리 장치(100)는 코사인 유사도 기법이나 편집 거리 기법 등을 이용하여 추출된 표현 설명 문장과 해당 스크립트 문장 내의 표현의 유사도를 측정하여 해당 표현 설명 문장이 설명하는 영어 표현을 추출할 수 있다.
- [62] 그런 후 문서 처리 장치(100)는 추출된 영어 표현과 이를 설명하는 표현 설명 문장을 저장한다(S570). 이때, 문서 처리 장치(100)는 영어 표현과 이를 설명하는 표현 설명 문장을 쌍으로 하여 하나의 문서로 저장할 수 있다.
- [63] 문서 처리 장치(100)는 사용자 단말기(300)로부터 질의어를 입력 받아 질의어를 분석하고 질의어에 가장 적합한 표현을 추출하는 등의 전처리 과정을 수행하고 해당 표현이 포함된 문서를 데이터베이스(170)에서 검색하고 검색된 결과를 사용자 단말기(300)에 전송한다.
- [64] 본 발명의 실시예에 따른 문서 처리 방법은 외국어 교육을 위한 표현 및 그 설명 문장 쌍을 수집하는 데 사용될 수 있어서 외국어 교육을 위한 다양한 방법에 포함될 수 있다. 이러한 방법의 예로서, 온라인 사전 제공 방법이나 대화형 교육 방법을 들 수 있으나 이에 한정되지 않는다.
- [65] 본 발명의 실시예는 다양한 컴퓨터로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터로 읽을 수 있는 매체를 포함한다. 이 매체는 지금까지 설명한 문서 처리 방법을 실행시키기 위한 프로그램을 기록한다. 이 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이러한 매체의 예에는 하드디스크, 플로피디스크 및 자기 테이프와 같은 자기 매체, CD 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(Floptical Disk)와 자기-광 매체, 롬, 램, 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 구성된 하드웨어 장치 등이 있다. 또는 이러한 매체는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수 있다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [66] 이상에서 본 발명의 바람직한 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.

청구범위

- [청구항 1] 웹으로부터 문서를 수집하고 상기 문서를 스크립트 부분과 설명 부분으로 분리하여 스크립트 문서 및 설명 문서를 생성하는 문서 수집부, 그리고
상기 설명 문서에 기초하여 표현 설명 문장을 추출하고 상기 표현 설명 문장이 상기 스크립트 문서 내에서 설명하고 있는 표현을 추출하는 표현 추출부를 포함하는 문서 처리 장치.
- [청구항 2] 제1항에서,
상기 스크립트 문서에서 문장을 분리하여 복수의 스크립트 문장을 추출하고 상기 설명 문서에서 문장을 분리하여 복수의 설명 문장을 추출하는 문장 추출부를 더 포함하는 문서 처리 장치.
- [청구항 3] 제2항에서,
상기 문장 추출부는 각 설명 문장이 상기 복수의 스크립트 문장 중 어느 스크립트 문장에 대한 설명인지 분류하는 문서 처리 장치.
- [청구항 4] 제3항에서,
상기 문장 추출부는 조건부 랜덤 필드(conditional random field) 분류기를 사용하여 상기 분류를 수행하는 문서 처리 장치.
- [청구항 5] 제2항에서,
상기 표현 추출부는 레이블드 순차 패턴(labeled sequential pattern) 기법을 이용하여 상기 복수의 설명 문장으로부터 상기 표현 설명 문장을 추출하는 문서 처리 장치.
- [청구항 6] 제2항에서,
상기 표현 추출부는 상기 표현 설명 문장과 상기 스크립트 문장 내의 표현의 유사도를 측정하여 상기 표현 설명 문장이 상기 스크립트 문장 내에서 설명하고 있는 표현을 추출하는 문서 처리 장치.
- [청구항 7] 제6항에서,
상기 유사도는 코사인 유사도(cosine similarity) 기법이나 편집 거리(edit distance) 기법을 사용하여 측정되는 문서 처리 장치.
- [청구항 8] 제1항에서,
상기 문서 수집부는 상기 스크립트 부분 및 상기 설명 부분의 시작 또는 끝을 나타내는 식별자 또는 문장에 기반하여 상기 스크립트 문서 및 상기 설명 문서를 생성하는 문서 처리 장치.
- [청구항 9] 제1항에서,
상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 저장하는 데이터베이스를 더 포함하는 문서 처리 장치.
- [청구항 10] 제1항에서,

사용자로부터의 요청에 따라 상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 상기 사용자에게 제공하는 자료 제공부를 더 포함하는 문서 처리 장치.

[청구항 11]

제1항에서,

상기 문서는 영어로 작성되어 있는 문서 처리 장치.

[청구항 12]

제1항 내지 제11항 중 어느 한 항의 문서 처리 장치를 포함하는 어학 교육 시스템.

[청구항 13]

웹으로부터 문서를 수집하고 상기 문서를 스크립트 부분과 설명 부분으로 분리하여 스크립트 문서 및 설명 문서를 생성하는 단계, 그리고

상기 설명 문서에 기초하여 표현 설명 문장을 추출하고 상기 표현 설명 문장이 상기 스크립트 문서 내에서 설명하고 있는 표현을 추출하는 단계

를 포함하는 문서 처리 방법.

[청구항 14]

제13항에서,

상기 스크립트 문서에서 문장을 분리하여 복수의 스크립트 문장을 추출하고 상기 설명 문서에서 문장을 분리하여 복수의 설명 문장을 추출하는 단계를 더 포함하는 문서 처리 방법.

[청구항 15]

제14항에서,

상기 문장 추출 단계는 각 설명 문장이 상기 복수의 스크립트 문장 중 어느 스크립트 문장에 대한 설명인지 분류하는 단계를 포함하는 문서 처리 방법.

[청구항 16]

제15항에서,

상기 분류 단계는 조건부 랜덤 필드(conditional random field) 분류기를 사용하는 단계를 포함하는 문서 처리 방법.

[청구항 17]

제14항에서,

상기 표현 추출 단계는 레이블드 순차 패턴(labeled sequential pattern) 기법을 이용하여 상기 복수의 설명 문장으로부터 상기 표현 설명 문장을 추출하는 단계를 포함하는 문서 처리 방법.

[청구항 18]

제14항에서,

상기 표현 추출 단계는 상기 표현 설명 문장과 상기 스크립트 문장 내의 표현의 유사도를 측정하여 상기 표현 설명 문장이 상기 스크립트 문장 내에서 설명하고 있는 표현을 추출하는 단계를 포함하는 문서 처리 방법.

[청구항 19]

제18항에서,

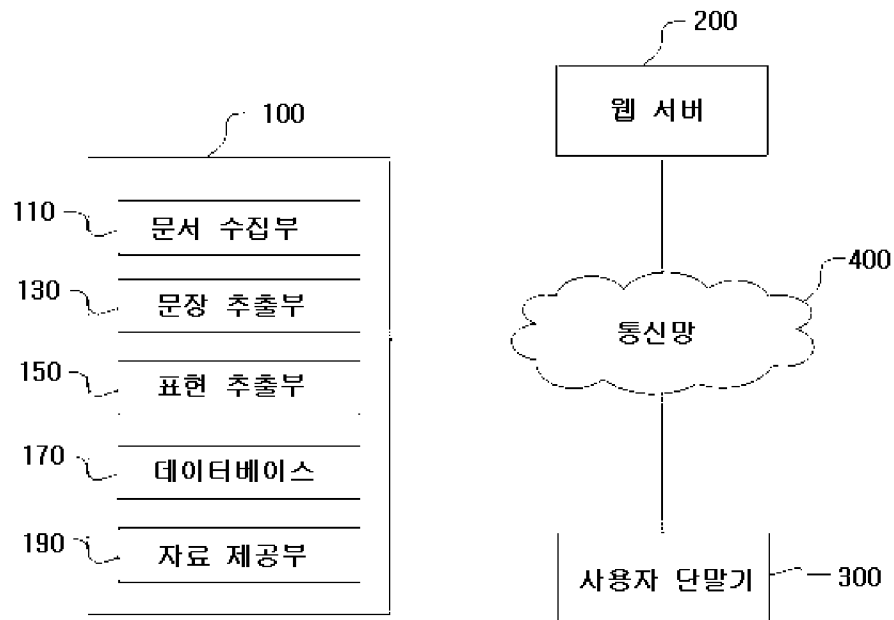
상기 유사도는 코사인 유사도(cosine similarity) 기법이나 편집 거리(edit distance) 기법을 사용하여 측정되는 문서 처리 방법.

[청구항 20]

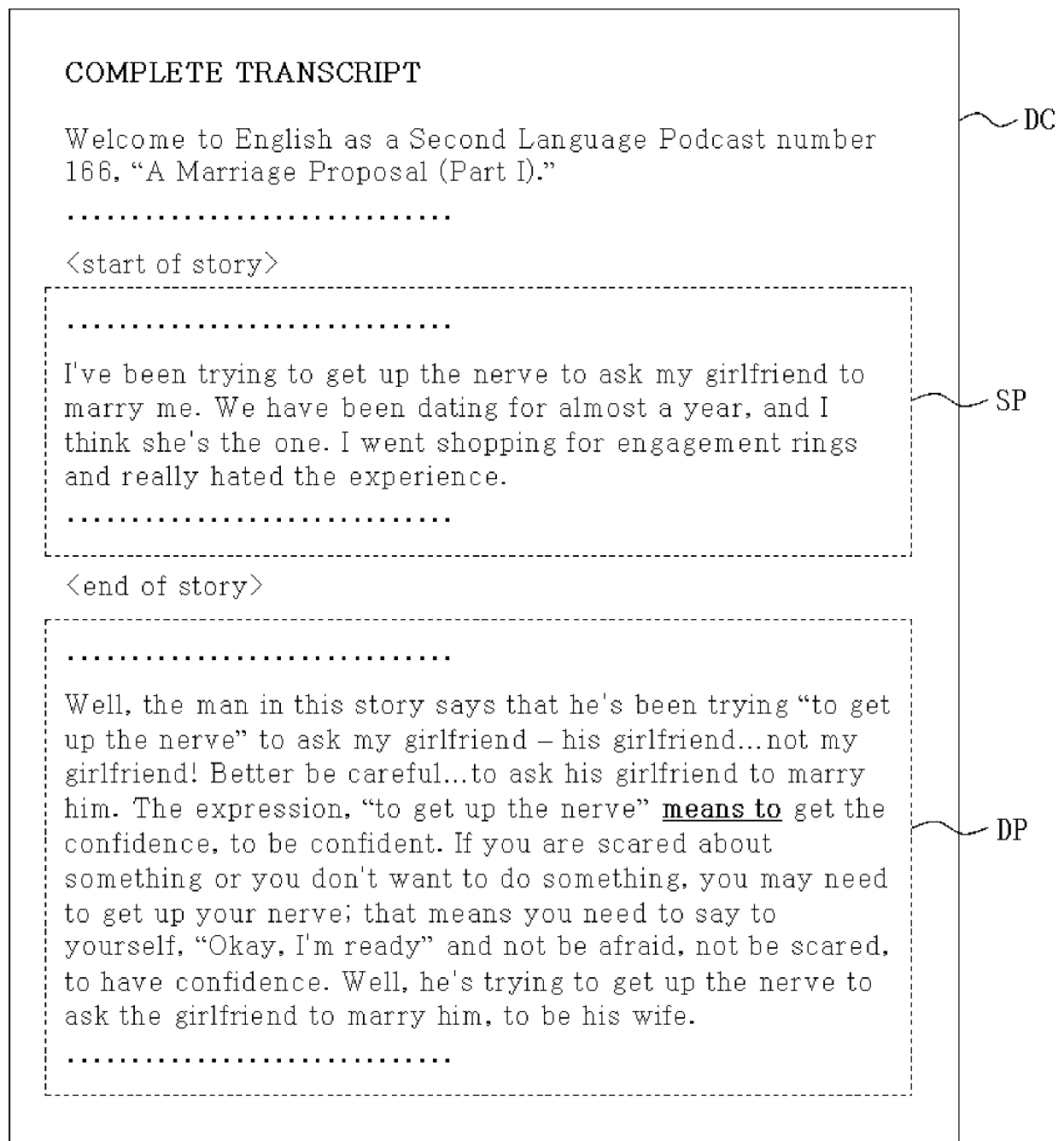
제13항에서,

- 상기 문서 생성 단계는 상기 스크립트 부분 및 상기 설명 부분의 시작 또는 끝을 나타내는 식별자 또는 문장에 기반하여 상기 스크립트 문서 및 상기 설명 문서를 생성하는 문서 처리 방법.
- [청구항 21] 제13항에서,
상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 저장하는 단계를 더 포함하는 문서 처리 방법.
- [청구항 22] 제13항에서,
사용자로부터의 요청에 따라 상기 추출된 표현과 상기 추출된 표현을 설명하는 표현 설명 문장을 상기 사용자에게 제공하는 단계를 더 포함하는 문서 처리 방법.
- [청구항 23] 제13항에서,
상기 문서는 영어로 작성되어 있는 문서 처리 방법.
- [청구항 24] 제13항 내지 제23항 중 어느 한 항의 문서 처리 방법을 포함하는 어학 교육 방법.
- [청구항 25] 컴퓨터에 제13항 내지 제23항 중 어느 한 항의 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 매체.

[Fig. 1]



[Fig. 2]



[Fig. 3]

