(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0048368
(43) 공개일자 2021년05월03일

(51) 국제특허분류(Int. Cl.)
G06F 16/33 (2019.01) *G06F 16/338* (2019.01)
G06F 40/20 (2020.01)
(52) CPC특허분류
G06F 16/3344 (2019.01)
G06F 16/338 (2019.01)
(21) 출원번호 10-2019-0132572
(22) 출원일자 2019년10월23일
심사청구일자 2019년10월23일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
박성배
서울특별시 송파구 오금로 396 에스케이허브파크 396동 1402호
김소언
경기도 수원시 영통구 덕영대로 1705, 1101호
(뒷면에 계속)
(74) 대리인
특허법인지원

전체 청구항 수 : 총 16 항

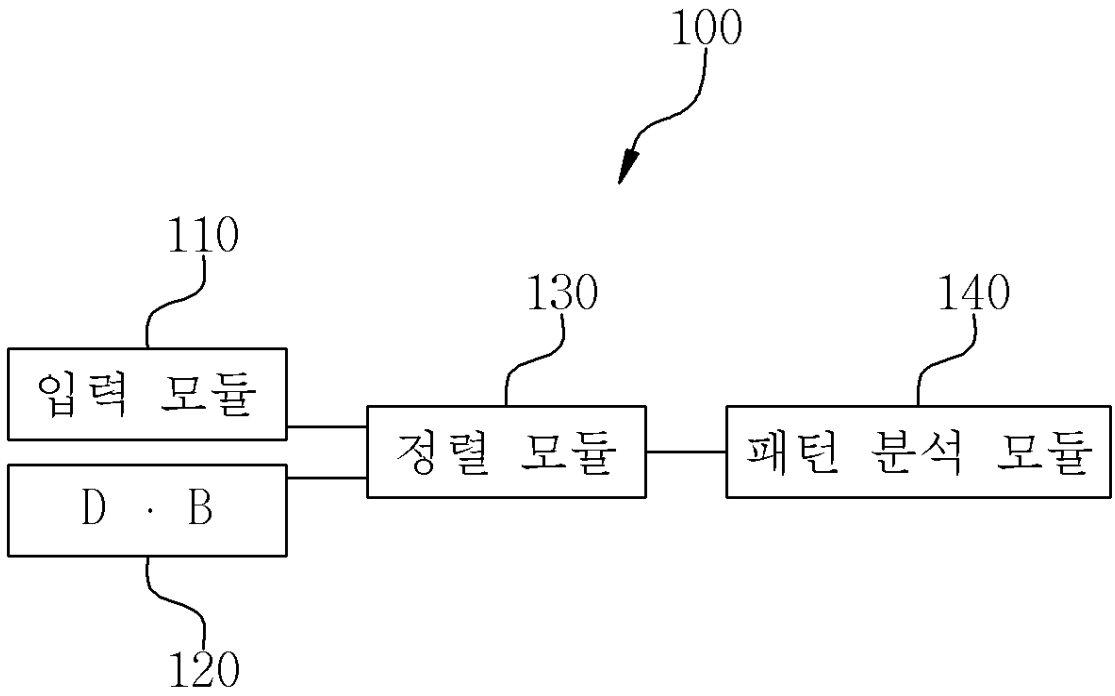
(54) 발명의 명칭 **유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법**

(57) 요 약

본 발명은 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법에 관한 것으로서, 상기 유사 문장 검색 시스템은 사용자로부터 분석 대상 문장을 입력받는 입력모듈과, 다수의 비교 대상 문장이 저장된 데이터 베이스와, 상기 입력모듈 및 데이터 베이스에서 제공되는 상기 분석 대상 문장 또는 비교 대상 문장의 형태소를 분석하

(뒷면에 계속)

대 표 도 - 도1



는 정렬모듈과, 상기 정렬모듈에서 제공되는 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 패턴 분석 모듈을 구비한다.

본 발명에 따른 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법은 문장의 형태소를 기준으로 문장패턴을 분석하여 분석 대상 문장과 유사한 비교 대상 문장을 선정하므로 사용자가 입력한 문장과 유사율이 비교적 높은 문장을 선정할 수 있다는 장점이 있다.

(52) CPC특허분류
G06F 40/205 (2020.01)
G06F 40/268 (2020.01)

(72) 발명자
김아영
경기도 성남시 분당구 성남대로171번길 8 청솔마을 서광영남아파트 105동 1402호

박형준
경기도 수원시 영통구 매영로 432, 506호

이 발명을 지원한 국가연구개발사업

과제고유번호	1711081022
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	지원사업
연구과제명	WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발
기 여 율	1/1
과제수행기관명	경희대학교 산학협력단
연구기간	2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

사용자로부터 분석 대상 문장을 입력받는 입력모듈;

다수의 비교 대상 문장이 저장된 데이터 베이스;

상기 입력모듈 및 데이터 베이스에서 제공되는 상기 분석 대상 문장 또는 비교 대상 문장의 형태소를 분석하는 정렬모듈; 및

상기 정렬모듈에서 제공되는 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 패턴 분석 모듈;을 구비하는,

유사문장 검색 시스템.

청구항 2

제1항에 있어서,

상기 정렬모듈은 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소의 유무를 분석하고, 분석 결과에 따라 상기 분석 대상 문장 및 비교 대상 문장들을 정렬하는,

유사문장 검색 시스템.

청구항 3

제2항에 있어서,

상기 정렬모듈은

상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소에 대응되게 가산점을 산출하고, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 불일치하는 형태소에 대응되게 패널티 점수를 산출하되, 상기 가산점 및 패널티 점수를 토대로 상기 비교 대상 문장들을 정렬하는,

유사문장 검색 시스템.

청구항 4

제1항에 있어서,

상기 패턴 분석 모듈은

상기 비교 대상 문장에 포함된 어휘 또는 형태소들 중 상기 분석 대상 문장의 어휘 또는 형태소와 일치하는 어휘 및 형태소의 갯수에 따라 가산점을 부여하고, 부여된 가산점을 합산하여 해당 비교 대상 문장의 유사도를 산출하는,

유사문장 검색 시스템.

청구항 5

제4항에 있어서,

상기 패턴 분석 모듈은

상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 어말어미 또는 선어말어미이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사문장 검색 시스템.

청구항 6

제4항에 있어서,

상기 패턴 분석 모듈은

상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 접미사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 시스템.

청구항 7

제4항에 있어서,

상기 패턴 분석 모듈은

상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사 또는 인용격 조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 시스템.

청구항 8

제4항에 있어서,

상기 패턴 분석 모듈은

상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치하고, 상기 비교 대상 문장의 어휘가 상기 분석 대상 문장의 어휘와 일치할 경우, 해당 형태소가 보조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 시스템.

청구항 9

사용자에 의해 입력모듈에 입력된 분석 대상 문장과, 데이터 베이스에 저장된 비교 대상 문장의 형태소를 분석하는 정렬단계; 및

상기 정렬단계에서 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 패턴 분석 단계;를 포함하는,

유사 문장 검색 방법.

청구항 10

제9항에 있어서,

상기 정렬단계에서는, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소의 유무를 분석하고, 분석 결과에 따라 상기 분석 대상 문장 및 비교 대상 문장들을 정렬하는,

유사 문장 검색 방법.

청구항 11

제10항에 있어서,

상기 정렬단계에서는, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소에 대응되게 가산점을 산출하고, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 불일치하는 형태소에 대응되게 패널티 점수를 산출하되, 상기 가산점 및 패널티 점수를 토대로 상기 비교 대상 문장들을 정렬하는,

유사 문장 검색 방법.

청구항 12

제9항에 있어서,

상기 패턴 분석 단계에서는, 상기 비교 대상 문장에 포함된 어휘 및 형태소들 중 상기 분석 대상 문장의 어휘 및 형태소와 일치하는 어휘 및 형태소의 갯수에 따라 가산점을 부여하고, 부여된 가산점을 합산하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 방법.

청구항 13

제12항에 있어서,

상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 어말어미 또는 선어말어미이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 방법.

청구항 14

제12항에 있어서,

상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 접미사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 방법.

청구항 15

제12항에 있어서,

상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치하고, 상기 비교 대상 문장의 어휘가 상기 분석 대상 문장의 어휘와 일치할 경우, 해당 형태소가 보조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 방법.

청구항 16

제12항에 있어서,

상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우 해당 형태소가 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사 또는 인용격 조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출하는,

유사 문장 검색 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법에 관한 것으로서, 더욱 상세하게는 형태소를 기준으로 문장패턴을 분석하여 분석 대상 문장과 유사한 비교 대상 문장을 선정하는 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법에 관한 것이다.

배 경 기 술

[0002] 사용자가 입력한 문장에 대해 적절한 응답을 하는 대화 모델은 크게 검색 기반과 생성 기반으로 나뉠 수 있다. 그 중 검색을 기반으로하는 대화 모델의 경우, 분석 대상 문장과 가장 유사한 발화를 검색하는 방법이 대화 모델의 성능과 직결된다.

[0003] 효과적인 응답 생성을 위해 유사 문장을 검색하는 기술 중 하나는 어휘적으로 비슷한 문장을 찾는 것이다. 이러한 방법으로 두 문장이 동일하게 가지고 있는 n-gram의 수를 측정하는 방법인 'BLEU(Bilingual Evaluation Understudy)'와 색인 기반의 검색엔진 'Lucene' 등이 있다. 하지만 이러한 방법은 짧은 문장에 비정상적으로 높은 점수가 부여되는 문제와 재현율이 낮더라도 유사도가 높다고 계산되는 문제가 존재한다.

[0004] 다른 기술로는 의미적으로 비슷한 문장을 찾는 것인데, 'Semantic Roles'가 이에 해당한다. 두 문장에 대해 'Semantic Role'을 생성하여 얼마나 유사한지 측정하는 방식이지만 'Semantic Role'의 한국어에 대한 표준 기술이 없고 성능이 매우 낮다는 문제가 존재한다.

[0005] 또 다른 기술로는 구조적으로 비슷한 문장을 찾는 것인데, 두 문장에 각각에 대한 구문 트리를 생성하고, 구문 트리 커널을 통해 트리 간 유사도를 측정하는 방식이다. 하지만 이 방법을 통해선 의미적 유사도가 고려되지 않아 대화 모델의 성능이 낮다는 문제가 존재한다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 등록특허공보 제10-1224660호: 유사 문장 검색 장치, 검색방법, 저장 매체, 자동 대화서비스 시스템 및 서비스 방법

발명의 내용

해결하려는 과제

[0007] 본 발명은 상기한 문제점을 해결하기 위해 창안된 것으로서, 형태소를 기준으로 문장패턴을 분석하여 분석 대상 문장과 유사한 비교 대상 문장을 선정하는 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법을 제공하는데 그 목적이 있다.

과제의 해결 수단

[0008] 상기 목적을 달성하기 위한 본 발명에 따른 유사 문장 검색 시스템은 사용자로부터 분석 대상 문장을 입력받는 입력모듈과, 다수의 비교 대상 문장이 저장된 데이터 베이스와, 상기 입력모듈 및 데이터 베이스에서 제공되는 상기 분석 대상 문장 또는 비교 대상 문장의 형태소를 분석하는 정렬모듈과, 상기 정렬모듈에서 제공되는 분석 정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 패턴 분석 모듈을 구비한다.

[0009] 상기 정렬모듈은 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소의 유무를 분석하고, 분석 결과에 따라 상기 분석 대상 문장 및 비교 대상 문장들을 정렬한다.

[0010] 상기 정렬모듈은 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소에 대응되게 가산점을 산출하고, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 불일치하는 형태소에 대응되게 패널티 점수를 산출하되, 상기 가산점 및 패널티 점수를 토대로 상기 비교 대상 문장들을 정렬한다.

[0011] 상기 패턴 분석 모듈은 상기 비교 대상 문장에 포함된 어휘 또는 형태소들 중 상기 분석 대상 문장의 어휘 또는 형태소와 일치하는 어휘 및 형태소의 갯수에 따라 가산점을 부여하고, 부여된 가산점을 합산하여 해당 비교 대상 문장의 유사도를 산출한다.

[0012] 상기 패턴 분석 모듈은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 어말어미 또는 선어말어미이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0013] 상기 패턴 분석 모듈은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 접미사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0014] 상기 패턴 분석 모듈은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사 또는 인용격 조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0015] 상기 패턴 분석 모듈은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치하고, 상기 비교 대상 문장의 어휘가 상기 분석 대상 문장의 어휘와 일치할 경우, 해당 형태소가 보조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0016] 한편, 본 발명에 따른 유사 문장 검색 방법은 사용자에 의해 입력모듈에 입력된 분석 대상 문장과, 데이터 베이스에 저장된 비교 대상 문장의 형태소를 분석하는 정렬단계와, 상기 정렬단계에서 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 패턴 분석 단계를 포함한다.

[0017] 상기 정렬단계에서는, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소의 유무를 분석하고, 분석 결과에 따라 상기 분석 대상 문장 및 비교 대상 문장들을 정렬한다.

[0018] 상기 정렬단계에서는, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소에 대응되게 가산점을 산출하고, 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 불일치하는 형태소에 대응되게 패널티 점수를 산출하되, 상기 가산점 및 패널티 점수를 토대로 상기 비교 대상 문장들을 정렬할 수 있다.

[0019] 상기 패턴 분석 단계에서는, 상기 비교 대상 문장에 포함된 어휘 및 형태소들 중 상기 분석 대상 문장의 어휘 및 형태소와 일치하는 어휘 및 형태소의 갯수에 따라 가산점을 부여하고, 부여된 가산점을 합산하여 해당 비교 대상 문장의 유사도를 산출할 수 있다.

[0020] 상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 어말어미 또는 선어말어미이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0021] 상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 접미사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0022] 상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치하고, 상기 비교 대상 문장의 어휘가 상기 분석 대상 문장의 어휘와 일치할 경우, 해당 형태소가 보조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

[0023] 상기 패턴 분석 단계에서, 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사 또는 인용격 조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출할 수도 있다.

발명의 효과

[0024] 본 발명에 따른 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법은 문장의 형태소를 기준으로 문장 패턴을 분석하여 분석 대상 문장과 유사한 비교 대상 문장을 선정하므로 사용자가 입력한 문장과 유사율이 비교적 높은 문장을 선정할 수 있다는 장점이 있다.

도면의 간단한 설명

[0025] 도 1은 본 발명에 따른 유사 문장 검색 시스템에 대한 블럭도이고,

도 2는 본 발명에 따른 유사 문장 검색 방법에 대한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0026] 이하, 첨부한 도면을 참조하여 본 발명의 실시예에 따른 유사 문장 검색 시스템 및 이를 이용한 유사 문장 검색 방법에 대해 설명한다. 본 발명은 다양한 변경을 가할 수 있고 여러 가지 형태를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 본문에 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 개시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다. 첨부된 도면에 있어서, 구조물들의 치수는 본 발명의 명확성을 기하기 위하여 실제보다 확대하여 도시한 것이다.

[0027] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다.

[0028] 본 출원에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 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.

[0029] 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 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0030] 도 1에는 본 발명에 따른 유사 문장 검색 시스템(100)이 도시되어 있다.

[0031] 도면을 참조하면, 상기 유사 문장 검색 시스템(100)은 사용자로부터 분석 대상 문장을 입력받는 입력모듈(110)과, 다수의 비교 대상 문장이 저장된 데이터 베이스(120)와, 상기 입력모듈(110) 및 데이터 베이스(120)에서 제공되는 상기 분석 대상 문장 또는 비교 대상 문장의 형태소를 분석하는 정렬모듈(130)과, 상기 정렬모듈(130)에서 제공되는 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 패턴 분석 모듈(140)을 구비한다.

[0032] 입력모듈(110)은 사용자로부터 분석 대상 문장을 입력받는 것으로서, 키보드, 키패드, 터치패널 등과 같은 입력수단이거나 사용자의 음성을 인식하여 해당 음성에 포함된 문장을 입력받는 음성인식수단이 적용될 수 있다.

[0033] 데이터 베이스(120)는 다수의 비교 대상 문장이 저장되어 있다. 상기 비교 대상 문장은 가비지(garbage) 제거,

문형 및 시제 확인, 키워드 추출, 용언 기본형 변환 등의 언어처리를 통해 데이터 베이스(120)로 저장될 수 있다.

[0034] 정렬모듈(130)은 입력모듈(110)에서 입력된 분석 대상 문장과, 데이터 베이스(120)에 저장된 비교 대상 문장들의 형태소를 분석한다. 여기서, 형태소는 국어에서 뜻을 가지고 있는 가장 작은 단위의 말을 의미한다. 이때, 정렬모듈(130)은 분석 대상 문장과 비교 대상 문장들에 형태소 태거를 수행한다. 여기서, 정렬모듈(130)은 형태소 태거가 완료된 비교 대상 문장들을 데이터 베이스(120)나 별도의 저장수단에 저장할 수 있다.

[0035] 한편, 정렬모듈(130)은 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소의 유무를 분석하고, 분석 결과에 따라 상기 분석 대상 문장 및 비교 대상 문장들을 정렬한다.

[0036] 즉, 상기 정렬모듈(130)은 비교 대상 문장에 포함된 형태소들 중 분석 대상 문장의 형태소와 동일한 형태소가 존재한 경우, 상기 동일한 형태소의 갯수에 따라 가산점을 부여하고, 분석 대상 문장의 형태소와 일치하지 않은 형태소의 갯수에 따라 패널티 점수를 부여한다. 이때, 정렬모듈(130)은 상기 가산점 및 패널티 점수를 토대로 상기 비교 대상 문장들을 정렬한다. 즉, 정렬모듈(130)은 해당 비교 대상 문장에 부여된 가산점에 패널티 점수를 감산하여 총점수를 산출하고, 산출된 총점수 순으로 상기 비교 대상 문장들을 정렬한다. 여기서, 정렬모듈(130)은 상술된 가산점, 패널티 점수 산출방식을 'Needleman-Wunsch Alignment Algorithm'에 적용하여 분석 대상 문장과 비교 대상 문장들을 정렬한다. 상술된 바와 같이 정렬모듈(130)은 비교 대상 문장들을 분석 대상 문장과의 형태소 일치여부에 따라 정렬하므로 보다 유사 문장을 선정하는데 있어서 정확성 및 신속성을 향상시킬 수 있다.

[0037] 패턴 분석 모듈(140)은 상기 정렬모듈(130)에서 제공되는 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정한다.

[0038] 여기서, 패턴 분석 모듈(140)은 상기 비교 대상 문장에 포함된 어휘 또는 형태소들 중 상기 분석 대상 문장의 어휘 또는 형태소와 일치하는 어휘 및 형태소의 갯수에 따라 가산점을 부여하고, 부여된 가산점을 합산하여 해당 비교 대상 문장의 유사도를 산출한다. 일예로, 비교 대상 문장이 분석 대상 문장에 포함된 어휘 또는 형태소와 일치한 어휘 또는 형태소의 갯수가 3개일 경우, 해당 비교 대상 문장은 3점의 가산점이 부여되고, 3 점의 유사도를 갖는다.

[0039] 한편, 패턴 분석 모듈(140)은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 어말어미 또는 선어말어미이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출한다. 여기서, 어말어미는 용언 어간의 가장 뒤에 오는 형태소이고, 선어말어미는 용언의 어간과 어말어미(語末語尾) 사이에 나타나는 문법형태소이다. 일예로, 비교 대상 문장이 분석 대상 문장과 일치한 형태소의 갯수가 3이고, 일치한 형태소 중 어말어미가 2개 일 경우, 부여된 3점에 추가로 2점을 부여하여 해당 비교 대상 문장은 총 5점의 유사도를 갖는다.

[0040] 또한, 패턴 분석 모듈(140)은 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 접미사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출한다. 여기서, 접미사는 접사(接辭)의 하나로 낱말의 끝에 붙어 의미를 첨가하여 다른 낱말을 이루는 말이다. 일예로, 비교 대상 문장이 분석 대상 문장과 일치한 형태소의 갯수가 3이고, 일치한 형태소 중 접미사가 1개 일 경우, 부여된 3점에 추가로 1점을 부여하여 해당 비교 대상 문장은 총 4점의 유사도를 갖는다.

[0041] 그리고, 패턴 분석 모듈(140)은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 해당 형태소가 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사 또는 인용격 조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출한다. 일예로, 비교 대상 문장이 분석 대상 문장과 일치한 형태소의 갯수가 3이고, 일치한 형태소 중 호격 조사가 1개, 인용격 조사가 1개일 경우, 부여된 3점에 추가로 2점을 부여하여 해당 비교 대상 문장은 총 5점의 유사도를 갖는다.

[0042] 또한, 패턴 분석 모듈(140)은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치하고, 상기 비교 대상 문장의 어휘가 상기 분석 대상 문장의 어휘와 일치할 경우, 해당 형태소가 보조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출한다. 여기서, 보조사는 체언이 어떤 문장성분으로 쓰이는데에 구별없이 그 체언에 어떤 뜻을 첨가하여 주는 조사의 하나이다. 일예로, 비교 대상 문장이 분석 대상 문장과 일치한 형태소의 갯수가 3이고, 일치한 형태소 중 보조사가 1개일 경우, 보조사의 어휘가 분석 대상 문장의 어휘와 일치한지 파악하고, 해당 보조사의 어휘가 상호 일치한다면, 부여된 3점에 추가로 1점을 부여하여 해당 비교 대상 문장은 총 4점의 유사도를 갖는다.

[0043] 여기서, 패턴 분석 모듈(140)은 기학습된 'Maximum Entropy Model'을 이용하여 상술된 바와 같이 데이터 베이스(120)에 저장된 비교 대상 문장들의 유사도를 계산하고, 가장 유사도가 높은 비교 대상 문장을 선정한다.

[0044] 상술된 바와 같이 구성된 본 발명에 따른 유사 문장 검색 시스템(100)은 문장의 형태소를 기준으로 문장패턴을 분석하여 분석 대상 문장과 유사한 비교 대상 문장을 선정하므로 사용자가 입력한 문장과 유사율이 비교적 높은 문장을 선정할 수 있다는 장점이 있다.

[0045] 한편, 도 2에는 본 발명에 따른 유사 문장 검색 시스템(100)을 이용한 유사 문장 검색 방법에 대한 순서도가 도시되어 있다.

[0046] 도면을 참조하면, 상기 유사 문장 검색 방법은 정렬단계(S110) 및 패턴 분석 단계(S120)를 포함한다.

[0047] 상기 정렬단계(S110)는 사용자에 의해 입력모듈(110)에 입력된 분석 대상 문장과, 데이터 베이스(120)에 저장된 비교 대상 문장의 형태소를 분석하는 단계이다. 정렬모듈(130)은 분석 대상 문장과 비교 대상 문장들에 형태소 태거를 수행한다.

[0048] 또한, 상술된 바와 같이 정렬모듈(130)은 상기 비교 대상 문장에 포함된 형태소들 중 상기 분석 대상 문장의 형태소와 동일한 형태소의 유무를 분석하고, 분석 결과에 따라 상기 분석 대상 문장 및 비교 대상 문장들을 정렬하는데, 비교 대상 문장에 포함된 형태소들 중 분석 대상 문장의 형태소와 동일한 형태소가 존재한 경우, 상기 동일한 형태소의 갯수에 따라 가산점을 부여하고, 분석 대상 문장의 형태소와 일치하지 않은 형태소의 갯수에 따라 패널티 점수를 부여하고, 상기 가산점 및 패널티 점수를 토대로 상기 비교 대상 문장들을 정렬한다.

[0049] 패턴 분석 단계(S120)는 정렬단계(S110)에서 분석정보를 토대로 형태소를 기준으로 상기 분석 대상 문장과 비교 대상 문장들의 상호 유사도를 분석하여 가장 높은 유사도를 갖는 상기 비교 대상 문장을 선정하는 단계이다.

[0050] 여기서, 패턴 분석 모듈(140)은 상술된 바와 같이 상기 비교 대상 문장에 포함된 어휘 또는 형태소들 중 상기 분석 대상 문장의 어휘 또는 형태소와 일치하는 어휘 및 형태소의 갯수에 따라 가산점을 부여하고, 부여된 가산점을 합산하여 해당 비교 대상 문장의 유사도를 산출한다.

[0051] 또한, 패턴 분석 모듈(140)은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치할 경우, 비교 대상 문장과 분석 대상 문장의 상호 일치하는 해당 형태소가 어말어미, 선어말어미, 접미사, 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사, 인용격 조사 또는 접미사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출한다. 한편, 패턴 분석 모듈(140)은 상기 비교 대상 문장의 형태소가 상기 분석 대상 문장의 형태소와 일치하고, 상기 비교 대상 문장의 어휘가 상기 분석 대상 문장의 어휘와 일치할 경우, 해당 형태소가 보조사이면, 추가로 가산점을 부여하여 해당 비교 대상 문장의 유사도를 산출한다.

[0052] 제시된 실시 예들에 대한 설명은 임의의 본 발명의 기술분야에서 통상의 지식을 가진 자가 본 발명을 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 발명의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이며, 여기에 정의된 일반적인 원리들은 본 발명의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 발명은 여기에 제시된 실시예들로 한정되는 것이 아니라, 여기에 제시된 원리들 및 신규한 특징들과 일관되는 최광의의 범위에서 해석되어야 할 것이다.

부호의 설명

[0053] 100: 유사 문장 검색 시스템

110: 입력모듈

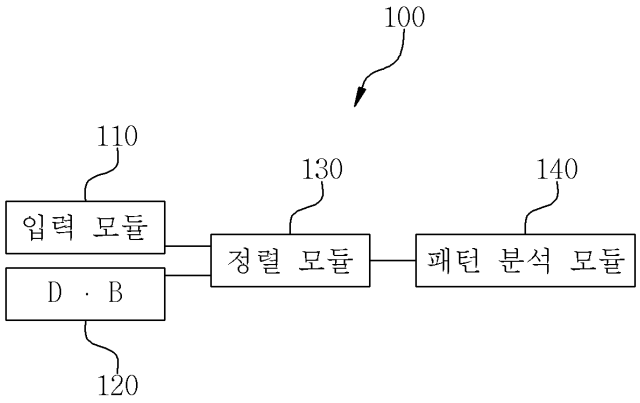
120: 데이터 베이스

130: 정렬모듈

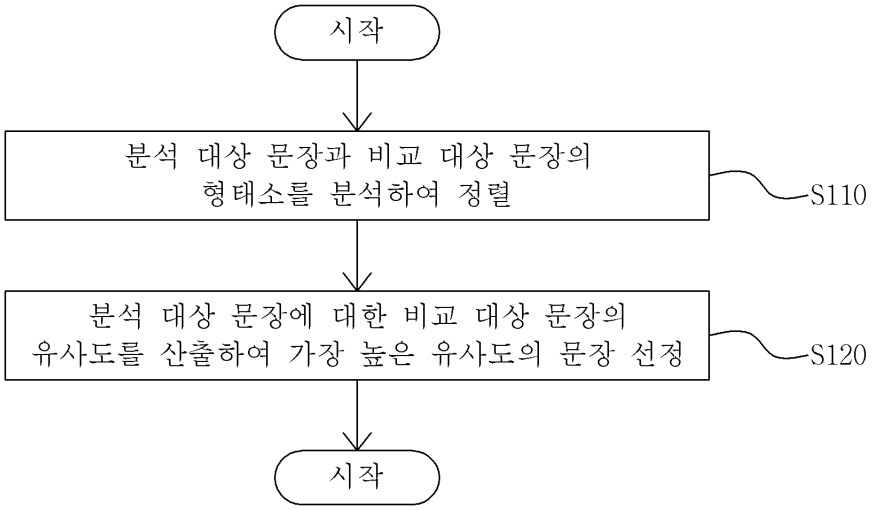
140: 패턴 분석 모듈

도면

도면1


100
110
130
140
입력 모듈
정렬 모듈
패턴 분석 모듈
D · B
120

도면2


시작
분석 대상 문장과 비교 대상 문장의
형태소를 분석하여 정렬
S110
분석 대상 문장에 대한 비교 대상 문장의
유사도를 산출하여 가장 높은 유사도의 문장 선정
S120
시작