(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0057983
(43) 공개일자 2021년05월24일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01) *G06N 20/00* (2019.01)
(52) CPC특허분류
G06F 40/205 (2020.01)
G06F 40/268 (2020.01)
(21) 출원번호 10-2019-0144846
(22) 출원일자 2019년11월13일
심사청구일자 2019년11월13일

(71) 출원인
창원대학교 산학협력단
경상남도 창원시 의창구 창원대학로 20(퇴촌동)
(72) 발명자
차정원
경상남도 창원시 성산구 동산로 115, 105동 1404호 (상남동, 대동아파트)
김정무
경상남도 창원시 성산구 대암로 22, 103동 1207호 (남양동, 우성아파트)
(74) 대리인
김종선, 이형석

전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **포인트 네트워크를 이용한 트리플 추출방법 및 그 추출장치**

(57) 요 약

본 발명은 포인터 네트워크에 강화학습을 적용하여 트리플을 형성하는 단어 추출을 위한 추출방법 및 장치에 관한 것이다. 본 발명은 입력된 학습 코퍼스의 문장의 분류하는 문장 분류부와, 상기 분류된 문장을 형태소 단위로 구분하는 형태소 분석부와, 상기 형태소를 기초로 하여 용언의 앞 단어를 이용하여 D_0을 형성하는 D_0 형성부, 및 상기 D_0를 활용한 포인터 네트워크를 사용하여 D_0와 관계있는 단어를 추출하는 트리플 추출부를 포함하여 구성된다.

대 표 도 - 도1



(52) CPC특허분류
G06F 40/284 (2020.01)
G06N 20/00 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	K-19-IR-13-04S-1
부처명	한국과학기술정보연구원(KISTI)
과제관리(전문)기관명	한국과학기술정보연구원(KISTI)
연구사업명	한국과학기술정보연구원 주요사업
연구과제명	기계학습 기반 이슈 이벤트 추출 및 탐지 기술 개발
기 여 율	1/1
과제수행기관명	한국과학기술정보연구원(KISTI)
연구기간	2019.04.01 ~ 2019.10.31

명 세 서

청구범위

청구항 1

문장 단위의 학습 코퍼스를 입력받고 문장을 분류하는 제1 단계;

상기 분류한 문장을 형태소 단위로 구분하는 제2 단계;

상기 형태소 단위에서 용언(VV, VA, XSV, XSA)의 앞 단어를 이용하여 D_0를 형성하는 제3 단계;

상기 D_0를 포인터 네트워크의 디코더의 입력으로 사용하는 제4 단계; 및

상기 입력된 D_0의 정보를 기초로 상기 D_0와 관련있는 단어에 대한 주의 집중(attention)을 수행하여 학습 코퍼스에서 트리플 단어를 추출하는 제5 단계를 포함하는 포인트 네트워크를 이용한 트리플 추출방법.

청구항 2

제 1 항에 있어서,

다음 학습을 위하여 상기 추출된 학습 관계를 대상으로 보상(reward)을 부여하여 학습모델을 업데이트 하는 제6 단계를 더 포함하는 포인트 네트워크를 이용한 트리플 추출방법.

청구항 3

제 2 항에 있어서,

상기 D_0는 용언과 결합한 단어인 포인트 네트워크를 이용한 트리플 추출방법.

청구항 4

제 2 항에 있어서,

상기 트리플 단어는, 학습문장에 D_0가 포함된 문장을 사용하고, 한글 문장에는 주어 및 목적어가 포함된 문장이며, 상기 용언과 특정 거리 내의 단어를 추출하는 포인트 네트워크를 이용한 트리플 추출방법.

청구항 5

제 4 항에 있어서,

상기 트리플 단어는, 장소나 시간에 해당하는 단어이고, 부사인 단어는 포함하지 않고, 품사에 따라 상이한 점수를 부여하여 보상하는 포인트 네트워크를 이용한 트리플 추출방법.

청구항 6

제 4 항에 있어서,

상기 트리플 단어는, 학습 코퍼스의 문장 중 상기 D_0와 관련있는 단어들에 대해 가중치를 부여하여 가장 높은 가중치를 가진 단어를 추출하는 포인트 네트워크를 이용한 트리플 추출방법.

청구항 7

입력된 학습 코퍼스의 문장의 분류하는 문장 분류부;

상기 분류된 문장을 형태소 단위로 구분하는 형태소 분석부;

상기 형태소를 기초로 하여 용언의 앞 단어를 이용하여 D_0을 형성하는 D_0 형성부; 및

상기 D_0를 활용한 포인터 네트워크를 사용하여 D_0와 관계있는 단어를 추출하는 트리플 추출부를 포함하는 포인트

네트워크를 이용한 트리플 추출장치.

청구항 8

제 7 항에 있어서,

상기 형태소 분석부는 'Espresso' 분석기인 포인트 네트워크를 이용한 트리플 추출장치.

청구항 9

제 7 항에 있어서,

상기 트리플 추출부는,

학습문장에 D_0가 포함된 문장을 사용하고, 한글 문장에는 주어 및 목적어가 포함된 문장이며, 상기 용언과 특정 거리 내의 단어를 추출하는 포인트 네트워크를 이용한 트리플 추출장치.

청구항 10

제 9 항에 있어서,

상기 트리플 추출부는,

학습 코퍼스의 문장 중 상기 D_0와 관련있는 단어들에 대해 가중치를 부여하여 가장 높은 가중치를 가진 단어를 추출하는 포인트 네트워크를 이용한 트리플 추출장치.

청구항 11

제 7 항에 있어서,

상기 추출된 단어를 대상으로 보상을 부여하여 새로운 관계를 추출하도록 학습모델을 업데이트하는 업데이트부를 더 포함하는 포인트 네트워크를 이용한 트리플 추출장치.

청구항 12

제 11 항에 있어서,

상기 업데이트부는,

상기 추출된 단어 품사에 따라 서로 상이한 점수를 부여하는 포인트 네트워크를 이용한 트리플 추출장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 용언과 관련된 단어 추출에 관한 것으로, 특히 포인터 네트워크에 강화학습을 적용하여 트리플을 형성하는 단어 추출을 위한 추출방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 현재 인터넷의 대부분을 차지하고 있는 정보들은 비구조 자연어로 표현되어 있으며, 이와 같이 자연어로 표현된 문서들은 일반 사용자들에게 매우 보편적인 표현법이다. 하지만, 사람이 아닌 컴퓨터가 이러한 자연어로 표현된 문서 즉, 비구조로 표현된 데이터를 활용하기에는 많은 어려움이 따른다.

[0003] 따라서, 최근에는 상술한 바와 같이 비구조 표현으로 구축된 정보를 구조화함으로써, 컴퓨터가 계산 가능한 형태로 정보를 구축하여 서비스를 제공하는 사례가 늘어나고 있다. 이렇게 구축된 정보를 활용하여 자연어 처리의 관계 추출문제를 수행한다. 이때 자연어 처리의 관계 추출 문제를 해결하기 위해서는 학습에 필요한 트리플 패턴의 코퍼스(corpus)가 필요하다. 그리고 정답 트리플 코퍼스는 문장이나 문서에서 적절한 관계를 지닌 단어들의 분석이 필요하다.

[0004] 그러나 정답 트리플 코퍼스는 사람이 직접 문장이나 문서를 분석하여 생성해야 하기 때문에 많은 시간과 비용이 소모된다는 문제가 있다.

발명의 내용

해결하려는 과제

[0005] 따라서 본 발명의 목적은 상기한 문제점을 해결하기 위한 것으로, 포인터 네트워크에 강화학습을 적용하여 트리플을 형성하는 단어를 자동으로 추출할 수 있는 포인트 네트워크를 이용한 이벤트 추출장치를 제공하는 것이다.

[0006] 즉 본 발명은 자연어 처리 중 다양한 문서 요약 분야에 적용될 수 있는 관계 추출 문제에 대해 정답 코퍼스 없이 학습하여 문서나 문장에서 용언과 관계된 관계단어를 추출하는 것이라 할 수 있다.

과제의 해결 수단

[0007] 이와 같은 목적을 달성하기 위한 본 발명은, 문장 단위의 학습 코퍼스를 입력받고 문장을 분류하는 제1 단계; 상기 분류한 문장을 형태소 단위로 구분하는 제2 단계; 상기 형태소 단위에서 용언(VV, VA, XSV, XSA)의 앞 단어를 이용하여 D_0를 형성하는 제3 단계; 상기 D_0를 포인터 네트워크의 디코더의 입력으로 사용하는 제4 단계; 및 상기 입력된 D_0의 정보를 기초로 상기 D_0와 관련있는 단어에 대한 주의 집중(attention)을 수행하여 학습 코퍼스에서 트리플 단어를 추출하는 제5 단계를 포함하는 포인트 네트워크를 이용한 트리플 추출방법을 제공한다.

[0008] 바람직하게, 다음 학습을 위하여 상기 추출된 학습 관계를 대상으로 보상(reward)을 부여하여 학습모델을 업데이트 하는 제6 단계를 더 포함한다.

[0009] 바람직하게, 상기 D_0는 용언과 결합한 단어를 의미한다.

[0010] 바람직하게, 상기 트리플 단어는, 학습문장에 D_0가 포함된 문장을 사용하고, 한글 문장에는 주어 및 목적어가 포함된 문장이며, 상기 용언과 특정 거리 내의 단어를 추출한다.

[0011] 바람직하게, 상기 트리플 단어는, 장소나 시간에 해당하는 단어이고, 부사인 단어는 포함하지 않고, 품사에 따라 상이한 점수를 부여하여 보상한다.

[0012] 바람직하게, 상기 트리플 단어는, 학습 코퍼스의 문장 중 상기 D_0와 관련있는 단어들에 대해 가중치를 부여하여 가장 높은 가중치를 가진 단어를 추출한다.

[0013] 본 발명의 다른 특징에 따르면, 입력된 학습 코퍼스의 문장의 분류하는 문장 분류부; 상기 분류된 문장을 형태소 단위로 구분하는 형태소 분석부; 상기 형태소를 기초로 하여 용언의 앞 단어를 이용하여 D_0을 형성하는 D_0 형성부; 및 상기 D_0를 활용한 포인터 네트워크를 사용하여 D_0와 관계있는 단어를 추출하는 트리플 추출부를 포함하는 포인트 네트워크를 이용한 트리플 추출장치를 제공한다.

[0014] 바람직하게, 상기 형태소 분석부는 'Espresso' 분석기이다.

[0015] 바람직하게, 상기 트리플 추출부는 학습문장에 D_0가 포함된 문장을 사용하고, 한글 문장에는 주어 및 목적어가 포함된 문장이며, 상기 용언과 특정 거리 내의 단어를 추출한다.

[0016] 바람직하게, 상기 트리플 추출부는 학습 코퍼스의 문장 중 상기 D_0와 관련있는 단어들에 대해 가중치를 부여하여 가장 높은 가중치를 가진 단어를 추출한다.

[0017] 바람직하게, 상기 추출된 단어를 대상으로 보상을 부여하여 새로운 관계를 추출하도록 학습모델을 업데이트하는 업데이트부를 더 포함한다.

[0018] 상기 업데이트부는, 상기 추출된 단어 품사에 따라 서로 상이한 점수를 부여한다.

발명의 효과

[0019] 이상과 같은 본 발명의 포인트 네트워크를 이용한 트리플 단어 추출방법 및 장치에 따르면, 자연어 처리 중 다양한 문서 요약 분야에 적용할 수 있는 관계 추출 분야에서, 포인터 네트워크(Pointer Network)에 강화 학습(Reinforcement Learning)을 적용하여 트리플을 형성하는 단어를 추출하고 있다.

[0020] 그리고 본 발명은 정답 코퍼스 없이 강화학습을 기반으로 학습하면서 특히 트리플 단어의 추출 효용성을 향상시

키기 위하여, 학습문장에 D_0가 포함된 문장을 사용하고, 한글 문장에는 주어 및 목적어가 포함된 문장 및 상기 용언과 특정 거리 내의 단어를 추출하고 있고, 주의 기법(Attention Mechanism)을 활용하여 학습 코퍼스의 문장 중 Do와 관련 있는 단어들에 대해 가중치를 부여하여 가장 높은 가중치를 가진 단어를 추출한다.

[0021] 따라서 본 발명은 정답 코퍼스 없이도 자연어 처리의 관계 추출 문제를 해결하기 필요한 트리플 코퍼스를 종래보다 시간과 비용을 절약하면서 추출할 수 있는 효과를 기대할 수 있다.

도면의 간단한 설명

[0022] 도 1은 본 발명의 바람직한 실시 예에 따른 포인트 네트워크를 이용한 트리플 추출장치의 블록 구성도

도 2는 본 발명의 바람직한 실시 예에 따른 포인트 네트워크를 이용한 트리플 추출방법을 보인 흐름도

도 3은 본 발명에 따른 문장 분류 및 형태소 분석 예를 나타낸 도면

도 4는 본 발명에 따른 트리플 추출방법의 실행 결과를 나타낸 도면

발명을 실시하기 위한 구체적인 내용

[0023] 본 발명의 목적 및 효과, 그리고 그것들을 달성하기 위한 기술적 구성들은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시 예들을 참조하면 명확해질 것이다. 본 발명을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.

[0024] 그리고 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다.

[0025] 그러나 본 발명은 이하에서 개시되는 실시 예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있다. 단지 본 실시 예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0026] 이하에서는 도면에 도시한 실시 예에 기초하면서 본 발명에 대하여 더욱 상세하게 설명하기로 한다.

[0027] 본 발명의 포인트 네트워크를 이용한 트리플 추출장치(100)를 보인 도 1을 참고하면, 트리플 단어 추출을 목적으로 하는 학습 코퍼스가 제공된다. 상기 학습 코퍼스는 언어 연구를 위해 텍스트(text)를 컴퓨터가 읽을 수 있는 형태로 모아놓은 언어 자료를 말하며, 본 실시 예는 이러한 학습 코퍼스의 예로 한글 뉴스 데이터를 이용한다. 그리고 한글 뉴스 데이터 중 관계를 구성하는 문장을 사용하기 위하여 인터넷 포털(portal)인 네이버(Naver)의 실시간 급상승 검색어를 활용한다.

[0028] 본 발명의 실시 예에는 한글 뉴스 데이터의 문장 중에서 급상승 검색어가 포함되어 있는 경우 해당 문장을 뉴스 데이터의 주요 문장으로 가정하여 분류하는 문장 분류부(110)가 구비된다. 즉 문장 분류부(110)는 학습 코퍼스를 입력으로 받는 것이고 이때 코퍼스는 문장단위로 입력된다.

[0029] 그리고 상기 문장 분류부(110)가 분류한 문장을 형태소 단위로 학습하기 위해 형태소 단위로 나누는 형태소 분석부(120)가 구비된다. 형태소 분석부는 'Espresso'를 활용할 것이다.

[0030] 또 상기 형태소 분석기(120)에는 D_0 형성부(130)가 연결된다. D_0 형성부(130)는 용언(VV, VA, XSV, XSA)의 앞 단어를 이용하여 D_0을 형성하는 것인데, 여기서 D_0는 용언과 결합된 단어를 의미한다.

[0031] 또 상기 D_0 형성부(130)가 형성한 D_0를 이용하여 그 D_0와 관련있는 단어를 추출하는 트리플 추출부(140)가 구비된다. 이때 트리플 추출부(140)는 포인터 네트워크의 디코더(decoder)에 상기 D_0를 입력하여 학습 코퍼스에 대해 D_0와 관련있는 단어를 추출하여 관계 코퍼스를 생성하는 역할을 하며, 관계를 형성하는 단어 추출은 다수의 관계로 단어 추출을 진행할 수 있을 것이다.

[0032] 그리고 본 발명은 정답 코퍼스를 사용하지 않기 때문에 트리플 추출부(140)의 신뢰성을 향상시킬 필요가 있다. 본 실시 예에는 학습 과정에서 3가지 가정을 통해 신뢰성을 향상시킬 것이다.

[0033] 한편, 본 발명에 따르면 다음 학습을 위해 상기 트리플 추출부(140)의 추출 결과에 대해 보상을 부여하여 새로

운 관계를 추출할 수 있도록 학습 모델을 업데이트하는 업데이트부(150)가 마련될 수 있다.

[0034] 다음에는 이와 같이 구성된 추출장치(100)를 이용한 트리플 단어의 추출방법에 대하여 도 2를 함께 참조하여 설명한다.

[0035] 먼저, 문장 분류부(110)는 학습 코퍼스를 입력받아 문장을 분류하게 된다(s100). 학습 코퍼스는 한글 뉴스 데이터이고, 이러한 한글 뉴스 데이터 중에서 관계를 구성하는 문장을 사용하기 위해 주요 인터넷 포털인 네이버의 실시간 급상승 검색어를 활용한다.

[0036] 이와 같이 분류된 문장은 형태소 분석부(120)로 입력된다. 그러면 형태소 분석부(120)는 상기 문장을 형태소 단위로 학습하기 위하여 형태소 단위로 나눈다(s110). 형태소 분석부(120)의 예로 본 실시 예는 'Espresso'를 이용한다. 그러나 한글 문장을 형태소로 나눌 수 있는 다른 어떠한 분석수단이 사용될 수 있음은 당연하다 할 것이다.

[0037] 이러한 급상승 검색어를 한글 뉴스 기사 문장으로 활용하고 문장 분류부(110)에 의한 문장 분류 및 형태소 단위로 나눈 실시 예를 도 3에 도시하였다.

[0038] 도 3을 보면, 실시간 급상승 검색어로, '지진', '아이폰 xs', 'ebsi 고등', '교보문고', '고속도로교통상황', '제주항공'이 검색되고 있고, 다음과 같은 한글 뉴스 기사문장인 "24호 태풍 '짜미'가 일본 열도를 관통 ~ 피해가 발생했다.", "문재인 대통령은 국군의 날이 1일 ~", "제25호 태풍 콩레이(Kong-Rey)의 ~ 기록했다." 에 급상승 검색어가 포함되어 있다면 문장 분류부 및 형태소 분석부는 도 3의 "뉴스 문장 분류 및 Espresso 실행" 란에 기재된 바와 같이 급상승 검색어를 이용하여 학습에 사용되는 문장을 분류한 후 형태소 분석부(10)의 'Espresso'를 실행한 결과를 제공한다.

[0039] 그리고 형태소 분석된 결과에 대하여 D_0 형성부(130)는 용언(VV, VA, XSV, XSA)의 앞 단어를 이용하여 용언과 결합된 단어인 D_0를 형성한다(s120).

[0040] 그러면 트리플 추출부(140)는, 상기 D_0를 포인터 네트워크의 디코더(Decoder)의 입력으로 사용하여(s130), 학습 코퍼스에 대해 Do와 관련 있는 단어를 추출한다(s140). 이때 단어 추출에는 포인트 네트워크가 사용되며, 포인트 네트워크는 주의 기법(attention mechanism)을 활용한다. 또한 트리플 추출부(140)는 학습 코퍼스의 문장 중 Do와 관련 있는 단어들에 대해 가중치를 부여하여 가장 높은 가중치를 가진 단어를 추출한다.

[0041] 이와 같이 상기 D_0을 형성하고, 학습 코퍼스 문장에서 관계 단어를 추출한 실행 결과는 도 4에 도시하였다.

[0042] 그리고 상기와 같이 트리플 단어를 추출한 다음에는, 업데이트(150)는 다음 학습을 위하여 상기 추출된 트리플 단어를 대상으로 보상(reward)을 부여하여 학습모델을 업데이트할 수 있다(s150).

[0043] 한편, 본 실시 예는 단어 추출시, 주어, 목적어를 비롯하여 장소나 시간에 해당하는 단어를 추출하며 부사인 단어는 추출 대상에 포함되지 않는다. 그리고 추출된 단어가 부사인 경우 0점을 부여하고, '은/는/이/가'의 주어 또는 '을/를'과 같은 목적어 등의 단어에는 1점을 부여하는 방식이다.

[0044] 또한, 본 실시 예는 학습문장에 D_0가 포함된 문장을 사용하고, 한글 문장에는 주어 및 목적어가 포함된 문장이며, 상기 용언과 특정 거리 내의 단어를 추출하여 트리플 추출부(140)의 신뢰성을 향상시킨다.

[0045] 이상과 같이 본 발명의 도시된 실시 예를 참고하여 설명하고 있으나, 이는 예시적인 것들에 불과하며, 본 발명이 속하는 기술 분야의 통상의 지식을 가진 자라면 본 발명의 요지 및 범위에 벗어나지 않으면서도 다양한 변형, 변경 및 균등한 타 실시 예들이 가능하다는 것을 명백하게 알 수 있을 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적인 사상에 의해 정해져야 할 것이다.

부호의 설명

[0046] 110: 문장 분류부

120: 형태소 분석부

130: D_0 형성부

140: 트리플 추출부

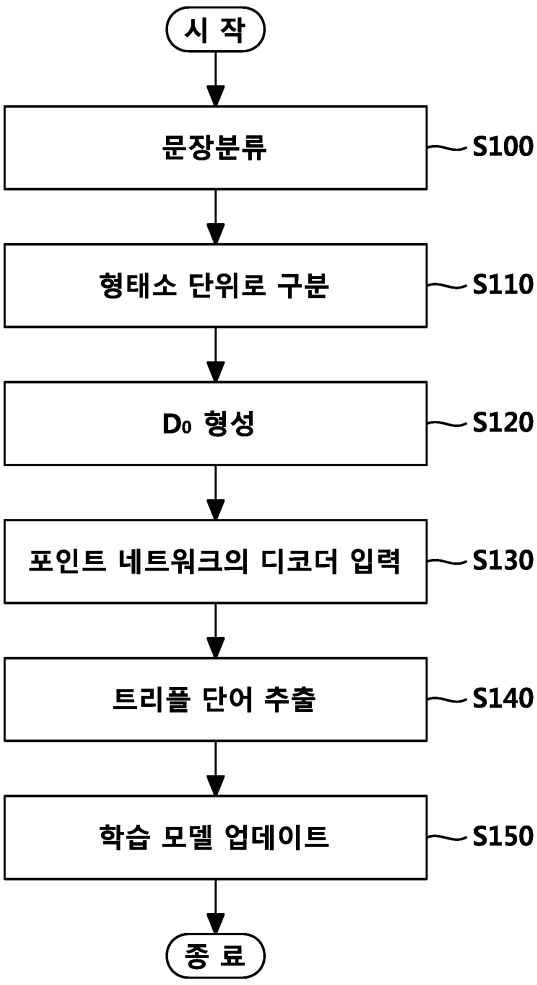
150: 업데이트부

도면

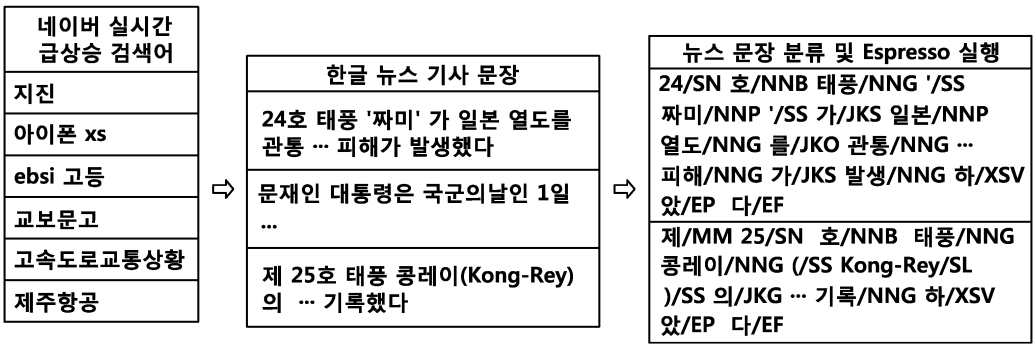
도면1



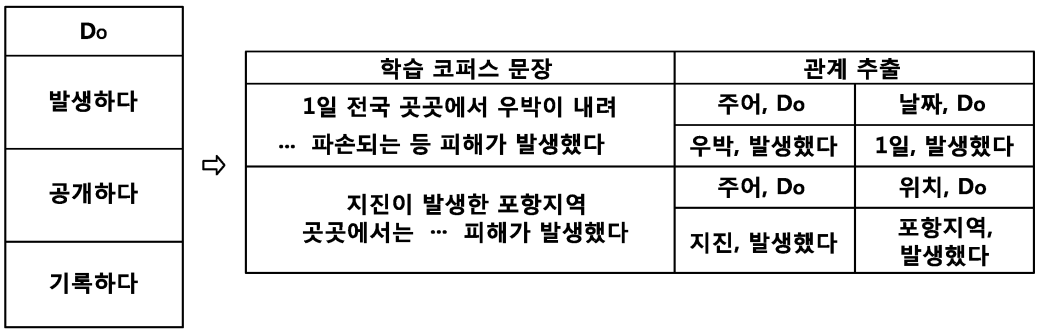
도면2



도면3


네이버 실시간 급상승 검색어
지진
아이폰 xs
ebsi 고등
교보문고
고속도로교통상황
제주항공
한글 뉴스 기사 문장
24호 태풍 '짜미' 가 일본 열도를 관통 … 피해가 발생했다
문재인 대통령은 국군의날인 1일 …
제 25호 태풍 콩레이(Kong-Rey) 의 … 기록했다
뉴스 문장 분류 및 Espresso 실행
24/SN 호/NNB 태풍/NNG '/SS 짜미/NNP '/SS 가/JKS 일본/NNP 열도/NNG 를/JKO 관통/NNG … 피해/NNG 가/JKS 발생/NNG 하/XSV 았/EP 다/EF
제/MM 25/SN 호/NNB 태풍/NNG 콩레이/NNG (/SS Kong-Rey/SL)/SS 의/JKG … 기록/NNG 하/XSV 았/EP 다/EF

도면4


Do
발생하다
공개하다
기록하다
학습 코퍼스 문장
1일 전국 곳곳에서 우박이 내려 … 파손되는 등 피해가 발생했다
지진이 발생한 포항지역 곳곳에서는 … 피해가 발생했다
관계 추출
주어, Do
날짜, Do
우박, 발생했다
1일, 발생했다
주어, Do
위치, Do
지진, 발생했다
포항지역, 발생했다