(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0056114
(43) 공개일자 2021년05월18일

(51) 국제특허분류(Int. Cl.)
G06F 16/332 (2019.01) *G06F 16/2458* (2019.01)
G06F 16/33 (2019.01) *G06F 40/20* (2020.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06F 16/3329 (2019.01)
G06F 16/2468 (2019.01)
(21) 출원번호 10-2019-0142801
(22) 출원일자 2019년11월08일
심사청구일자 2019년11월08일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
차다은
서울특별시 성북구 개운사길 21-3, 303호
박기남
서울특별시 강서구 마곡중앙로 33, 1406동 101호
(뒷면에 계속)
(74) 대리인
김홍석

전체 청구항 수 : 총 6 항

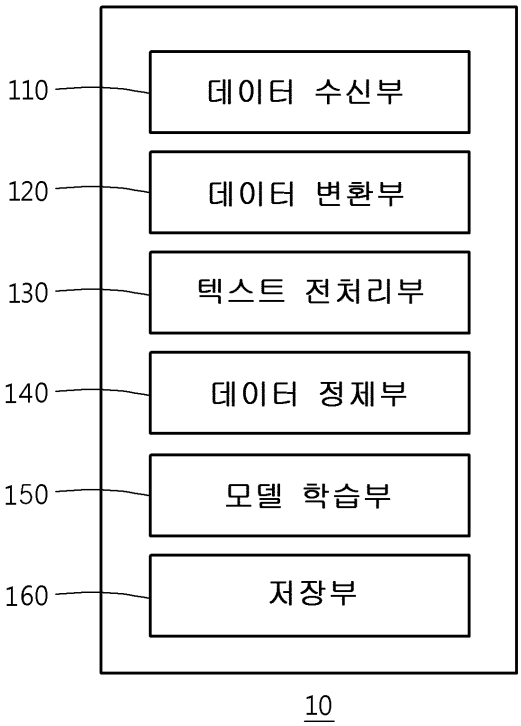
(54) 발명의 명칭 **자동 질의응답 장치**

(57) 요 약

자동 질의응담 모델 생성 장치 및 자동 질의응답 장치가 개시된다. 상기 자동 질의응답 모델 생성 장치는 질의응답 데이터에 포함된 복수의 질의들 중 복수의 질문들을 포함하는 적어도 하나의 질의를 복수의 질문들로 분할하고, 상기 질의응답 데이터에 포함된 복수의 응답들 중 복수의 답변들을 포함하는 적어도 하나의 응답을 복수의

(뒷면에 계속)

대 표 도 - 도1



답변들로 분할하여 복수의 질문들과 복수의 답변들에 대한 카티전 곱(cartesian product)으로 이루어진 질의응답 쌍들을 생성하는 데이터 변환부, 상기 질의응답 쌍들에 포함된 복수의 질문들과 복수의 답변들에서 불완전한 음절로 구성된 음절들과 사용자 식별 정보를 제거하고 형태소 단위로 분석하는 텍스트 전처리부, 텍스트 랭크(TextRank) 알고리즘을 이용하여 상기 질의응답 쌍들에 포함된 복수의 답변들 중에서 미리 정해진 중요도 값보다 낮은 중요도 값을 갖는 답변을 제거하고, TF-IDF(Term Frequency-Inverse Document Frequency)를 이용하여 상기 질의응답 쌍들 중에서 잘못 매칭된 질의응답 쌍을 제거하여 제2 질의응답 쌍들을 생성하는 데이터 정제부, 및 상기 제2 질의응답 쌍들을 학습 데이터로 이용하여 자동 질의응답 모델을 학습시키는 모델 학습부를 포함한다.

(52) CPC특허분류
G06F 16/3344 (2019.01)
G06F 40/268 (2020.01)
G06N 3/08 (2013.01)

(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
황태선
대전광역시 유성구 왕가봉로 23, 1101동 302호
양기수
경기도 하남시 초이로99번길 49-12, 초이화평교회 지하1층

이 발명을 지원한 국가연구개발사업

과제고유번호	1711093067
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	대학ICT연구센터지원사업
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2019.03.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

질의응답 데이터에 포함된 복수의 질의들 중 복수의 질문들을 포함하는 적어도 하나의 질의를 복수의 질문들로 분할하고, 상기 질의응답 데이터에 포함된 복수의 응답들 중 복수의 답변들을 포함하는 적어도 하나의 응답을 복수의 답변들로 분할하여 복수의 질문들과 복수의 답변들에 대한 카티전 곱(cartesian product)으로 이루어진 질의응답 쌍들을 생성하는 데이터 변환부;

상기 질의응답 쌍들에 포함된 복수의 질문들과 복수의 답변들에서 불완전한 음절로 구성된 음절들과 사용자 식별 정보를 제거하고 형태소 단위로 분석하는 텍스트 전처리부;

텍스트 랭크(TextRank) 알고리즘을 이용하여 상기 질의응답 쌍들에 포함된 복수의 답변들 중에서 미리 정해진 중요도 값보다 낮은 중요도 값을 갖는 답변을 제거하고, TF-IDF(Term Frequency-Inverse Document Frequency)를 이용하여 상기 질의응답 쌍들 중에서 잘못 매칭된 질의응답 쌍을 제거하여 제2 질의응답 쌍들을 생성하는 데이터 정제부; 및

상기 제2 질의응답 쌍들을 학습 데이터로 이용하여 자동 질의응답 모델을 학습시키는 모델 학습부를 포함하는 자동 질의응답 모델 생성 장치.

청구항 2

제1항에 있어서,

상기 텍스트 전처리부는 미리 정해진 빈도 이상의 동시 출현 빈도를 갖는 합성어를 하나의 형태소로 분석하는,

자동 질의응답 모델 생성 장치.

청구항 3

제1항에 있어서,

데이터 정제부는 상기 질의응답 쌍들에 포함된 복수의 질문들에 대하여 단어들의 TF-IDF 값을 질문의 특징 벡터로 이용하고 상기 질의응답 쌍들에 포함된 복수의 답변들에 대하여 단어들의 TF-IDF 값을 답변의 특징 벡터로 이용하여 질문과 답변의 코사인 유사도를 계산한 후 미리 정해진 유사도 값보다 낮은 유사도 값을 갖는 질의응답 쌍을 제거하는,

자동 질의응답 모델 생성 장치.

청구항 4

제1항에 있어서,

상기 모델 학습부는 COALA(COmpare-Aggregate for Long Answers) 모델을 학습시킴으로써 상기 자동 질의응답 모델을 생성하는,

자동 질의응답 모델 생성 장치.

청구항 5

사용자 단말로부터 질의를 수신하는 데이터 수신부;

상기 사용자 단말로부터 수신된 질의에서 불완전한 음절로 구성된 음절들과 사용자 식별 정보를 제거하고 형태소 단위로 분석하는 텍스트 전처리부;

TF-IDF를 이용하여 상기 질의와 기준치 이상의 유사도를 보이는 미리 저장되어 있는 답변들 중에서 복수의 답변들을 선택하여 답변 후보군을 생성하는 답변 후보군 선택부; 및

제1항에 기재된 자동 질의응답 모델 생성 장치에 의해 생성된 자동 질의응답 모델을 이용하여 상기 답변 후보군 중에서 적어도 하나의 답변을 선택하는 최종 답변 선택부를 포함하는 자동 질의응답 장치.

청구항 6

제5항에 있어서,

상기 텍스트 전처리부는 미리 정해진 빈도 이상의 동시 출현 빈도를 갖는 합성어를 하나의 형태소로 분석하는,

자동 질의응답 모델 생성 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 자동 질의응답 시스템에 관한 것으로, 특히 자연어처리 기술과 딥러닝(Deep Learning) 기반 질의응답 기술을 활용하여 소량의 정제되지 않은 기존 질의응답 데이터로부터 질의응답 시스템을 구축하는 장치 및 방법과 질의응답 장치 및 방법에 관한 것이다.

배 경 기 술

[0003] 여러 질문자가 비슷한 질문을 반복하는 질의응답 상황에서 인간이 질문에 대한 답변을 직접 제시하는 것은 많은 노동력을 요구하며 비효율적이다. 딥러닝 기반 자동 질의응답 기술의 발달로 이러한 문제가 해결될 것으로 기대되었으나, 많은 사용자들은 대량의 정제된 학습 데이터 확보의 어려움으로 인해 자동 질의응답 시스템 구축에 어려움을 겪는다.

[0004] 이에 따라 소량의 질의응답 데이터로부터 자동 질의응답 시스템을 구축하기 위한 연구들이 cQA(community based Question Answering) 분야에서 진행되어왔다. 비특허문헌 1, 2, 3에서는 LSP 등 인간이 직접 정의한 특징에 따라 질의응답 데이터를 카테고리별로 분류한 뒤, 카테고리 내에서 문서 유사도를 기반으로 적합한 답변을 선택하도록 하였다. 이후 비특허문헌 4, 5에서는 CRF를 적용하여 질문의 핵심부를 추출하고 SVM을 적용하여 질문의 카테고리를 분류함으로써, 다양한 도메인에 적용 가능한 질의응답 기술을 제안하였으나 세부 모델의 학습을 위한 데이터 구축이 요구된다는 한계를 보인다. 비특허문헌 6에서는 음절 및 형태소 단위의 Okapi BM25 모델과 딥러닝 모델을 통합하여 질의응답 데이터로부터 다양한 도메인의 질의응답 시스템을 구축할 수 있는 기술을 제안하였다.

[0005] 그러나 질의응답 시스템을 필요로 하는 사용자들이 보유하고 있는 데이터는 주로 Q&A 게시판 데이터나 이메일 질의응답 데이터 등으로, 소량임과 더불어 정제되지 않은 데이터라는 문제가 있다. 이에 따라 본 발명에서는 소량의 정제되지 않은 질의응답 데이터로부터 자동으로 질의응답 시스템을 구축할 수 있는 기술을 제안한다.

[0006] 제안 기술은 '데이터 전처리', '데이터 정제', '문서 유사도 기반 답변 후보군 선택', '딥러닝 기반 답변 선택'의 단계들로 나뉘며, 자연어처리 기술과 딥러닝 기반 질의응답 기술을 활용하여 주어진 데이터를 정제하고 이로부터 새로운 질문에 대한 정확한 답변을 빠르게 제시한다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 등록특허 제1851786호 (2018.04.24. 공고)

(특허문헌 0002) 대한민국 공개특허 제2018-0129886호 (2018.12.05. 공개)

비특허문헌

[0009] (비특허문헌 0001) S. J. Kwon, D. H. Yoo, D. S. Oh & J. Y. Seo. (2014). A cQA System on Korean Language and Grammar using LSP Classification. Journal of Korea Information Science Society, 1263-1265.

(비특허문헌 0002) S. J. Do, Y. S. Kim, H. S. Yeom, S. Y. Jung, K. J. Kim & J. Y. Seo. (2014). Korean Information cQA System using Target-predicator Analysis and Topic Extraction. Journal of Korea Information Science Society, 1290-1292.

(비특허문헌 0003) J. M. Mun, Y. H. Song, J. H. Jin, H. S. Lee & H. A. Lee. (2015). Similar Question Search System for online Q&A for the Korean Language Based on Topic Classification, Journal of Korean Society for Cognitive Science, 263-278.

(비특허문헌 0004) H. B. Shin, W. C. An, Y. J. Jeong, S. J. Kwon, J. A. Kim & J. Y. Seo. (2015). A cQA System of Korean Information Method Based on a Machine Learning Using CRF and SVM, Journal of Korea Information Science Society, 1677-1679.

(비특허문헌 0005) S. J. Kwon, J. A. Kim, S. W. Kang & J. Y. Seo. (2016). A Korean Community-based Question Answering System Using Multiple Machine Learning Methods. Journal of Korea Information Science Society, 1085-1093.

(비특허문헌 0006) H. S. Hwang, K. H. Choi, G. Y. Kim, J. H. Oh & C. K. Lee. (2015). CQA System Using Deep Learning. Journal of Korea Information Science Society, 572-574.

(비특허문헌 0007) T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado & J. Dean. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 3111-3119.

(비특허문헌 0008) R. Mihalcea & P, Tarau (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

(비특허문헌 0009) A. Ruckle, N. S. Moosavi & I. Gurevych (2019). COALA: ANeural Coverage-Based Approach for Long Answer Selection with Small Data, Association for the Advancement of Artificial Intelligence.

발명의 내용

해결하려는 과제

[0010] 본 발명이 이루고자 하는 기술적인 과제는 소량의 미정제 데이터를 대상으로 자동 질의응답 모델을 생성하는 장치 및 방법과 이를 통해 생성된 자동 질의응답 모델을 이용하여 자동 질의응답 서비스를 제공하는 자동 질의응답 장치를 제공하는 것이다.

과제의 해결 수단

[0012] 본 발명의 실시예에 따른 자동 질의응답 모델 생성 장치는 질의응답 데이터에 포함된 복수의 질의들 중 복수의 질문들을 포함하는 적어도 하나의 질의를 복수의 질문들로 분할하고, 상기 질의응답 데이터에 포함된 복수의 응답들 중 복수의 답변들을 포함하는 적어도 하나의 응답을 복수의 답변들로 분할하여 복수의 질문들과 복수의 답변들에 대한 카티전 곱(cartesian product)으로 이루어진 질의응답 쌍들을 생성하는 데이터 변환부, 상기 질의응답 쌍들에 포함된 복수의 질문들과 복수의 답변들에서 불완전한 음절로 구성된 음절들과 사용자 식별 정보를 제거하고 형태소 단위로 분석하는 텍스트 전처리부, 텍스트 랭크(TextRank) 알고리즘을 이용하여 상기 질의응답

쌍들에 포함된 복수의 답변들 중에서 미리 정해진 중요도 값보다 낮은 중요도 값을 갖는 답변을 제거하고, TF-IDF(Term Frequency-Inverse Document Frequency)를 이용하여 상기 질의응답 쌍들 중에서 잘못 매칭된 질의응답 쌍을 제거하여 제2 질의응답 쌍들을 생성하는 데이터 정제부, 및 상기 제2 질의응답 쌍들을 학습 데이터로 이용하여 자동 질의응답 모델을 학습시키는 모델 학습부를 포함한다.

[0013] 본 발명의 실시예에 따른 자동 질의응답 장치는 사용자 단말로부터 질의를 수신하는 데이터 수신부, 상기 사용자 단말로부터 수신된 질의에서 불완전한 음절로 구성된 음절들과 사용자 식별 정보를 제거하고 형태소 단위로 분석하는 텍스트 전처리부, TF-IDF를 이용하여 상기 질의와 기준치 이상의 유사도를 보이는 미리 저장되어 있는 답변들 중에서 복수의 답변들을 선택하여 답변 후보군을 생성하는 답변 후보군 선택부, 및 제1항에 기재된 자동 질의응답 모델 생성 장치에 의해 생성된 자동 질의응답 모델을 이용하여 상기 답변 후보군 중에서 적어도 하나의 답변을 선택하는 최종 답변 선택부를 포함한다.

발명의 효과

[0015] 본 발명의 실시 예에 따른 자동 질의응답 장치에 의할 경우, 소량의 미정제 데이터를 이용하여 학습함으로써 자동 질의응답 모델을 생성하고 이를 통해 자동 질의응답을 수행할 수 있다.

도면의 간단한 설명

[0017] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른 모델 생성 장치의 기능 블럭도이다.

도 2는 본 발명의 일 실시예에 따른 모델 생성 방법을 설명하기 위한 흐름도이다.

도 3에는 데이터 변환 과정 및 데이터 정제 과정이 도시되어 있다.

도 4는 텍스트 전처리부에 의한 텍스트 전처리 단계를 설명하기 위한 도면이다.

도 5는 질의-응답 매칭 필터를 설명하기 위한 도면이다.

도 6은 본 발명의 일 실시예에 따른 자동 질의응답 장치의 기능 블럭도이다.

도 7은 TF-IDF 코사인 유사도 기반 답변 후보군 선택 과정을 설명하기 위한 도면이다.

도 8은 COALA의 전체 구조도를 도시한다.

도 9는 본 발명의 다른 실시예에 따른 자동 질의응답 장치의 기능 블럭도이다.

발명을 실시하기 위한 구체적인 내용

[0018] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0019] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0020] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0021] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어"

있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0022] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0023] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0024] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시 예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0026] 도 1은 본 발명의 일 실시예에 따른 모델 생성 장치의 기능 블럭도이고, 도 2는 본 발명의 일 실시예에 따른 모델 생성 방법을 설명하기 위한 흐름도이다.

[0027] 도 1을 참조하면, 질의응답 모델 생성 장치, 자동 질의응답 모델 생성 장치 등으로 명명될 수도 있는 모델 생성 장치(10)는 자동 질의응답 모델로 명명될 수도 있는 질의응답 모델을 생성할 수 있다. 여기서, 상기 질의응답 모델은 챗봇(chatbot)이라 불리기도 한다. 모델 생성 장치(10)는 데이터 변환부(120), 텍스트 전처리부(130), 데이터 정제부(140), 및 모델 학습부(150)를 포함한다. 실시예에 따라, 모델 생성 장치(10)는 데이터 수신부(110) 및/또는 저장부(160)를 더 포함할 수도 있다.

[0028] 모델 생성 장치(10)는 적어도 하나의 프로세서를 이용하여 구현될 수 있으며, 여기서, 프로세서는 중앙 처리 장치(CPU, Central Processing Unit), 마이크로 컨트롤러 유닛(MCU, Micro Controller Unit), 애플리케이션 프로세서(AP, Application Processor), 마이컴(Micom, Micro Processor), 전자 제어 유닛(ECU, Electronic Controlling Unit) 및/또는 각종 연산 처리 및 제어 신호의 생성이 가능한 다른 전자 장치 등을 포함할 수 있다. 이들 장치는, 예를 들어, 하나 또는 둘 이상의 반도체 칩 및 관련 부품을 이용하여 구현될 수 있다. 일 실시예에 의하면, 프로세서는 저장부(160)에 저장된 적어도 하나의 애플리케이션(소프트웨어, 프로그램이나 앱 등으로 표현 가능하다)을 구동시켜, 미리 정의된 연산, 판단, 처리 및/또는 제어 동작 등을 수행할 수도 있다. 여기서, 저장부(160)에 저장된 애플리케이션은, 설계자에 의해 직접 작성되어 저장부(160)에 입력 및 저장된 것일 수도 있고, 또는 유선 또는 무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망을 통하여 획득 또는 갱신된 것일 수도 있다.

[0029] 또한, 모델 생성 장치(10)는 이와 같은 프로세서가 하나 이상 설치된 적어도 하나의 정보 처리 장치를 이용하여 구현될 수도 있으며, 전자 정보 처리 장치는 데스크톱 컴퓨터, 랩톱 컴퓨터, 서버용 컴퓨터, 스마트 폰, 태블릿 피씨, 스마트 시계, 내비게이션 장치, 휴대용 게임기, 두부 장착형 디스플레이(HMD, Head Mounted Display) 장치, 인공지능 음향 재생 장치, 디지털 텔레비전, 가전기기, 기계 장치 및/또는 전자적으로 정보의 연산/처리 및 이와 관련된 제어가 가능하고 에너지 관리를 위해 특별히 제작된 적어도 하나의 장치를 포함할 수 있다.

[0031] 데이터 수신부(110)는 모델의 학습에 이용되는 데이터 및/또는 응답을 요하는 질의 데이터 등을 수신할 수 있다. 데이터 수신부(110)에 의해 수신된 데이터는 저장부(160)에 저장될 수 있다. 실시예에 따라, 데이터 수신부(110)는 유무선 통신망을 통하여 별도의 전자 장치 및/또는 서버 등으로부터 데이터를 수신할 수도 있다. 여기서 모델의 학습에 이용되는 데이터는 소정의 서비스를 제공하는 서버에 저장되어 있는 기존의 질의 및/또는 응답 데이터를 의미할 수 있다. 구체적으로, 질문 상담을 위한 게시판 데이터와 Q&A 게시판 데이터 등과 같이 소량의 정제되지 않은 기존 질의응답 데이터가 이용될 수 있다. 실시예에 따라, 모델 생성 장치(10)는 상기 서

버에 구현될 수도 있는 바, 이 경우에 데이터 수신부(110)는 생략될 수도 있다. 서버 내에 질의 및/또는 응답 데이터가 저장되어 있기 때문이다. 또한, 응답을 요하는 질의 데이터는 사용자 단말, 예컨대 스마트폰, 데스크탑 컴퓨터, 노트북 등과 같은 전자 장치로부터 수신될 수도 있다.

[0033] 데이터 변환부(120)는 질의응답 데이터를 변환할 수 있다. Q&A 게시판 데이터 등의 정제되지 않은 질의응답 데이터에는 하나의 질의응답 쌍에 여러 개의 질문과 이들 각각에 대한 대답이 혼재되어 있을 수 있다. 반면, 자동 질의응답 시스템에서 질문자가 제시하는 질문은 하나의 질문 내용으로 이루어지며, 답변 또한 이에 대한 하나의 답변 내용을 제시하여야 한다. 이에 따라, 기존 질의응답 데이터에 대한 적절한 데이터 변환이 필요하다.

[0034] 데이터 변환부(120)는 질문(또는 질의)과 답변(또는 응답)을 일정 단위로 분할하여 이들이 조합으로 만들어진 질의응답 쌍들을 생성함으로써 변환된 질의응답 데이터를 생성할 수 있다. 다시 말하면, 데이터 변환부(120)는 질문(또는 응답)에 하나의 질문(또는 응답)만이 포함되도록 일정 단위로 분할할 수 있다. 이때, 잘못 짝지어진 질의응답 쌍은 후술되는 데이터 정제부(140)에 의해 제거될 수 있다. 일정 단위의 예는 문장, 줄바꿈, 문단, 문장 부호 등이 될 수 있다. 예컨대, 데이터 변환부(120)는 질문(또는 답변)에 포함된 여러 문장을 문장 단위로 분할거나, 질문(또는 답변)을 줄바꿈을 기준으로 분할하거나, 질문(또는 답변)을 문단 단위로 분할하거나, 질문(또는 답변)을 문장 부호(예컨대, 물음표)를 기준으로 분할할 수 있다. 데이터 변환부(120)에 의해 변환된 데이터, 즉 질의응답 쌍들은 저장부(160)에 저장될 수 있다. 즉, 데이터 변환부(120)는 분할된 질문들과 분할된 답변들에 대한 카티전 곱(cartesian product)으로 만들어진 질의응답 쌍들을 생성할 수 있다. 도 3에는 데이터 변환 과정 및 데이터 정제 과정이 도시되어 있다.

[0036] 텍스트 전처리부(130)는 변환된 질의응답 데이터를 전처리할 수 있다. 구체적으로, 텍스트 전처리부(130)는 데이터 변환부(120)에 의해 생성된 질의응답 쌍들 에 포함된 질의와 응답에서 불완전한 음절로 구성된 음절들을 제거할 수 있다. 불완전한 음절의 예로 'ㅋㅋ', 'ㅠㅠ' 등이 있다. 후술할 TF-IDF(Term Frequency-Inverse Document Frequency) 코사인 유사도 기반 답변 후보군 선택 과정에서 각 단어의 문서 내 중요도만을 고려하며 단어의 의미를 고려하지 않는다. 따라서, 'ㅋㅋ', 'ㅠㅠ' 등의 불완전한 음절로 구성된 단어를 정제하지 않을 경우 해당 단어들의 문서 내 중요도를 높이 평가하여 답변 후보군 선택에 문제가 발생할 수 있다.

[0037] 또한, 텍스트 전처리부(130)는 질의자의 개인정보 및/또는 식별 정보(예컨대, 닉네임, 이름 등)가 포함되었을 가능성이 높은 어절을 삭제할 수 있다. 기존 질의응답 데이터에는 데이터의 특성상 질문자의 이름과 같이 질문자를 식별할 수 있는 식별 정보가 포함될 가능성이 크다. 그러한 정보가 챗봇 서비스의 답변으로 노출되는 것은 바람직하지 않으므로, 형태소 분석기를 활용하여 이름 등(식별 정보)이 포함된 것으로 추측되는 어절을 다음의 규칙 중 적어도 하나에 따라 삭제할 수 있다.

[0038] 1. '님', '고객님' 등 이름과 결합하는 접미사(이하, '키워드'라 지칭할 수 있음)가 포함된 어절 대상 형태 분석 수행

[0039] 2. 형태 분석 결과 키워드가 하나의 형태소로 분석될 경우 해당 어절 삭제

[0040] 3. 형태 분석 결과 알 수 없는 형태소로 분석된 형태소가 포함되어 있을 경우 해당 어절 삭제(어절의 형태 분석 실패로 인해 이름이 노출되는 문제를 방지하기 위함)

[0041] 4. 2 또는 3의 조건이 충족되었을 때, 앞 어절이 데이터셋 전체에서 일정 횟수 이상 출현하지 않는 어절일 경우 앞 어절 삭제(이름과 키워드가 분리된 경우의 처리를 위함)

[0042] 위와 같이, 텍스트 전처리부(130)는 적어도 하나의 규칙에 따라 사용자 식별 정보(예컨대, 이름)를 질의와 답변에서 삭제할 수 있다.

[0043] 텍스트 전처리부(130)에 의해 형태소 단위로 분석된 질의응답 데이터는 저장부(160)에 저장될 수 있다. 다만, 질의응답 도메인의 특성에 따라 학과명, 전형 이름 등과 같이 높은 동시 출현 빈도를 보이는 합성어 처리가 요구된다. 따라서, 텍스트 전처리부(130)는 비특허문헌 7에 제안된 방법으로 추출한 절(phrase)을 하나의 형태소로 취급하여 처리할 수 있다. 구체적으로 미리 정해진 빈도 이상의 동시 출현 빈도를 보이는 합성어가 하나의 형태소로 취급될 수 있다. 미리 정해진 빈도는 일정 횟수(예컨대, 3회)를 의미할 수도 있다. 도 4에는 텍스트 전처리부에 의한 텍스트 전처리 단계를 설명하기 위한 도면이 도시되어 있다.

[0045] 데이터 정제부(140)는 텍스트 전처리부(130)에 의해 전처리된 질의응답 데이터 중 무의미하거나 잘못된 데이터를 제거함으로써 데이터를 정제할 수 있다. 정제 단계에서는 효율성과 성능을 높이기 위해 형태소 분석 결과 명사 등 특정 품사의 형태소들만을 대상으로 진행될 수도 있다.

[0046] 구체적으로, 데이터 정제부(140)는 텍스트 랭크(TextRank) 알고리즘을 이용하여 무의미한 답변들을 제거할 수 있다. 텍스트 랭크 알고리즘은 텍스트에 대한 그래프 기반 중요도 평가 모델(ranking 모델)로, 각 문장을 정점으로 하고 문장 간 유사도로 간선을 구축하여 문서 내 문장의 중요도 평가를 수행한다. 데이터 정제부(140)는 질의응답 데이터의 답변들 전체에 대해 텍스트 랭크 알고리즘을 적용하여 각 답변의 중요도를 계산하고, 미리 정해진 중요도 값보다 낮은 값을 기록하는 답변을 무의미하거나 잘못된 데이터로 인식하여 제거할 수 있다. 다시 말하면, 답변들 중 미리 정해진 유사도 값보다 낮은 유사도 값을 갖는 답변을 무의미하거나 잘못된 데이터로 인식하여 제거할 수 있다. 예컨대, '메일로 답변드렸습니다'와 같이 다른 답변들과의 유사도가 낮은 무의미한 답변들이 제거될 수 있다. 텍스트 랭크 알고리즘에 대한 상세한 설명은 비특허문헌 8이 참조될 수 있다.

[0047] 또한, 데이터 정제부(140)는 질의-응답 매칭 필터(Question-Answer Matching Filter)를 이용하여 잘못 짝지어진 질의응답 쌍을 제거할 수 있다. 구체적으로, 데이터 정제부(140)는 질문과 답변의 유사도에 기초하여 데이터 변환 과정에서 잘못 짝지어진 질의응답 쌍을 제거할 수 있다. 일 예로, 데이터 정제부(140)는 질의응답 데이터에서 질문의 TF-IDF 특징 벡터와 답변의 TF-IDF 특징 벡터의 코사인 유사도(cosine similarity)를 계산할 결과가 일정 수준 이하일 경우, 질문과 답변의 키워드가 일치하지 않는 것으로 판단하여 제거할 수 있다. 도 5에는 질의-응답 매칭 필터를 설명하기 위한 도면이 도시되어 있다.

[0048] TF-IDF는 각 단어가 특정 문서(본 발명에서는 질문 또는 답변) 내에서 갖는 중요도를 나타내는 통계적 수치로 정보검색 분야의 다양한 과제에 활용되고 있다. 여러 단어들의 TF-IDF 값을 특정 문서를 나타내는 특징 벡터로 사용한 뒤 문서의 특징 벡터 간 코사인 유사도를 계산함으로써, 간단한 연산을 통해 빠르게 문서 간 유사도를 계산할 수 있다. 즉, 본 발명에서는 질문(또는 답변)을 이루는 각 단어의 TF-IDF 값을 질문(또는 답변)의 특징 벡터로 사용함으로써, 질문과 답변의 코사인 유사도를 계산할 수 있다.

[0050] 모델 학습부(150)는 데이터 정제부(140)에 의해 정제된 데이터를 이용하여 질의응답 모델을 학습시킬 수 있다. 즉, 모델 학습부(150)는 COALA(COmpare-Aggregate for Long Answers)를 학습시킴으로써 질의응답 모델을 생성할 수 있다. COALA는 질문의 각 내용이 답변에 모두 포함되어 있는지 판단하는 방식으로 모델링되었으며 기존의 일반적인 딥러닝 기반 챗봇 기술들과 비교하여 구조를 크게 단순화한 딥러닝 기반 답변 선택 모델이다. 이에 따라 요구되는 학습 데이터의 수가 적으며 답변 텍스트의 길이가 긴 질의응답 상황의 처리에 용이하다. COALA에 대한 상세한 설명에 관하여는 비특허문헌 9가 참조될 수 있다.

[0052] 저장부(160)에는 모델의 학습이 이용되는 데이터, 응답을 요하는 질의 데이터, 데이터 변환부(120)에 의하여 분할된 질문(또는 응답) 데이터, 텍스트 전처리부(130)에 의해 전처리된 질의응답 데이터, 데이터 정제부(140)에 의해 정제된 데이터, 모델 학습부(150)에 의해 학습된(또는 생성된) 질의응답 모델 등이 저장될 수 있다.

[0054] 도 6은 본 발명의 일 실시예에 따른 자동 질의응답 장치(20)의 기능 블럭도이다.

[0055] 도 6을 참조하면, 자동 질의응답 장치(20)는 사용자(질의자 또는 질문자라는 용어가 사용될 수도 있음) 단말로부터 질의를 수신하고 이에 대한 응답(답변)을 상기 사용자 단말로 송신할 수 있다. 자동 질의응답 장치(20)는 데이터 수신부(210), 텍스트 전처리부(230), 답변 후보군 선택부(240), 및 최종 답변 선택부(250)를 포함한다. 실시예에 따라, 자동 질의응답 장치(20)는 저장부(260)를 더 포함할 수도 있다.

[0056] 자동 질의응답 장치(20)는 적어도 하나의 프로세서를 이용하여 구현될 수 있으며, 여기서, 프로세서는 중앙 처리 장치(CPU, Central Processing Unit), 마이크로 컨트롤러 유닛(MCU, Micro Controller Unit), 애플리케이션 프로세서(AP, Application Processor), 마이컴(Micom, Micro Processor), 전자 제어 유닛(ECU, Electronic Controlling Unit) 및/또는 각종 연산 처리 및 제어 신호의 생성이 가능한 다른 전자 장치 등을 포함할 수 있다. 이들 장치는, 예를 들어, 하나 또는 둘 이상의 반도체 칩 및 관련 부품을 이용하여 구현될 수 있다. 일 실시예에 의하면, 프로세서는 저장부(260)에 저장된 적어도 하나의 애플리케이션(소프트웨어, 프로그램이나 앱

등으로 표현 가능하다)을 구동시켜, 미리 정의된 연산, 판단, 처리 및/또는 제어 동작 등을 수행할 수도 있다. 여기서, 저장부(260)에 저장된 애플리케이션은, 설계자에 의해 직접 작성되어 저장부(260)에 입력 및 저장된 것일 수도 있고, 또는 유선 또는 무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망을 통하여 획득 또는 갱신된 것일 수도 있다.

[0057] 또한, 자동 질의응답 장치(20)는 이와 같은 프로세서가 하나 이상 설치된 적어도 하나의 정보 처리 장치를 이용하여 구현될 수도 있으며, 전자 정보 처리 장치는 데스크톱 컴퓨터, 랩톱 컴퓨터, 서버용 컴퓨터, 스마트 폰, 태블릿 피씨, 스마트 시계, 내비게이션 장치, 휴대용 게임기, 두부 장착형 디스플레이(HMD, Head Mounted Display) 장치, 인공지능 음향 재생 장치, 디지털 텔레비전, 가전기기, 기계 장치 및/또는 전자적으로 정보의 연산/처리 및 이와 관련된 제어가 가능하고 에너지 관리를 위해 특별히 제작된 적어도 하나의 장치를 포함할 수 있다.

[0059] 데이터 수신부(210)는 유무선 통신망을 이용하여 사용자 단말로부터 질의를 수신할 수 있다. 데이터 수신부(210)에 의해 수신된 질의는 저장부(260)에 저장될 수 있다.

[0061] 텍스트 전처리부(230)는 사용자 단말로부터 수신된 질의를 전처리할 수 있다. 구체적으로, 텍스트 전처리부(230)는 불완전한 음절로 구성된 음절들을 질의에서 제거할 수 있다. 또한, 텍스트 전처리부(230)는 질의자의 개인정보 및/조는 식별 정보(예컨대, 닉네임, 이름 등)가 포함되었을 가능성이 높은 어절을 삭제할 수 있다. 텍스트 전처리부(230)에 의해 형태소 단위로 분석된 질의는 저장부(260)에 저장될 수 있다.

[0062] 또한, 질의응답 도메인의 특성에 따라 학과명, 전형 이름 등과 같이 높은 동시 출현 빈도를 보이는 합성어 처리가 요구된다. 따라서, 텍스트 전처리부(230)는 미리 정해진 빈도 이상의 동시 출현 빈도를 보이는 합성어를 하나의 형태소로 취급할 수 있다. 미리 정해진 빈도는 일정 횟수(예컨대, 3회)를 의미할 수도 있다.

[0064] 답변 후보군 선택부(240)는 텍스트 전처리부(230)에 의해 전처리된 질의를 이용하여 미리 저장되어 있는 답변들 중에서 답변 후보군을 선택할 수 있다. 즉, 문서 유사도에 기반하여 새로운 질문에 대하여 기존 질의응답 데이터 전체로부터 답변 후보군으로 사용할 데이터의 부분집합을 선택한다. 이후, 후술되는 최종 답변 선택부(250)에 의해 답변 후보군 중 적어도 하나(예컨대, 5개)의 답변이 선택되고, 선택된 답변은 사용자 단말로 송신될 수 있다.

[0065] 새로 주어진 질문과 유사한 기존 질문에 대한 답변은 새로운 질문에 대한 적합한 답변일 가능성이 높다. 이에 따라, 답변 후보군 선택을 위해 새로 주어진 질문과 기존 질의응답 데이터 질문의 TF-IDF의 특징 벡터 간 코사인 유사도를 계산하여 주어진 질문과 유사한 기존 질문들을 선택하고, 이 질문들에 대한 기존 답변 상위 n(n은 2 이상의 자연수) 개를 답변 후보군으로 선택할 수 있다. 이를 위해, 답변 후보군 선택부(240)는 사용자 단말로부터 수신된 질문을 이루는 각 단어의 TF-IDF 값을 질문의 특징 벡터로 사용함으로써, 새롭게 수신된 질문과 기존의 질문들 각각 과의 코사인 유사도를 계산할 수 있다. 결국, 유사도가 높은 순으로 상위 n 개(예컨대, 50개)의 기존 질문 각각에 대한 답변을 답변 후보군으로 선택할 수 있다. 도 7에는 TF-IDF 코사인 유사도 기반 답변 후보군 선택 과정을 설명하기 위한 도면이 도시되어 있다.

[0067] 최종 답변 선택부(250)는 도 1을 통해 설명한 모델 생성 장치(10)에 의해 생성된 질의응답 모델을 이용하여 최종 답변을 선택할 수 있다. 예컨대, 최종 답변 선택부(250)는 COALA를 이용하여 답변 후보군으로부터 가장 적절한 적어도 하나의 답변을 선택할 수 있다.

[0068] 수학식 1은 COALA의 기본적인 구조를 나타낸 식이다.

[0069] [수학식 1]

[0070] $$f(Q,A) = \Omega(\Phi(Q),\Phi(A))$$

[0071] 수학식 1에서, Q는 질문 텍스트, A는 답변 텍스트를 나타낸다. 그리고, f 는 A가 Q의 적합한 답변일 점수

를 구하는 함수이며, COALA에서는 이를 텍스트의 특징을 추출하는 함수 Φ와 텍스트 간 연관성을 평가하는 함수 Ω의 결합으로 모델링한다. COALA의 전체 구조도는 도 8에 도시되어 있다.

[0072] COALA는 크게 'Aspect Identification', 'Relevance Matching', 'Aggregation'의 세 부분으로 나뉜다.

[0073] Aspect Identification 단계(제1 단계)에서는 합성곱 연산(convolution operation)을 통해 질문 텍스트와 답변 후보 텍스트의 각 n-gram 부분들의 지역적 문맥을 특징으로 추출한다. 텍스트 전체 문맥을 파악하는 대신 여러 부분의 지역적 문맥을 파악하도록 모델링함으로써 긴 답변 텍스트에 대한 표현력을 높이고 요구되는 학습 데이터 양을 절감할 수 있다. 이를 수식으로 표현하면 수학식 2와 같다.

[0074] [수학식 2]

[0075] $$\Phi(Q) = CNN(Q)$$
$$\Phi(A) = CNN(A)$$

[0076] Relevance Matching(제2 단계)에서는 질문 텍스트와 답변 텍스트의 연관성을 평가하기 위하여 행렬곱 연산을 통해 'Aspect Identification'에서 추출한 지역적 특징 간의 유사도를 계산한다. 이를 수식으로 표현하면 수학식 3과 같다.

[0077] [수학식 3]

[0078] $$H = \Phi(Q)\Phi(A)^T$$

[0079] H의 i번째 행, j번째 열의 원소 H_{ij}는 Q의 i번째 지역적 특징과 A의 j번째 지역적 특징의 유사도를 나타낸다. H의 모든 행에서 각각의 최댓값을 선택함으로써 질문 텍스트의 각 내용이 답변 텍스트에 나타나는 정도를 나타내는 벡터 c를 계산한다. 이를 수식으로 표현하면 수학식 4와 같다.

[0080] [수학식 4]

[0081] $$[c]_i = \max_j([H]_{i,j})$$

[0082] Aggregation(제3 단계)에서는 'Relevance Matching'에서 계산한 c로부터 질문 텍스트 전체에 대한 답변 텍스트의 연관성을 평가하기 위해 p-mean(power-mean, 멱평균) 기반의 aggregation을 수행한다. 이를 수식으로 표현하면 수학식 5와 같다.

[0083] [수학식 5]

[0084] $$powermean(x,p) = \left(\frac{[x]_1^p + \cdots + [x]_n^p}{n}\right)^{1/p}$$

[0085] 더불어 COALA 모델에서는 c로부터 상호보완 가능한 여러 aggregation 결과값을 활용하기 위하여, 여러 p 값에 대한 p-mean 계산 결과를 신경망에 입력하여 최종 연관성 평가 점수를 계산한다.

[0086] 또한, 최종 답변 선택부(250)는 최종적으로 선택된 적어도 하나의 답변을 사용자 단말로 송신할 수 있다.

[0088] 저장부(260)에는 데이터 수신부(210)에 의해 수신된 질의, 텍스트 전처리부(230)에 의해 전처리된 질의, 답변 후보군 선택부(240)에 의해 선택된 답변 후보군, 최종 답변 선택부(250)에 의해 선택된 적어도 하나의 답변, 최종 답변 선택부(250)가 이용하는 자동 응답 모델 등이 저장될 수 있다.

[0090] 도 9는 본 발명의 다른 실시예에 따른 자동 질의응답 장치의 기능 블럭도이다. 도 9에 도시된 자동 질의응답 장치를 설명함에 있어 앞선 기재와 중복되는 내용에 관하여는 그 기재를 생략하기로 한다.

[0091] 도 9를 참조하면, 자동 질의응답 장치(30)는 사용자 단말로부터 질의를 수신하고 이에 대한 응답(답변)을 상기 사용자 단말로 송신할 수 있다. 자동 질의응답 장치(20)는 데이터 수신부(310), 모델 생성부(340), 및 질의 응답부(350)를 포함한다. 실시예에 따라, 자동 질의응답 장치(30)는 저장부(360)를 더 포함할 수도 있다.

[0092] 자동 질의응답 장치(30)는 적어도 하나의 프로세서를 이용하여 구현될 수 있으며, 여기서, 프로세서는 중앙 처리 장치(CPU, Central Processing Unit), 마이크로 컨트롤러 유닛(MCU, Micro Controller Unit), 애플리케이션 프로세서(AP, Application Processor), 마이컴(Micom, Micro Processor), 전자 제어 유닛(ECU, Electronic Controlling Unit) 및/또는 각종 연산 처리 및 제어 신호의 생성이 가능한 다른 전자 장치 등을 포함할 수 있다. 이들 장치는, 예를 들어, 하나 또는 둘 이상의 반도체 칩 및 관련 부품을 이용하여 구현될 수 있다. 일 실시예에 의하면, 프로세서는 저장부(360)에 저장된 적어도 하나의 애플리케이션(소프트웨어, 프로그램이나 앱 등으로 표현 가능하다)을 구동시켜, 미리 정의된 연산, 판단, 처리 및/또는 제어 동작 등을 수행할 수도 있다. 여기서, 저장부(360)에 저장된 애플리케이션은, 설계자에 의해 직접 작성되어 저장부(360)에 입력 및 저장된 것일 수도 있고, 또는 유선 또는 무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망을 통하여 획득 또는 갱신된 것일 수도 있다.

[0093] 또한, 자동 질의응답 장치(30)는 이와 같은 프로세서가 하나 이상 설치된 적어도 하나의 정보 처리 장치를 이용하여 구현될 수도 있으며, 전자 정보 처리 장치는 데스크톱 컴퓨터, 랩톱 컴퓨터, 서버용 컴퓨터, 스마트 폰, 태블릿 피씨, 스마트 시계, 내비게이션 장치, 휴대용 게임기, 두부 장착형 디스플레이(HMD, Head Mounted Display) 장치, 인공지능 음향 재생 장치, 디지털 텔레비전, 가전기기, 기계 장치 및/또는 전자적으로 정보의 연산/처리 및 이와 관련된 제어가 가능하고 에너지 관리를 위해 특별히 제작된 적어도 하나의 장치를 포함할 수 있다.

[0094] 데이터 수신부(310)는 모델의 학습에 이용되는 데이터 및/또는 응답을 요하는 질의 데이터 등을 수신할 수 있다. 수신된 데이터는 저장부(360)에 저장될 수 있다.

[0095] 모델 생성부(340)는 자동 질의응답 모델을 생성할 수 있다. 실시예에 따라, 모델 생성부(340)는 데이터 변환부, 텍스트 전처리부, 데이터 정제부, 및 모델 학습부를 포함할 수 있는데, 각 구성의 기능은 도 1에 도시된 모델 생성 장치(10)의 구성의 기능과 동일할 수 있다.

[0096] 질의 응답부(350)는 사용자 단말로부터 수신된 질의에 대한 적어도 하나의 응답을 선택하고, 선택된 응답을 사용자 단말로 송신할 수 있다. 실시예에 따라, 질의 응답부(350)는 텍스트 전처리부, 답변 후보군 선택부, 및 최종 답변 선택부가 포함할 수 있는데, 각 구성의 기능은 도 6에 도시된 자동 질의응답 장치(20)의 구성의 기능과 동일할 수 있다.

[0097] 저장부(360)에는 데이터 수신부(310)에 의해 수신된 데이터, 모델 생성부(340)에 의해 생성된 자동 질의응답 모델, 모델 생성 중에 생성된 데이터, 질의 응답부(350)에 의해 선택된 답변 후보군, 최종 답변 등이 저장될 수 있다.

[0099] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시 예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0100] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서,

분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0101] 실시 예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 좋ㅂ하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시 예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시 예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0103] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0105] 10 : 모델 생성 장치

20, 30 : 자동 질의응답 장치

110, 210, 310 : 데이터 수신부

120 : 데이터 변환부

130, 230 : 텍스트 전처리부

140 : 데이터 정제부

150 : 모델 학습부

160, 260, 360 : 저장부

240 : 답변 후보군 선택부

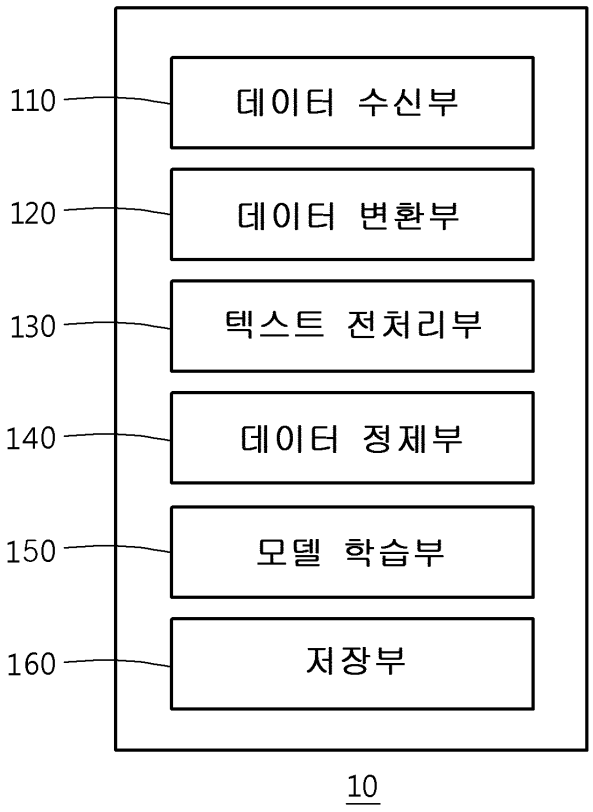
250 : 최종 답변 선택부

340 : 모델 생성부

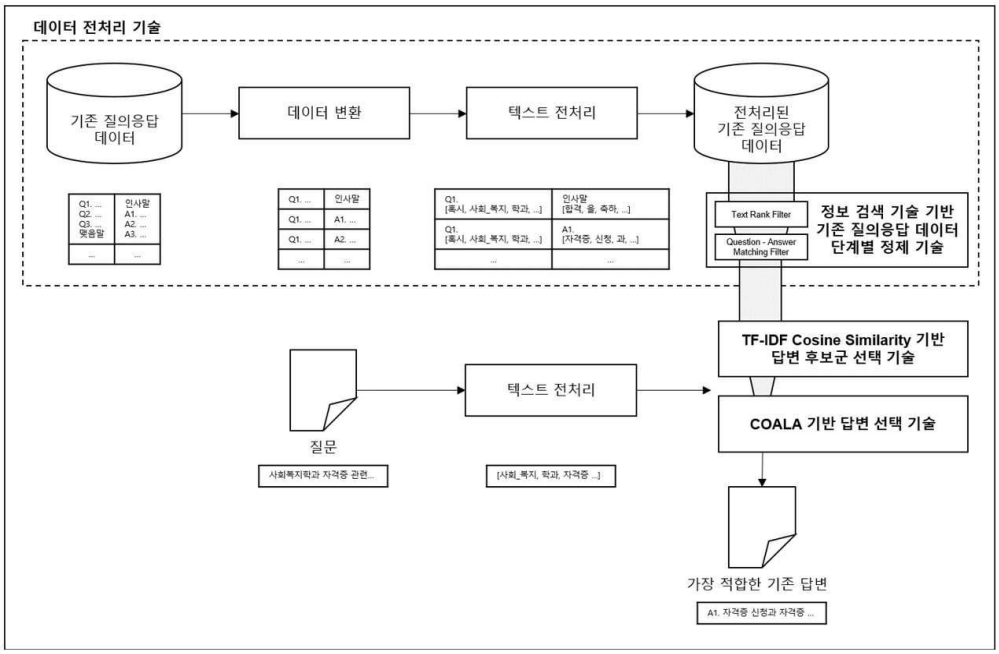
350 : 질의 응답부

도면

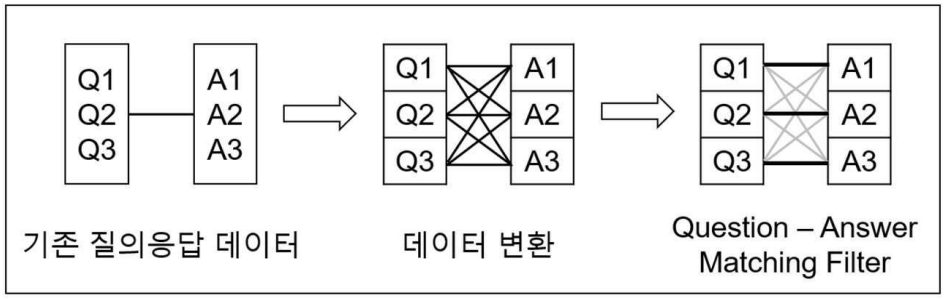
도면1


110
데이터 수신부
120
데이터 변환부
130
텍스트 전처리부
140
데이터 정제부
150
모델 학습부
160
저장부
10

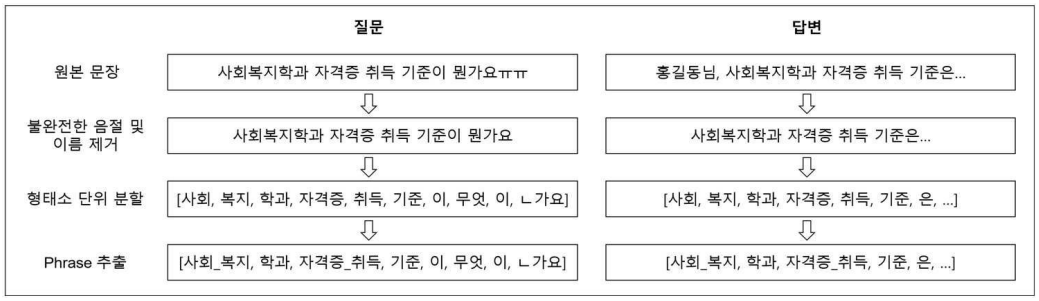
도면2


데이터 전처리 기술
기존 질의응답 데이터
데이터 변환
텍스트 전처리
전처리된 기존 질의응답 데이터
Text Rank Filter
Question - Answer Matching Filter
정보 검색 기술 기반 기존 질의응답 데이터 단계별 정제 기술
TF-IDF Cosine Similarity 기반 답변 후보군 선택 기술
COALA 기반 답변 선택 기술
텍스트 전처리
질문
가장 적합한 기존 답변

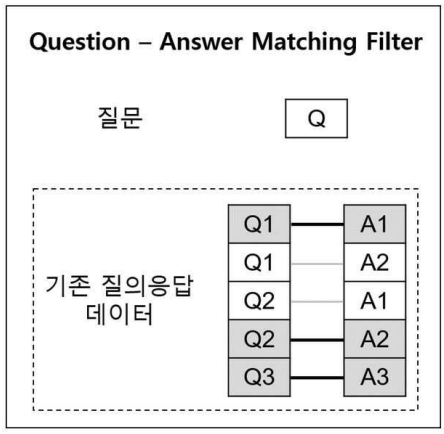
도면3


Q1
Q2
Q3
A1
A2
A3
기존 질의응답 데이터
데이터 변환
Question – Answer Matching Filter

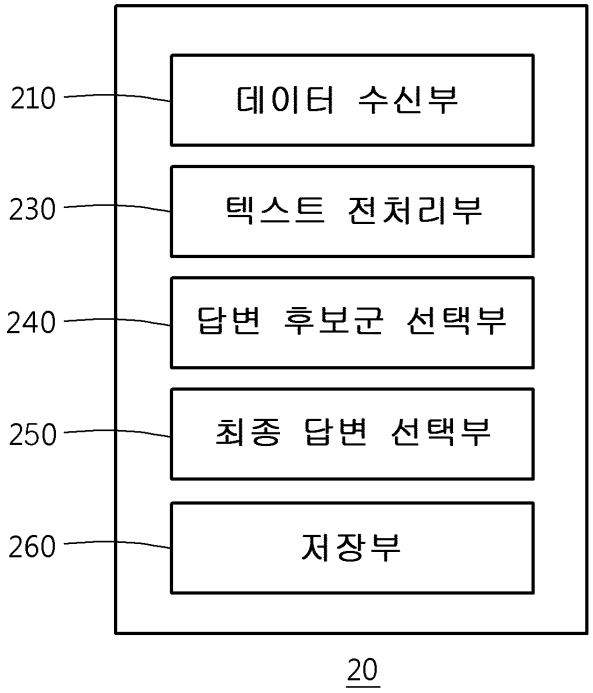
도면4


질문
답변
원본 문장
사회복지학과 자격증 취득 기준이 뭔가요ㅠㅠ
홍길동님, 사회복지학과 자격증 취득 기준은...
불완전한 음절 및 이름 제거
사회복지학과 자격증 취득 기준이 뭔가요
사회복지학과 자격증 취득 기준은...
형태소 단위 분할
[사회, 복지, 학과, 자격증, 취득, 기준, 이, 무엇, 이, ㄴ가요]
[사회, 복지, 학과, 자격증, 취득, 기준, 은, ...]
Phrase 추출
[사회_복지, 학과, 자격증_취득, 기준, 이, 무엇, 이, ㄴ가요]
[사회_복지, 학과, 자격증_취득, 기준, 은, ...]

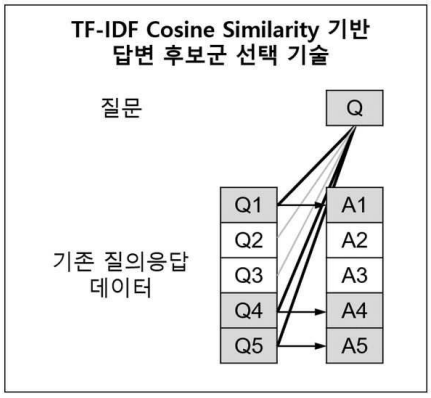
도면5


Question – Answer Matching Filter
질문
Q
기존 질의응답 데이터
Q1
Q1
Q2
Q2
Q3
A1
A2
A1
A2
A3

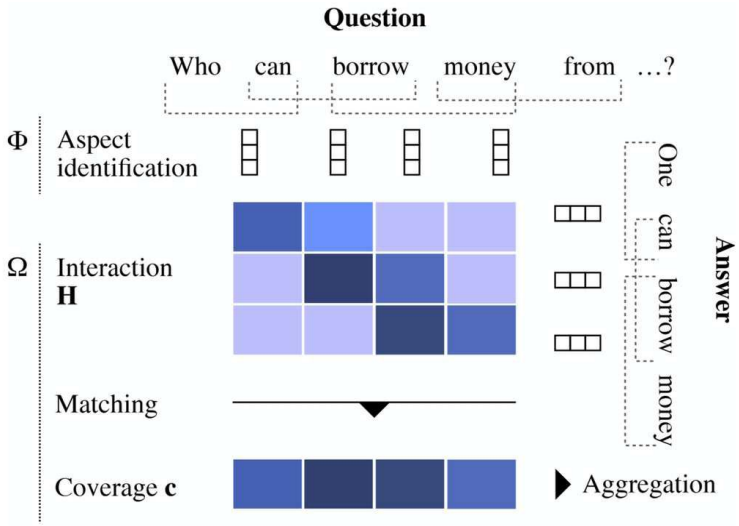
도면6


210
데이터 수신부
230
텍스트 전처리부
240
답변 후보군 선택부
250
최종 답변 선택부
260
저장부
20

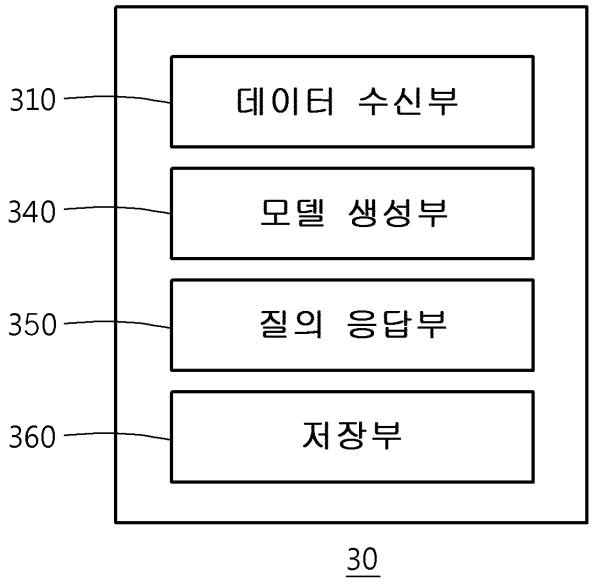
도면7


TF-IDF Cosine Similarity 기반
답변 후보군 선택 기술
질문
Q
기존 질의응답
데이터
Q1
Q2
Q3
Q4
Q5
A1
A2
A3
A4
A5

도면8


Question
Who
can
borrow
money
from
…?
Φ
Aspect identification
Ω
Interaction H
Matching
Coverage c
Aggregation
Answer
One
can
borrow
money

도면9


310
데이터 수신부
340
모델 생성부
350
질의 응답부
360
저장부
30