



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0169175
(43) 공개일자 2022년12월27일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) G06F 16/332 (2019.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06F 16/3329 (2019.01)
(21) 출원번호 10-2021-0079261
(22) 출원일자 2021년06월18일
심사청구일자 2021년06월18일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
박성배
서울특별시 송파구 오금로 396, 1402호 (가락동, 에스케이허브파크)
이준범
경기도 용인시 기흥구 서천동로21번길 6-1, 301호 (서천동)
(뒷면에 계속)
(74) 대리인
유미특허법인

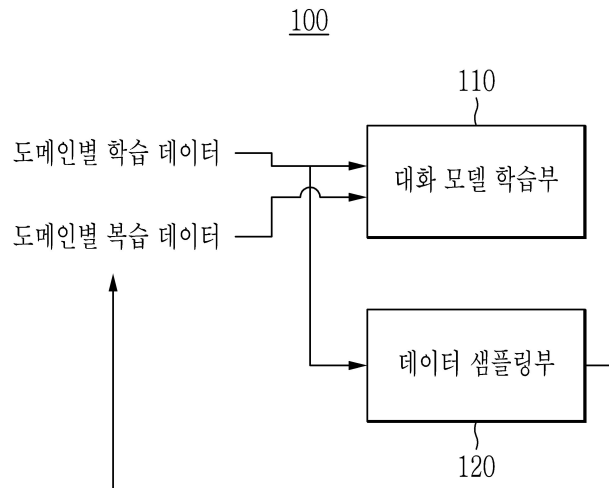
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 단일 도메인 대화 모델을 이용한 다중 도메인 대화 모델 생성 시스템 및 방법

(57) 요약

적어도 하나의 프로세서에 의해 동작하는 대화 모델 생성 시스템이 다중 도메인 대화 모델을 생성하는 방법으로, 첫 번째 도메인에서부터 n-1번째 도메인까지 학습이 순차적으로 완료된 단일 도메인 대화 모델을, n번째 도메인에 해당하는 복수의 학습 데이터들과, 첫 번째 도메인에서부터 n-1번째 도메인까지 추출한 복수의 복습 데이터들로 단일 도메인 대화 모델을 학습시키고, 학습 데이터들을 클러스터링하여 n번째 도메인에 대한 복수의 복습 데이터들을 추출한다.

대표도 - 도1



(72) 발명자

박형준

경기도 수원시 영통구 청명북로7번길 7-3, 301호
(영통동)

권성구

경기도 수원시 영통구 에듀타운로 101, 107동 403
호 (이의동)

김소연

경기도 용인시 기흥구 서천동로21번길 6-1, 301호
(서천동)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711103335

과제번호 2013-2-00109-008

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 혁신성장동력프로젝트(R&D)(과기정통부)

연구과제명 (엑소브레인-2세부) WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기
술 개발

기 여 율 1/1

과제수행기관명 주식회사 솔트룩스

연구기간 2020.01.01~2020.12.31

명세서

청구범위

청구항 1

적어도 하나의 프로세서에 의해 동작하는 대화 모델 생성 시스템이 다중 도메인 대화 모델을 생성하는 방법으로서,

첫 번째 도메인에서부터 $n-1$ 번째 도메인까지 학습이 순차적으로 완료된 단일 도메인 대화 모델을, n 번째 도메인에 해당하는 복수의 학습 데이터들과, 상기 첫 번째 도메인에서부터 상기 $n-1$ 번째 도메인까지 추출한 복수의 학습 데이터들로 단일 도메인 대화 모델을 학습시키는 단계, 그리고

상기 학습 데이터들을 클러스터링하여, 상기 n 번째 도메인에 대한 복수의 학습 데이터들을 추출하는 단계를 포함하는, 다중 도메인 대화 모델 생성 방법.

청구항 2

제1항에 있어서,

상기 단일 도메인 대화 모델을 학습시키는 단계는,

상기 n 번째 도메인의 복수의 학습 데이터들과 상기 복수의 학습 데이터들에서 미리 설정한 비율로 데이터들을 선택하여 상기 단일 도메인 대화 모델을 학습시키는, 다중 도메인 대화 모델 생성 방법.

청구항 3

제2항에 있어서,

상기 복수의 학습 데이터들을 추출하는 단계는,

상기 학습 데이터들을 클러스터링하여 앵커를 계산하는 단계,

상기 학습 데이터들을 상기 앵커 근처의 중심 데이터들과 상기 앵커에서 떨어진 경계 데이터들로 구분하는 단계, 그리고

상기 경계 데이터들을 상기 n 번째 도메인의 학습 데이터들로 선택하는 단계

를 포함하는, 다중 도메인 대화 모델 생성 방법.

청구항 4

제3항에 있어서,

상기 복수의 학습 데이터들을 추출하는 단계 이후에,

상기 단일 도메인 대화 모델이 상기 n 번째 도메인의 학습 데이터들에 의해 학습이 종료되었는지 확인하는 단계, 그리고

학습이 종료되었다면, $n+1$ 번째 도메인의 학습 데이터들과 상기 첫 번째 도메인에서부터 상기 n 번째 도메인까지의 학습 데이터들을 이용하여, 상기 단일 도메인 대화 모델을 학습시키는 단계

를 포함하는, 다중 도메인 대화 모델 생성 방법.

청구항 5

제4항에 있어서,

상기 학습이 종료되었는지 확인하는 단계는,

상기 n 번째 도메인의 모든 학습 데이터들로 상기 단일 도메인 대화 모델을 학습시키면 학습이 종료되었다고 판

단하는, 다중 도메인 대화 모델 생성 방법.

청구항 6

제4항에 있어서,

상기 학습이 종료되었는지 확인하는 단계는,

상기 n 번째 도메인의 학습 데이터에 포함된 복수의 질의들에 대한 정답 데이터들과, 해당 질의가 입력되면 상기 단일 도메인 대화 모델이 생성한 답변이 얼마나 매칭되는지 나타내는 BLEU 수치가 미리 설정된 임계 수치 이상이면, 학습이 종료되었다고 판단하는, 다중 도메인 대화 모델 생성 방법.

청구항 7

단일 도메인 대화 모델로부터 다중 도메인 대화 모델을 생성하는 다중 도메인 대화 모델 생성 시스템으로서,

첫 번째 도메인에서부터 $n-1$ 번째 도메인까지 학습이 순차적으로 완료된 단일 도메인 대화 모델을, n 번째 도메인의 복수의 학습 데이터들과 상기 첫 번째 도메인에서부터 상기 $n-1$ 번째 도메인까지의 복습 데이터들로 학습시켜, 다중 도메인 대화 모델로 생성하는 대화 모델 학습부, 그리고

상기 n 번째 도메인의 복수의 학습 데이터들을 클러스터링하고, 상기 n 번째 도메인의 복습 데이터들을 추출하는 데이터 샘플링부

를 포함하는, 다중 도메인 대화 모델 생성 시스템.

청구항 8

제7항에 있어서,

상기 대화 모델 학습부는,

상기 n 번째 도메인의 복수의 학습 데이터들과 상기 복수의 복습 데이터들에서 미리 설정한 비율로 데이터들을 선택하여 상기 단일 도메인 대화 모델을 학습시키는, 다중 도메인 대화 모델 생성 시스템.

청구항 9

제8항에 있어서,

상기 대화 모델 학습부는,

상기 n 번째 도메인의 학습 데이터들에 의해 상기 단일 도메인 대화 모듈의 학습이 종료되면, $n+1$ 번째 도메인의 학습 데이터들과 상기 첫 번째 도메인에서부터 상기 n 번째 도메인까지의 복습 데이터들을 이용하여 상기 학습된 단일 도메인 대화 모델을 학습시키는, 다중 도메인 대화 모델 생성 시스템.

청구항 10

제9항에 있어서,

상기 데이터 샘플링부는,

상기 n 번째 도메인의 학습 데이터들을 클러스터링하여 앵커를 계산하고, 상기 앵커에서 떨어져 분포된 경계 데이터들을 상기 n 번째 도메인의 복습 데이터들로 선택하는, 다중 도메인 대화 모델 생성 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 단일 도메인 대화 모델에 연속 학습 기법을 결합하여 다중 도메인 대화 모델을 생성하는 시스템 및 방법에 대한 기술이다.

배경 기술

[0002] 성공적인 대화 모델은 여러 영역을 포괄하는 말뭉치들로 학습되어 있으며, 다양한 질문에 응답할 수 있다. 이를

위해서는 대화 모델을 큰 말뭉치로 학습시켜야 하는데, 다음과 같은 몇 가지 문제들이 발생한다.

- [0003] 먼저, 대화 모델에 질의한 내용이 대화 모델을 학습시키기 위해 사용된 학습 데이터 중 데이터 수가 적은 도메인에 대한 질의인 경우, 대화 모델에 질의를 하더라도 응답이 좋지 않다는 것이다. 또한, 새 도메인에 대한 질의가 입력되면 적절한 응답을 생성하지 못하며, 질의에 대해 "응, 그래" 또는 "잘 모르겠어"등 매우 일반적인 대답을 생성하는 경향이 있다.
- [0004] 이러한 문제를 해결할 수 있는 해결책은 학습 데이터의 수를 늘리는 것이나, 학습 데이터를 무한히 늘리는 것이 불가능하다. 또 다른 해결책으로는 모델이 구체적인 응답을 생성할 수 있도록, 도메인 선택 모듈을 도입하여 응답을 생성할 대화 모델을 선택하고, 선택한 대화 모델을 이용하여 응답하도록 학습시키는 것이다. 그러나, 이 구조의 문제점은, 새로운 도메인이 추가될수록 도메인 선택 모듈과 대화 모델을 둘 다 새롭게 학습시켜야 하기 때문에, 확장성에 문제점이 발생한다.
- [0005] 이러한 문제점을 해결하기 위해 대부분 단일 도메인 대화 모델을 사용한다. 대화 모델이 다루는 도메인의 수가 점진적으로 증가함에 따라, 모델 학습에서 가장 중요한 문제는 모델이 새 도메인을 학습하는 동안 이전 도메인을 계속 기억하도록 만드는 것이다. 그러나 기계 학습 방법을 기반으로 하는 대부분의 대화 모델은 새로운 것을 배울 때 이전 지식이나 영역을 잊게 되는데, 이를 치명적인 망각(catastrophic forgetting)이라고 한다.
- [0006] 치명적인 망각을 극복하기 위한 초기 해결책으로, 대화 모델을 학습시키는 모든 영역의 모든 훈련 인스턴스를 보존하는 데 중점을 두었다. 따라서 모델이 새 도메인을 학습하는 시기에 관계없이, 이전에 학습한 도메인과 새 도메인에 대한 모든 인스턴스들로 대화 모델을 학습시키는 것이다.
- [0007] 하지만, 이러한 방식은 도메인 수가 증가함에 따라 비용이 많이 든다. 따라서, 훈련 인스턴스의 수를 극적으로 늘리지 않고도 치명적인 망각 문제를 해결하는 방식들이 연구되었다.
- [0008] 그 중 하나의 방식은 도메인 지식을 보존하기 위해 모델 매개 변수를 정규화하는 방식이다. 기존 도메인에 대해 훈련된 모델이 새 도메인에 대해 다시 훈련되면 해당 매개 변수는 새 도메인에 대해서만 최적화되는 문제점이 있다.
- [0009] 또 다른 접근법은 훈련된 모델의 구조를 동적으로 변경하여 새로운 도메인을 학습하는 것이다. 그러나 대화 모델을 학습시킬 도메인의 수가 늘어날수록 학습 복잡도 역시 증가되는 문제점이 있다.

발명의 내용

해결하려는 과제

- [0010] 따라서, 본 발명은 단일 도메인에서 잘 동작하는 대화 모델을 연속 학습 기법을 적용하여 순차적으로 학습시켜, 도메인 특화 능력을 유지하면서 다중 도메인을 수용할 수 있는, 단일 도메인 대화 모델을 이용한 다중 도메인 대화 모델 생성 시스템 및 방법을 제공한다.

과제의 해결 수단

- [0011] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 하나의 특징인 적어도 하나의 프로세서에 의해 동작하는 대화 모델 생성 시스템이 다중 도메인 대화 모델을 생성하는 방법으로서,
- [0012] 첫 번째 도메인에서부터 $n-1$ 번째 도메인까지 학습이 순차적으로 완료된 단일 도메인 대화 모델을, n 번째 도메인에 해당하는 복수의 학습 데이터들과, 상기 첫 번째 도메인에서부터 상기 $n-1$ 번째 도메인까지 추출한 복수의 학습 데이터들로 단일 도메인 대화 모델을 학습시키는 단계, 그리고 상기 학습 데이터들을 클러스터링하여, 상기 n 번째 도메인에 대한 복수의 학습 데이터들을 추출하는 단계를 포함한다.
- [0013] 상기 단일 도메인 대화 모델을 학습시키는 단계는, 상기 n 번째 도메인의 복수의 학습 데이터들과 상기 복수의 학습 데이터들에서 미리 설정한 비율로 데이터들을 선택하여 상기 단일 도메인 대화 모델을 학습시킬 수 있다.
- [0014] 상기 복수의 학습 데이터들을 추출하는 단계는, 상기 학습 데이터들을 클러스터링하여 앵커를 계산하는 단계, 상기 학습 데이터들을 상기 앵커 근처의 중심 데이터들과 상기 앵커에서 떨어진 경계 데이터들로 구분하는 단계, 그리고 상기 경계 데이터들을 상기 n 번째 도메인의 학습 데이터들로 선택하는 단계를 포함할 수 있다.
- [0015] 상기 복수의 학습 데이터들을 추출하는 단계 이후에, 상기 단일 도메인 대화 모델이 상기 n 번째 도메인의 학습 데이터들에 의해 학습이 종료되었는지 확인하는 단계, 그리고 학습이 종료되었다면, $n+1$ 번째 도메인의 학습 데

이터들과 상기 첫 번째 도메인에서부터 상기 n번째 도메인까지의 복습 데이터들을 이용하여, 상기 단일 도메인 대화 모델을 학습시키는 단계를 포함할 수 있다.

- [0016] 상기 학습이 종료되었는지 확인하는 단계는, 상기 n번째 도메인의 모든 학습 데이터들로 상기 단일 도메인 대화 모델을 학습시키면 학습이 종료되었다고 판단할 수 있다.
- [0017] 상기 학습이 종료되었는지 확인하는 단계는, 상기 n번째 도메인의 학습 데이터에 포함된 복수의 질의들에 대한 정답 데이터들과, 해당 질의가 입력되면 상기 단일 도메인 대화 모델이 생성한 답변이 얼마나 매칭되는지 나타내는 BLEU 수치가 미리 설정된 임계 수치 이상이면, 학습이 종료되었다고 판단할 수 있다.
- [0018] 기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 또 다른 특징인 단일 도메인 대화 모델로부터 다중 도메인 대화 모델을 생성하는 다중 도메인 대화 모델 생성 시스템으로서,
- [0019] 첫 번째 도메인에서부터 n-1번째 도메인까지 학습이 순차적으로 완료된 단일 도메인 대화 모델을, n번째 도메인의 복수의 학습 데이터들과 상기 첫 번째 도메인에서부터 상기 n-1번째 도메인까지의 복습 데이터들로 학습시켜, 다중 도메인 대화 모델로 생성하는 대화 모델 학습부, 그리고 상기 n번째 도메인의 복수의 학습 데이터들을 클러스터링하고, 상기 n번째 도메인의 복습 데이터들을 추출하는 데이터 샘플링부를 포함한다.
- [0020] 상기 대화 모델 학습부는, 상기 n번째 도메인의 복수의 학습 데이터들과 상기 복수의 복습 데이터들에서 미리 설정한 비율로 데이터들을 선택하여 상기 단일 도메인 대화 모델을 학습시킬 수 있다.
- [0021] 상기 대화 모델 학습부는, 상기 n번째 도메인의 학습 데이터들에 의해 상기 단일 도메인 대화 모듈의 학습이 종료되면, n+1번째 도메인의 학습 데이터들과 상기 첫 번째 도메인에서부터 상기 n번째 도메인까지의 복습 데이터들을 이용하여 상기 학습된 단일 도메인 대화 모델을 학습시킬 수 있다.
- [0022] 상기 데이터 샘플링부는, 상기 n번째 도메인의 학습 데이터들을 클러스터링하여 앵커를 계산하고, 상기 앵커에서 떨어져 분포된 경계 데이터들을 상기 n번째 도메인의 복습 데이터들로 선택할 수 있다.

발명의 효과

- [0023] 본 발명에 따르면, 다중 도메인 대화 모델을 만들 때, 도메인에 특화된 대답이 줄어드는 문제점을 해결할 수 있다.
- [0024] 또한, 단일 도메인 대화 모델을 사용하여 다중 도메인 모델을 만들 때, 순차적으로 학습하면서 발생하는 치명적인 망각 문제를 연속 학습 기법을 이용하여 해결할 수 있다.
- [0025] 또한, 단일 도메인에서 잘 동작하는 대화 모델을 이용하기 때문에, 새로운 도메인을 추가하기 위한 재학습 시간을 줄여줄 수 있다.
- [0026] 또한, 데이터의 부분만 샘플링하여 기존의 대화 모델을 재사용할 수 있으며, 질의에 적절하고 구체적인 응답 생성이 가능하다.

도면의 간단한 설명

- [0027] 도 1은 본 발명의 실시예에 따른 다중 도메인 대화 모델 생성 시스템의 구조도이다.
- 도 2는 본 발명의 실시예에 따라 다중 도메인 대화 모델을 생성하는 방법의 흐름도이다.
- 도 3은 본 발명의 실시예에 따른 다중 도메인 대화 모델 생성 시스템의 예시도이다.
- 도 4는 본 발명의 실시예에 따른 복습 데이터로 샘플링된 데이터들의 예시도이다.
- 도 5는 본 발명의 실시예에 따른 컴퓨팅 장치의 구조도이다.

발명을 실시하기 위한 구체적인 내용

- [0028] 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.

- [0029] 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.
- [0030] 이하, 도면을 참조로 하여, 본 발명의 실시예에 따른 단일 도메인 대화 모델을 이용한 다중 도메인 대화 모델 생성 시스템 및 방법에 대해 설명한다.
- [0031] 도 1은 본 발명의 실시예에 따른 다중 도메인 대화 모델 생성 시스템의 구조도이다.
- [0032] 도 1에 도시된 바와 같이, 단일 도메인 대화 모델에 연속 학습 기법을 적용하여 다중 도메인 대화 모델을 생성하는 대화 모델 생성 시스템(100)은 대화 모델 학습부(110)와 데이터 샘플링부(120)를 포함한다.
- [0033] 대화 모델 학습부(110)는 복수의 도메인(예를 들어, 날씨, 음식, 뷰티, 오락 등)별 수집한 학습 데이터들로 대화 모델을 학습시킨다. 대화 모델 학습부(110)에 의해 학습된 대화 모델은, 도메인에 상관없이 질의가 입력되면 질의에 해당하는 답변을 생성하여 제공할 수 있다.
- [0034] 데이터 샘플링부(120)는 각 도메인별 학습 데이터들을 샘플링하여, 각 도메인별 복습 데이터를 생성한다. 이렇게 생성된 도메인별 복습 데이터는 다른 도메인의 학습 데이터로 대화 모델 학습부(110)가 대화 모델을 학습시킬 때, 다른 도메인의 학습 데이터와 함께 복습 데이터로 대화 모델을 학습시키는데 사용된다.
- [0035] 즉, 복수의 학습 데이터들이 복수의 도메인별로 각각 있다고 가정한다. 그러면, 순차적으로 복수의 도메인들 중 어느 하나의 도메인에 해당하는 복수의 학습 데이터들을 이용하여 대화 모델 학습부(110)는 대화 모델을 학습시킨다. 여기서, 각 학습 데이터들의 형태나, 대화 모델 학습부(110)가 학습 데이터들로 대화 모델을 학습시키는 방법은 어느 하나로 한정하지 않는다.
- [0036] 대화 모델 학습부(110)가 대화 모델을 특정 도메인의 학습 데이터들로 학습시키면, 데이터 샘플링부(120)는 해당 도메인의 학습 데이터들을 입력으로 받아 클러스터링하여 앵커(anchor)를 계산한다. 데이터 샘플링부(120)는 다양한 방법으로 학습 데이터들을 클러스터링하여 앵커를 계산할 수 있으며, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.
- [0037] 데이터 샘플링부(120)가 학습 데이터들로부터 앵커를 계산하면, 학습 데이터들은 앵커 근처의 중심 데이터들(centroid)과 앵커에서 멀리 떨어진 경계 데이터들(boundary)로 구분된다. 데이터 샘플링 모듈(120)은 클러스터링한 학습 데이터들 중, 경계 데이터들로 구분된 학습 데이터들을 복습 데이터로 설정한다.
- [0038] 데이터 샘플링부(120)에서 설정한 복습 데이터는 대화 모델 학습부(110)가 다른 도메인의 학습 데이터들로 대화 모델을 학습시킬 때, 함께 대화 모델을 학습시키는데 사용된다. 이때, 대화 모델 학습부(110)가 제일 처음 대화 모델을 학습시킬 때에는, 이전 학습 데이터의 복습 데이터가 없기 때문에 학습 데이터만 이용하여 대화 모델을 학습시킨다.
- [0039] 이와 같이 이전에 학습시킨 도메인들의 학습 데이터들에서부터 추출된 복습 데이터들을 새로운 도메인의 학습 데이터와 함께 대화 모델을 학습시키는데 사용함으로써, 이전에 학습된 도메인에 대한 치명적인 망각이 발생하는 것을 방지할 수 있다.
- [0040] 다음은 상술한 대화 모델 생성 시스템(100)이 대화 모델을 학습시키는 방법에 대해 도 2를 참조로 설명한다.
- [0041] 도 2는 본 발명의 실시예에 따라 다중 도메인 대화 모델을 생성하는 방법의 흐름도이다.
- [0042] 도 2에 도시된 바와 같이, 대화 모델 생성 시스템(100)은 n 번째 도메인의 학습 데이터로 대화 모델을 학습시킨다(S100). 여기서 n 은 정수로, $1 \leq n \leq N$ 에 해당하며, 도메인의 수를 어느 하나의 수로 한정하지 않는다.
- [0043] 이때, 대화 모델 생성 시스템(100)은 $n-1$ 번째 도메인까지의 복습 데이터와 n 번째 도메인의 학습 데이터를 이용하여 대화 모델을 학습시킨다.
- [0044] 만약, n 이 1인 경우에는, 이전 도메인이 없기 때문에 첫 번째 도메인의 학습 데이터만 이용하여 대화 모델을 학습시킨다. 그러나, n 이 2인 경우에는 두 번째 도메인의 학습 데이터와 첫 번째 도메인에서 추출한 복습 데이터로 대화 모델을 학습시킨다. 또한, n 이 3인 경우에는 세 번째 도메인의 학습 데이터와 첫 번째 도메인에서 추출한 복습 데이터 및 두 번째 도메인에서 추출한 복습 데이터를 이용하여 대화 모델을 학습시킨다.
- [0045] 여기서, 학습 데이터는 질의와 해당 질의에 대한 정답 문장의 쌍으로 이루어지며, 질의가 대화 모델에 입력될 때 정답 문장이 생성되어 출력되도록 대화 모델을 학습시킨다. 대화 모델 생성 시스템(100)이 학습 데이터로 대화 모델을 학습시키는 방법은 다양한 방법으로 수행할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법

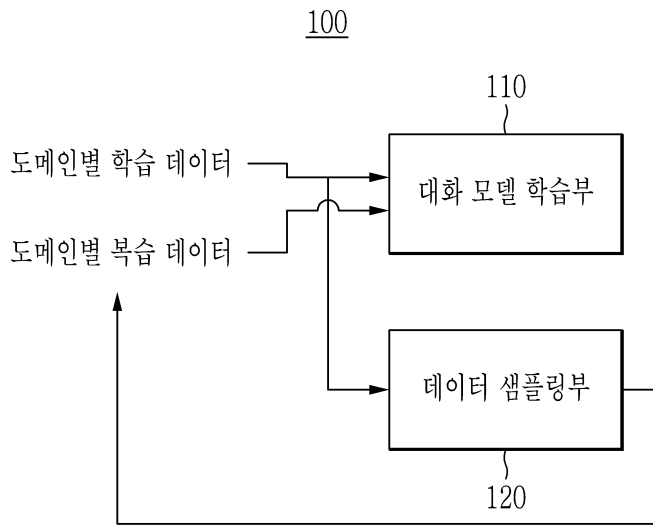
으로 한정하지 않는다.

- [0046] n번째 도메인에 대한 대화 모델을, 대화 모델 생성 시스템(100)은 대화 모델을 학습시킨 n번째 도메인의 학습 데이터에서 복습 데이터를 추출한다(S110). 이를 위해, 대화 모델 생성 시스템(100)은 복수의 학습 데이터에 해당하는 질의 문장들에서, 해당 도메인에서 자주 발화되는 중심 데이터들과, 자주 발화되지는 않으나 해당 도메인에 해당하는 학습 데이터들인 경계 데이터들로 분류한다. 그리고, 경계 데이터들을 복습 데이터로 선택한다. 대화 모델 생성 시스템(100)이 학습 데이터들을 클러스터링하여 중심 데이터와 경계 데이터로 분류하는 방법은 다양한 방법으로 실행할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.
- [0047] 대화 모델 생성 시스템(100)은 n번째 도메인에 대한 학습이 종료되었는지 확인한다(S120). 대화 모델 생성 시스템(100)은 다양한 방법으로 학습 데이터의 학습이 종료되었는지 확인할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.
- [0048] 즉, 대화 모델 생성 시스템(100)은 n번째 도메인에서 수집된 모든 학습 데이터들로 대화 모델을 학습시킨 경우 학습이 종료되었다고 확인할 수 있다. 또한, 대화 모델 생성 시스템(100)은 학습 데이터에 포함된 복수의 질의들에 대한 정답 데이터들과, 해당 질의가 입력되면 대화 모델이 생성한 답변이 얼마나 매칭되는지 나타내는 BLEU 수치가 미리 설정된 임계 수치 이상일 경우 학습이 종료되었다고 판단할 수도 있다.
- [0049] S120 단계에서 확인한 결과 n번째 도메인에 대한 학습이 종료되지 않았다면, S100 단계 이후의 절차를 반복 수행한다.
- [0050] 그러나, n번째 도메인에 대한 학습이 종료되었다면, 대화 모델 생성 시스템(100)은 n+1번째 도메인에 대한 학습 데이터를 입력으로 받아 n번째 도메인까지의 복습 데이터들과 함께 대화 모델을 학습시킨다(S130). 이와 함께, 대화 모델 생성 시스템(100)은 n+1번째 도메인의 복습 데이터를 추출한다(S140).
- [0051] 이때, 대화 모델 생성 시스템(100)은 새로 학습시킬 n+1번째 도메인에 대한 학습 데이터의 수와 첫 번째 도메인에서부터 n번째 도메인까지 각각 추출한 복습 데이터들의 수를 미리 설정한 비율로 선택하여 대화 모델을 학습시킨다.
- [0052] 예를 들어, n+1번째 도메인의 학습 데이터들 중 80%의 학습 데이터들과 첫 번째 도메인에서부터 n번째 도메인까지 각각 추출한 복습 데이터들 중 20%의 복습 데이터들로 대화 모델을 학습시킬 수 있다. 여기서, 비율은 어느 하나의 수치로 한정하지 않는다.
- [0053] 이상의 절차에 따라 단일 대화 모델을 연속 학습 기법을 이용하여 각 도메인의 학습 데이터들과 복습 데이터들로 학습시켜 다중 도메인 대화 모델로 생성하는 예에 대해 도 3을 참조로 설명한다.
- [0054] 도 3은 본 발명의 실시예에 따른 다중 도메인 대화 모델 생성 시스템의 예시도이다.
- [0055] 대화 모델을 학습시키기 위한 세 가지 도메인으로 음식, 미용, 오락을 예로 하여 설명한다. 도 3에 도시된 바와 같이, 대화 모델 생성 시스템(100)은 먼저 음식 도메인에 대한 학습 데이터로 대화 모델을 학습시켜 대화 모델 1을 생성한다. 그리고, 음식 도메인에 대한 학습 데이터를 클러스터링하여, 음식 도메인에 대한 복습 데이터(음식')를 생성한다.
- [0056] 대화 모델 생성 시스템(100)은 음식 도메인에 대한 학습이 완료되면, 미용 데이터와 음식 복습 데이터로 대화 모델 1을 학습시켜 대화 모델 2를 생성한다. 그리고, 미용 도메인에 대한 학습 데이터를 클러스터링하여, 미용 도메인에 대한 복습 데이터(미용')를 생성한다.
- [0057] 대화 모델 생성 시스템(100)은 미용 도메인에 대한 학습이 완료되면, 오락 데이터, 음식 복습 데이터 그리고 미용 복습 데이터로 대화 모델 2를 학습시켜 대화 모델 3을 생성한다. 그리고, 오락 도메인에 대한 학습 데이터를 클러스터링하여, 오락 도메인에 대한 복습 데이터(오락')를 생성한다.
- [0058] 이러한 절차로 대화 모델 생성 시스템(100)은 n개의 도메인에 대한 학습 데이터와 첫 번째 도메인인 음식 도메인의 복습 데이터에서부터 n-1번째 도메인의 복습 데이터까지 모두 사용하여 대화 모델을 학습시켜, 최종 대화 모델을 생성한다.
- [0059] 다음은 대화 모델 생성 시스템(100)의 데이터 샘플링부(120)가 복습 데이터를 추출하는 예에 대해 도 4를 참조로 설명한다.
- [0060] 도 4는 본 발명의 실시예에 따른 복습 데이터로 샘플링된 데이터들의 예시도이다.

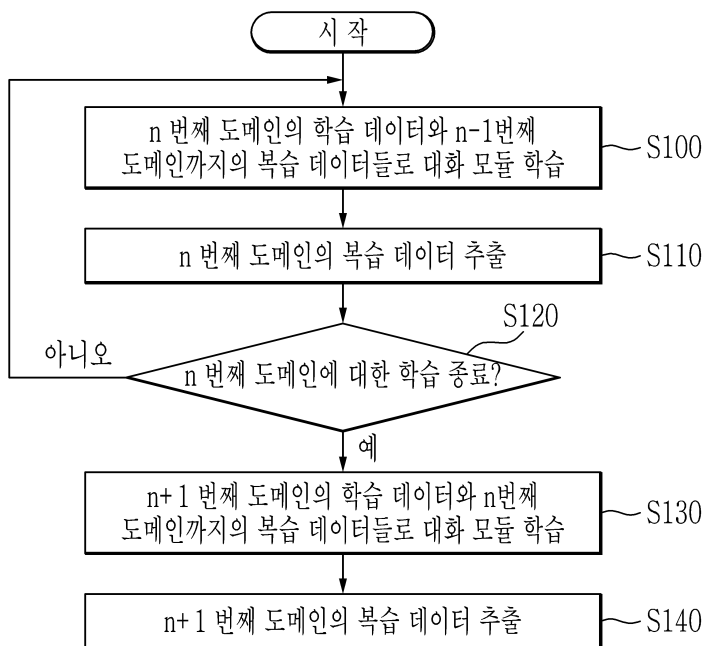
- [0061] 도 4에 도시된 바와 같이, 각 도메인의 학습 데이터는 같은 대화 패턴을 가진다고 가정하면, 데이터 샘플링부(120)는 복수의 학습 데이터들을 클러스터링하여 앵커를 계산한다. 그리고, 데이터 샘플링부(120)는 학습 데이터들을, 계산한 앵커 근처의 데이터들인 중심 데이터들(①)과 앵커에서 멀리 떨어져 분포되어 있는 데이터들인 경계 데이터들(②)로 구분할 수 있다.
- [0062] 데이터 샘플링부(120)는 경계 데이터들(②)을 복습 데이터로 선택하는 것을 예로 하여 설명한다. 그러나, 반드시 이와 같이 한정되는 것은 아니다.
- [0063] 예를 들어, 음식 도메인의 경우 "계산해주세요" 또는 "쿠폰 찍어드릴까요?" 등의 질의에 "만원입니다." 또는 "찍어주세요" 등의 답변이 발생하는 학습 데이터는 중심 데이터들에 해당한다. 그러나, "삼겹살 1근 주세요" 또는 "홍시 스무디 있나요?" 등과 같은 질의는 음식 도메인에 해당하지만 자주 질의가 발생하지 않거나 일부 특화된 질의에 해당하는 경우, 경계 데이터들로 분류하여 복습 데이터로써 대화 모델을 복습시키는데 사용한다.
- [0064] 도 5는 본 발명의 실시예에 따른 컴퓨팅 장치의 구조도이다.
- [0065] 도 5에 도시된 바와 같이, 대화 모델 생성 시스템(100)은 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치(200)에서, 본 발명의 동작을 실행하도록 기술된 명령들(instructions)이 포함된 프로그램을 실행한다.
- [0066] 컴퓨팅 장치(200)의 하드웨어는 적어도 하나의 프로세서(210), 메모리(220), 스토리지(230), 통신 인터페이스(240)를 포함할 수 있고, 버스를 통해 연결될 수 있다. 이외에도 입력 장치 및 출력 장치 등의 하드웨어가 포함될 수 있다. 컴퓨팅 장치(200)는 프로그램을 구동할 수 있는 운영 체제를 비롯한 각종 소프트웨어가 탑재될 수 있다.
- [0067] 프로세서(210)는 컴퓨팅 장치(200)의 동작을 제어하는 장치로서, 프로그램에 포함된 명령들을 처리하는 다양한 형태의 프로세서(210)일 수 있고, 예를 들면, CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 등 일 수 있다.
- [0068] 메모리(220)는 본 발명의 동작을 실행하도록 기술된 명령들이 프로세서(210)에 의해 처리되도록 해당 프로그램을 로드한다. 메모리(220)는 예를 들면, ROM(read only memory), RAM(random access memory) 등 일 수 있다.
- [0069] 스토리지(230)는 본 발명의 동작을 실행하는데 요구되는 각종 데이터, 프로그램 등을 저장한다. 통신 인터페이스(240)는 유/무선 통신 모듈일 수 있다.
- [0070] 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

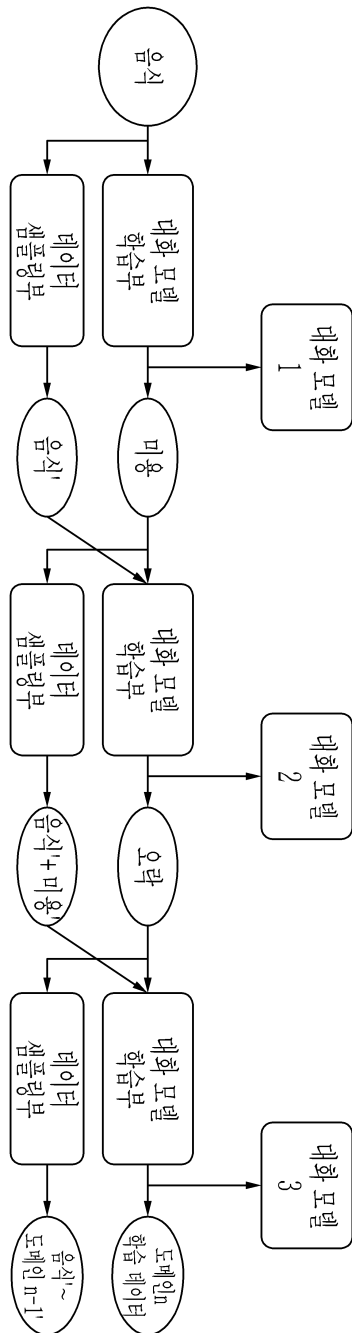
도면1



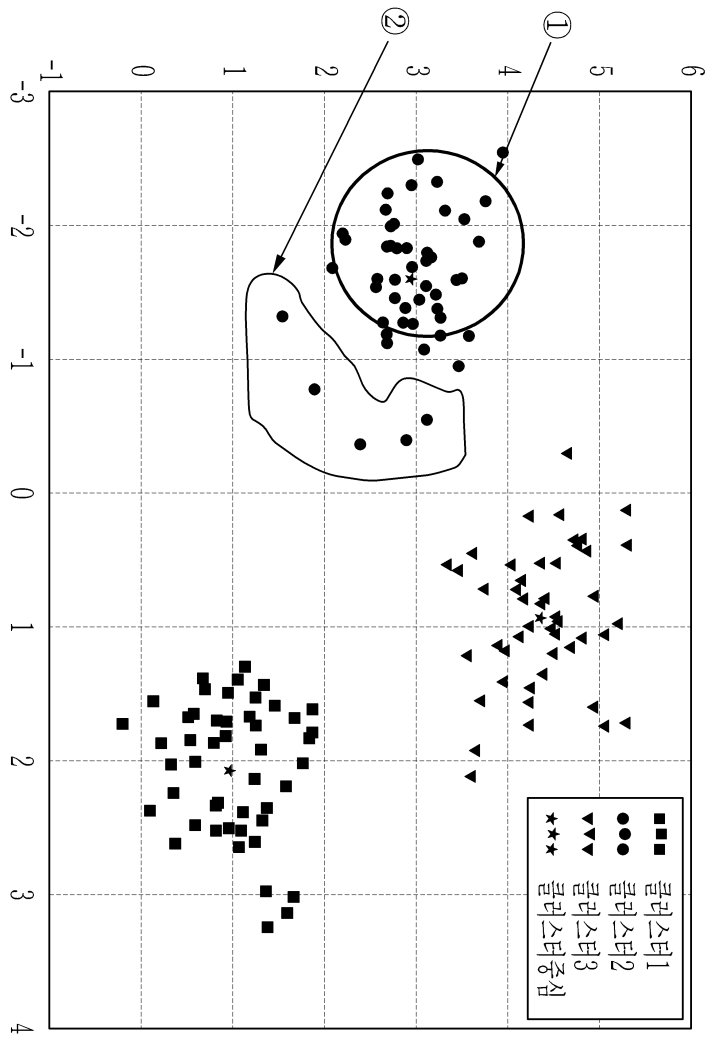
도면2



도면3



도면4



도면5

