



- (51) 국제특허분류:
G06Q 50/20 (2012.01) G06Q 50/30 (2012.01)
- (21) 국제출원번호: PCT/KR2014/012741
- (22) 국제출원일: 2014년 12월 23일 (23.12.2014)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
10-2013-0162902 2013년 12월 24일 (24.12.2013) KR
- (71) 출원인: 서강대학교 산학협력단 (SOGANG UNIVERSITY RESEARCH FOUNDATION) [KR/KR]; 121-742 서울시 마포구 백범로 35 서강대학교 데이아르관 306호, Seoul (KR). 강원대학교 산학협력단 (KNU-INDUSTRY COOPERATION FOUNDATION) [KR/KR]; 200-701 경기도 춘천시 강원대학길 1, Gyeonggi-do (KR).
- (72) 발명자: 서정연 (SEO, Jung Yun); 137-933 서울시 서초구 서초중앙로 206, 삼호가든맨션 B동 203호, Seoul (KR). 구명완 (KOO, Myoung-Wan); 158-773 서울시 양천구 목동동로 100, 목동신시가지아파트 13단지 1312동 1304호, Seoul (KR). 강상우 (KANG, Sang Woo); 136-110 서울시 성북구 길음로 20, 삼성래미안 3차 607

동 302호, Seoul (KR). 김학수 (KIM, Hark Soo); 137-765 서울시 서초구 신반포로 137, 경남아파트 5동 909호, Seoul (KR). 최맹식 (CHOI, Maeng Sik); 245-938 강원도 삼척시 남양길 14-4, Gangwon-do (KR). 송영길 (SONG, Yeong Kil); 200-729 강원도 춘천시 모수물길 60, 현대아파트 105동 804호, Gangwon-do (KR). 전원표 (JEON, Won Pyo); 200-948 강원도 춘천시 미려골길 3, 109호, Gangwon-do (KR).

(74) 대리인: 이지연 (LEE, Ji-Yeon); 151-832 서울시 관악구 남부순환로 1922, 청동빌딩 301호, 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, BN, 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, IR, IS, JP, KE, KG, KN, KP, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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.

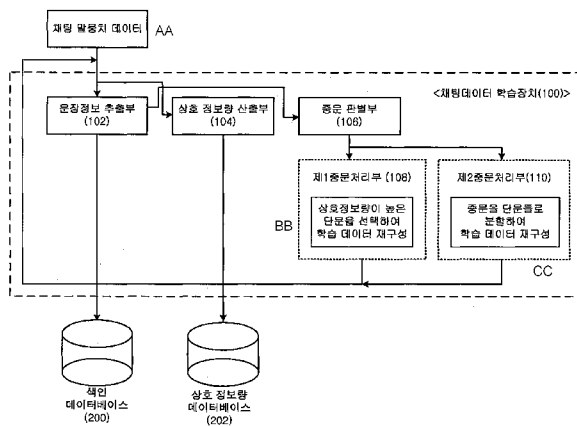
(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM,

[다음 쪽 계속]

(54) Title: CHATTING DATA LEARNING AND SERVICE METHOD AND SYSTEM THEREFOR

(54) 발명의 명칭: 채팅 데이터 학습 및 서비스 방법 및 그에 따른 시스템

도 1



- 100 ... Chatting data learning device
102 ... Sentence information extraction unit
104 ... Mutual information amount calculation unit
106 ... Compound sentence identification unit
108 ... First compound processing unit
110 ... Second compound processing unit
200 ... Index database
202 ... Mutual information amount database
AA ... Chatting corpus data
BB ... Select simple sentence with high mutual information amount and reconfigure learning data
CC ... Divide compound sentence into simple sentences and reconfigure learning data

(57) Abstract: According to the present invention, chatting corpus data, which comprises user utterance data and system response data, is received as learning data; learning is conducted to generate index information between the user utterance data and the system response data; when the user utterance data is a compound sentence, a mutual information amount in connection with the system response data is calculated with regard to each of simple sentences that constitute the compound sentence; one of the simple sentences is selected on the basis of the mutual information amount; learning data is generated using data regarding the selected simple sentence and the system response data; and re-learning is conducted, thereby providing system response data which is reliable in connection with compound sentences as well. In addition, according to the present invention, when user utterance data is input, system response data corresponding to the user utterance data is detected and output; or, when no system response data is detected and when the user utterance data is a compound sentence, system response data corresponding to each of simple sentences constituting the compound sentence is detected; a mutual information amount between the detected system response data and the simple sentences is calculated; and one of pieces of system response data is selected and output on the basis of the mutual information amount, thereby providing system response data which is reliable in connection with compound sentences as well.

(57) 요약서:

[다음 쪽 계속]



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

공개:

— 국제조사보고서와 함께 (조약 제 21 조(3))

본 발명은 사용자 발화 데이터와 시스템 응답 데이터로 이루어진 채팅 말뭉치 데이터를 학습 데이터로 입력받아 사용자 발화 데이터와 시스템 응답 데이터 사이의 색인정보를 생성하는 학습을 이행하며, 사용자 발화 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 시스템 응답 데이터와의 상호정보량을 산출하고, 그 상호정보량을 토대로 단문들 중 어느 하나를 선택하여 그 선택된 단문에 대한 데이터와 시스템 응답 데이터로 학습 데이터를 생성하여 재학습을 이행함으로써, 중문에 대해서도 신뢰할 수 있는 시스템 응답 데이터를 제공할 수 있게 한다. 또한 본 발명은 사용자 발화 데이터가 입력되면, 그 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 검출하여 출력하거나, 시스템 응답 데이터가 검출되지 않으며 사용자 발화 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 대응되는 시스템 응답 데이터들을 검출하고, 그 검출된 시스템 응답 데이터들과 단문 사이의 상호정보량을 산출하고, 그 상호정보량을 토대로 시스템 응답 데이터들 중 어느 하나를 선택하여 출력함으로써, 중문에 대해서도 신뢰할 수 있는 시스템 응답 데이터를 제공할 수 있다.

【명세서】

【발명의 명칭】

채팅 데이터 학습 및 서비스 방법 및 그에 따른 시스템

5

【기술분야】

본 발명은 채팅 시스템에 관한 것으로, 더욱 상세하게는 채팅 데이터 학습 및 서비스 방법 및 그에 따른 시스템에 관한 것이다.

10 【배경기술】

채팅 시스템에 관한 연구로는 패턴 매칭 방법, 키워드 인식, 마르코프 모델과 유전 알고리즘 등을 이용한 방법이 있다.

상기 패턴 매칭 방법은 사용자가 입력한 문장에 포함된 키워드로부터 미리 정의된 문장으로 응답하는 간단한 패턴 매칭 방법과 어휘를 치환하는 등의 방법을 이용한다. 이에 대한 대표적인 시스템으로는 최초의 채팅 시스템인 ELIZA가 있다. 상기한 패턴 매칭 방법은 구현이 쉬운 장점이 있지만 비슷하거나 동일한 대화가 계속되어 채팅의 유연성이 부족하고, 채팅의 수준을 향상시키기 위해서는 대량의 말뭉치가 필요하다는 단점이 있었다.

20 상기 마르코프 모델과 유전 알고리즘을 이용한 방법은 검색이 아닌 새로운 문장을 생성하는 방법이다. 상기 마르코프 모델을 이용한 방법은 인간의 대화 내용을 마르코프 모형으로 저장한 후에 사용자의 입력에 대해 데이터를 조회하여 가장 적절한 문장을 생성하는 문장 생성 기법 중 하나이다. 그리고 상기 유전 알고리즘은 데이터베이스에 저장되어 있는
25 데이터로부터 새로운 데이터를 생성하는 방법이다. 하지만 이 방법은 문법이 잘못된 문장을 생성할 수 있기 때문에 적절한 필터링이 요구되는 단점이 있었다.

【발명의 상세한 설명】

30 【기술적 과제】

본 발명은 사용자 발화 데이터와 시스템 응답 데이터로 이루어진 채팅

말뭉치 데이터를 학습 데이터로 입력받아 사용자 발화 데이터와 시스템
 응답 데이터 사이의 색인정보를 생성하는 학습을 이행하며, 사용자 발화
 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 시스템
 응답 데이터와의 상호정보량을 산출하고, 그 상호정보량을 토대로 단문들
 5 중 어느 하나를 선택하여 그 선택된 단문에 대한 데이터와 시스템 응답
 데이터로 학습 데이터를 생성하여 재학습을 이행하는 채팅 데이터 학습
 방법 및 그에 따른 시스템을 제공하는 것을 그 목적으로 한다.

또한 본 발명의 다른 목적은 사용자 발화 데이터가 입력되면, 그
 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 검출하여 출력하거나,
 10 시스템 응답 데이터가 검출되지 않으며 사용자 발화 데이터가 중문인
 경우에는 그 중문을 구성하는 단문들 각각에 대해 대응되는 시스템 응답
 데이터들을 검출하고, 그 검출된 시스템 응답 데이터들과 단문 사이의
 상호정보량을 산출하고, 그 상호정보량을 토대로 시스템 응답 데이터들 중
 어느 하나를 선택하여 출력하는 채팅 데이터 서비스 방법 및 그에 따른
 15 시스템을 제공하는 것이다.

【기술적 해결방법】

상기한 목적을 달성하기 위한 본 발명에 따르는 색인 데이터베이스와
 상호정보량 데이터베이스와 채팅 데이터 학습장치로 구성되는 채팅 데이터
 20 학습 시스템에서의 채팅 데이터 학습 장치에 의해 구현되는 채팅 데이터
 학습방법은, (a) 사용자 발화 데이터와 시스템 응답 데이터로 구성되는
 학습 데이터를 입력받는 단계; (b) 상기 학습 데이터를 이용하여 학습하여,
 사용자 발화 데이터의 문장 정보들과 이에 대응되는 시스템 응답 데이터의
 색인 정보를 생성하여 색인 데이터베이스에 저장하고, 사용자 발화
 25 데이터의 형태소들과 상기 시스템 응답 데이터의 형태소들 사이의
 상호정보량을 산출하여 상호정보량 데이터베이스에 저장하는 단계; (c)
 상기 학습 데이터의 사용자 발화 데이터가 중문인지 여부를 판단하는 단계;
 (d) 상기 사용자 발화 데이터가 중문이면, 사용자 발화 데이터를 단문으로
 분할하고, 분할된 단문들을 이용하여 학습 데이터를 재구성하고, 재구성된
 30 학습 데이터를 이용하여 재학습하여 색인 데이터베이스 및 상호정보량
 데이터베이스를 갱신하는 단계;를 구비하는 것을 특징으로 한다.

또한 본 발명에 따르는 색인 데이터베이스와 상호 정보량 데이터베이스와 채팅 데이터 서비스 장치를 포함하는 채팅 데이터 서비스 시스템에서의 채팅 데이터 서비스 장치에 의해 구현되는 채팅 데이터 서비스 방법은, (a) 외부로부터 사용자 발화 데이터를 입력받는 단계; (b) 5 상기 입력된 사용자 발화 데이터의 문장정보를 추출하고, 그 문장정보에 대응되는 시스템 응답 데이터를 색인 데이터베이스에 저장된 색인정보에서 검출하여 출력하는 단계; (c) 상기 문장정보에 대응되는 시스템 응답 데이터가 검출되지 않으면, 그 사용자 발화 데이터가 중문인지 여부를 판별하는 단계; (d) 상기 사용자 발화 데이터가 중문이면, 사용자 발화 10 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그 문장정보에 대응되는 시스템 응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 그 시스템 응답 데이터들과 단문들 사이의 상호정보량을 검출하고, 상기 상호정보량을 이용하여 시스템 응답 데이터를 선택하여 출력하는 단계;를 구비하는 것을 특징으로 한다.

15

【유리한 효과】

본 발명은 사용자 발화 데이터와 시스템 응답 데이터로 이루어진 채팅 말뭉치 데이터를 학습 데이터로 입력받아 사용자 발화 데이터와 시스템 응답 데이터 사이의 색인정보를 생성하는 학습을 이행하며, 사용자 발화 20 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 시스템 응답 데이터와의 상호정보량을 산출하고, 그 상호정보량을 토대로 단문들 중 어느 하나를 선택하여 그 선택된 단문에 대한 데이터와 시스템 응답 데이터로 학습 데이터를 생성하여 재학습을 이행함으로써, 중문에 대해서도 신뢰할 수 있는 시스템 응답 데이터를 제공할 수 있게 학습할 수 있는 25 효과가 있다.

또한 본 발명은 사용자 발화 데이터가 입력되면, 그 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 검출하여 출력하거나, 시스템 응답 데이터가 검출되지 않으며 사용자 발화 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 대응되는 시스템 응답 데이터들을 30 검출하고, 그 검출된 시스템 응답 데이터들과 단문 사이의 상호정보량을 산출하고, 그 상호정보량을 토대로 시스템 응답 데이터들 중 어느 하나를

선택하여 출력함으로써, 중문에 대해서도 신뢰할 수 있는 시스템 응답 데이터를 제공할 수 있는 효과가 있다.

【도면의 간단한 설명】

5 도 1은 본 발명의 바람직한 실시예에 따른 채팅 데이터 학습 시스템의 구성도.

 도 2는 본 발명의 바람직한 실시예에 따른 채팅 데이터 학습 방법의 절차도.

10 도 3은 본 발명의 바람직한 실시예에 따른 채팅 데이터 서비스 시스템의 구성도.

 도 4는 본 발명의 바람직한 실시예에 따른 채팅 데이터 서비스 방법의 절차도.

【발명의 실시를 위한 최선의 형태】

15 본 발명은 사용자 발화 데이터와 시스템 응답 데이터로 이루어진 채팅 말뭉치 데이터를 학습 데이터로 입력받아 사용자 발화 데이터와 시스템 응답 데이터 사이의 색인정보를 생성하는 학습을 이행하며, 사용자 발화 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 시스템 응답 데이터와의 상호정보량을 산출하고, 그 상호정보량을 토대로 단문들
20 중 어느 하나를 선택하여 그 선택된 단문에 대한 데이터와 시스템 응답 데이터로 학습 데이터를 생성하여 재학습을 이행함으로써, 중문에 대해서도 신뢰할 수 있는 시스템 응답 데이터를 제공할 수 있게 한다.

 또한 본 발명은 사용자 발화 데이터가 입력되면, 그 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 검출하여 출력하거나, 시스템
25 응답 데이터가 검출되지 않으며 사용자 발화 데이터가 중문인 경우에는 그 중문을 구성하는 단문들 각각에 대해 대응되는 시스템 응답 데이터들을 검출하고, 그 검출된 시스템 응답 데이터들과 단문 사이의 상호정보량을 산출하고, 그 상호정보량을 토대로 시스템 응답 데이터들 중 어느 하나를
30 선택하여 출력함으로써, 중문에 대해서도 신뢰할 수 있는 시스템 응답 데이터를 제공할 수 있다.

<채팅 데이터 학습 시스템>

상기한 본 발명의 바람직한 실시예에 따른 채팅 데이터 학습 시스템의 구성을 도 1을 참조하여 상세히 설명한다.

상기 채팅 데이터 학습 시스템은 채팅 데이터 학습장치(100)와 색인
5 데이터베이스(200)와 상호 정보량 데이터베이스(202)로 구성된다.

상기 채팅 데이터 학습장치(100)는 문장정보 추출부(102)와 상호
정보량 산출부(104)와 중문 판별부(106)와 제1중문 처리부(108)와 제2중문
처리부(110)로 구성된다.

10 상기 문장정보 추출부(102)는 학습 데이터를 입력받아 학습 데이터 중
사용자 발화 데이터에서 문장정보를 추출하여 색인정보를 생성하고 그
색인정보를 색인 데이터베이스(200)에 저장한다.

상기 사용자 발화 데이터에서 추출하는 문장정보는 [모든 품사에 대한
형태소들], [체언 및 용언, 부사에 대한 형태소들], [명사류에 대한
15 형태소들], [문장 특성정보], [연결어미정보] 등으로 구성된다.

상기 문장정보는 축약수준에 따라 제1 내지 제3색인정보로 구성되며,
상기 제1색인정보는 모든 품사에 대한 형태소들로 구성되고, 제2색인정보는
체언 및 용언, 부사에 대한 형태소들과 문장특성정보로 구성되고, 상기
제3색인정보는 명사류에 대한 형태소들과 문장특성정보로 구성된다. 상기한
20 제1 내지 제3색인정보는 축약수준에 따라 구별된 것으로, 축약 수준이
높아질수록 색인되는 정보가 적게되어 검색속도를 향상시킬 수 있다.

상기 문장 특성정보는 문장 유형정보와 양상정보와 긍정 또는
부정정보 등으로 구성된다.

상기 문장 특성정보의 문장유형정보는 12개로 구성되며, 이는 표 1에
25 나타낸 바와 같다.

【표 1】

	태그	분류	설명
1	Declarative_sentence	평서문	...
2	Imperative_sentence	명령문	어떻게 되었어?/얼마나 좋아해?/...
3	Exclamatory_sentence	감탄문	정말 좋겠다./와 신나는구나./...
4	Hortatory_sentence	청유문	~해줘./예약해줘./...
5	How_question	WH 의문문	어떻게 되었어?/얼마나 좋아해?/...
6	YN_question	YN 의문문	어떻게 되었어?/얼마나 좋아해?/...
7	What_question	WH 의문문	점심때 뭐 먹었나?/뭐 입었는데?/...
8	Who_question	WH 의문문	누가 그렇게 말했지?/누가 주연인데?/...
9	Where_question	WH 의문문	어디서 하는데?/어디가 제일 좋아?/...
10	When_question	WH 의문문	언제 방송했어?/언제 그렇게 됐지?/...
11	Which_question	WH 의문문	여름하고 겨울중에 어떤게 좋아?/...
12	Why_question	WH 의문문	왜 그러는데?/왜 좋아하지?/...

상기 표 1을 참조하면, 상기 문장유형은 평서문, 명령문, 감탄문, 청유문, WH 의문문, YN 의문문, WH 의문문, WH 의문문, WH 의문문, WH 의문문, WH 의문문, WH 의문문으로 구성된다.

그리고 상기 문장 특성정보의 양상정보는 35개로 구성되며, 이는 표 2에 나타낸 바와 같다. 상기 양상정보는 의도, 확신, 시도, 가능, 사동, 원인, 종결, 완료, 희망, 불허, 요행, 가합, 추측, 습관, 일념, 불가능, 불가피, 극한, 부정, 당위, 피동, 시인, 가식, 진행, 규정, 재귀, 보유, 예정, 봉사, 시발, 처지, 강조, 경향, 가치, 기타로 구성된다.

【표 2】

	태그	분류	설명
1	Affair	의도	~려고 하~
2	Assurance	확신	~게 마련이~
3	Attempt	시도	~어 보~
4	Available	가능	~르 수 있~
5	Causative	사동	~게 하~
6	Cause	원인	~기 때문이~
7	Closing	종결	~고 말~
8	Complete	완료	~어 있~
9	Desire	희망	~기 바라~
10	Disallowance	불허	~서는 안 되~
11	Fluke	요행	~르 뻔하~
12	Appropriation	가합	~어도 좋~
13	Guess	추측	~ㄴ 것 같~
14	Habit	습관	~곤 하~
15	Hopes	일념	~르 따름이~
16	Impossibility	불가능	~르 수 없~
17	Inevitable	불가피	~르 수밖에 없~
18	Limit	극한	~르 지경이~
19	Negative	부정	~지 않~
20	Appropriateness	당위	~어야 하~
21	Passive	피동	~당하~
22	Admit	시인	~기도 하~
23	Pretense	가식	~ㄴ 척하~
24	Progress	진행	~고 있~
25	Provision	규정	~ㄴ 바 있~
26	Recursive	재귀	~게 되~
27	Retention	보유	~어 놓~
28	Schedule	예정	~르 것이~
29	Service	봉사	~어 주~
30	Start	시발	~어 들~
31	Situation	처지	~ㄴ 터이~
32	Stress	강조	~어 대~
33	Tendency	경향	~ㄴ 편이~
34	Value	가치	~르 만하~
35	ETC	기타	

그리고 상기 연결어미정보는 사용자 발화 데이터가 중문인지 아닌지를 판별하기 위한 연결어미가 존재하는지, 그 연결어미가 중문을 구성하는 단문을 대등적으로 연결하는지, 종속적으로 연결하는지에 대한 식별정보로 구성된다. 즉, 연결어미가 존재하면 중문으로 판단하고, 그 연결어미가 -고, 5 -며, -자, -면서, -지만, -락, -거나, 든지 등이면 대등적 연결 어미로 연결된 중문으로 판단하고, 그 연결어미가 -면, -니, -나, --려고, -러, -자, -뿐더러, -니까 등이면 종속적 연결어미로 연결된 중문으로 판단한다.

즉, 상기 문장정보 추출부(102)는 학습 데이터를 입력받아 학습 10 데이터 중 사용자 발화 데이터에 대한 제1 내지 제3색인정보와 연결어미정보와 그 사용자 발화 데이터에 대응되는 시스템 응답 데이터로 구성되는 색인정보를 생성하고 그 색인정보를 색인 데이터베이스(200)에 저장한다.

상기 상호정보량 산출부(104)는 상기 학습 데이터를 입력받아 사용자 발화 데이터와 시스템 응답 데이터를 구성하는 형태소들 사이의 상호정보량을 산출하고, 상기 학습 데이터에 속한 사용자 발화 데이터와 시스템 응답 데이터 사이의 상호 정보량 산출값에 대한 상호 정보량 정보를 상호 정보량 데이터베이스(202)에 저장한다. 여기서, 상기 상호 정보량 20 정보에는 해당 학습 데이터의 색인정보를 지시하는 어드레스정보가 포함되며, 이는 채팅 데이터 서비스시에 이용될 수 있다.

상기 상호 정보량 산출식은 수학식 1과 같다.

【수학식 1】

$$MI(u, f) = \frac{p(u \cap f)}{p(u) \times p(f)}$$

상기 상호 정보량은 두 사건 사이에 얼마나 밀접한 관계가 있는지를 수치화한 값으로, 본 발명에서는 사용자 발화 데이터와 시스템 응답 데이터를 구성하는 형태소가 얼마나 밀접한 관계를 가지는지를 수치화하는 데에 사용된다.

상기 수학적 식 1에서 $p(u)$ 는 학습 데이터의 사용자 발화 데이터에서 u_i (형태소)가 등장하는 확률로써, 학습 데이터의 사용자 발화 데이터 전체에 대해 형태소 분석하여, 현재 사용자 발화 데이터에서 해당 형태소가 나타난 수를 학습 데이터의 사용자 발화 데이터 전체의 형태소 수로 나눈 것이다.

5 그리고 $p(f)$ 는 학습 데이터의 시스템 응답 데이터에서 f_i (형태소)가 등장하는 확률로써, 학습 데이터의 시스템 응답 데이터 전체에 대해 형태소 분석하여, 현재 시스템 응답 데이터에서 해당 형태소가 나타난 수를 학습 데이터의 시스템 응답 데이터 전체의 형태소 수로 나눈 것이다.

10 그리고 $p(u \cap f)$ 는 학습 데이터의 사용자 발화 데이터와 시스템 응답 데이터 전체에서 구한 것으로, 전체 학습 데이터에서 사용자 발화 데이터에서 u_i 가 시스템 응답 데이터에서 f_i 가 동시에 출현할 확률을 의미한다. 따라서 사용자 발화 데이터에서 u_i 가 시스템 응답 데이터에서 f_i 가 동시에 나타난 수를 사용자 발화 데이터와 시스템 응답 데이터의 형태소 조합의 수로 나눈 것이다.

15 즉, 상기 수학적 식 1은 학습 데이터 전체에서 사용자 발화 데이터에 속한 제1형태소들과 시스템 응답 데이터에 속한 제2형태소들이 동시에 출현할 확률 $p(u \cap f)$ 을, 현재 입력된 학습 데이터의 사용자 발화 데이터에서 제1형태소들이 등장할 확률 $p(u)$ 과 현재 입력된 학습 데이터의 시스템 응답 데이터에서 제2형태소들이 등장할 확률 $p(f)$ 을 곱한 값으로
20 나누어, 상호 정보량을 산출하는 것을 나타낸다.

상기 중문 판별부(106)는 문장정보 추출부(102)로부터 학습 데이터 중 사용자 발화 데이터가 중문으로 판단된 학습 데이터를 제공받아 연결어미에 따라 제1중문처리부(108)에 제공하거나 제2중문처리부(110)에 제공한다.

25 상기 제1중문처리부(108)는 종속적 연결어미로 연결된 중문으로 판단된 학습 데이터의 사용자 발화 데이터를 연결어미를 기준으로 단문들로 분할하고, 그 분할된 단문들 각각에 대해 시스템 응답 데이터와의 상호 정보량을 산출하고, 그 상호 정보량 산출값이 가장 큰 단문과 시스템 응답 데이터를 결합하여 재학습 데이터를 생성하여 상기 문장 추출부(102)에
30 제공하여 재학습 데이터로 재학습되게 한다.

이와 같이 장문의 사용자 발화 데이터에 대해 문장을 분리하고,

밀접한 연관이 있는 것으로 판단되는 단문만으로 재학습하여 색인정보를 생성하는 것은, 시스템의 재현율을 높일 수 있는 효과를 야기한다.

상기 단문들 각각에 대한 상호 정보량 산출식은 수학식 2 및 수학식 3과 같다.

5

【수학식 2】

$$MI_{u,f} = \begin{bmatrix} MI(u_1, f_1) & MI(u_1, f_m) \\ MI(u_n, f_1) & MI(u_n, f_m) \end{bmatrix}$$

상기 수학식 2는 사용자 발화 데이터에 포함된 형태소들($u_1 \sim u_n$)과
10 시스템 응답 데이터에 포함된 형태소들($f_1 \sim f_n$) 사이의 상호 정보량을 나타낸 것이다.

【수학식 3】

$$S(u, f) = \frac{\sum_{i=1}^n \max(MI(u_i, f_1), \dots, MI(u_i, f_m))}{n}$$

15

상기 수학식 3은 사용자 발화 데이터를 단문들로 분할하고, 분할된 단문들 각각에 포함된 형태소들과 시스템 응답 데이터에 포함된 형태소들 사이의 상호 정보량을 산출하여 더한 것을, 해당 단문에 포함된 형태소들의 수로 나눈 것을 의미한다. 즉 수학식 3에 따라 단문들 각각에 대해 시스템
20 응답 데이터와의 상호 정보량이 산출된다.

그리고 제2중문처리부(110)는 대등적 연결어미로 연결된 중문으로 판단된 학습 데이터를 입력받아, 그 학습 데이터의 사용자 발화 데이터를 연결어미를 기준으로 단문들로 분할하고, 그 분할된 단문들 각각과 학습
25 데이터의 시스템 응답 데이터를 결합하여 재학습 데이터들을 생성하여 상기

문장 추출부(102)에 제공하여 재학습 데이터들에 의해 재학습되게 한다. 이와 같이 대등적 연결어미로 연결된 중문으로 판단된 사용자 발화 데이터의 경우에는 주어진 학습 데이터를 최대한 활용할 수 있게 함으로써 시스템의 재현율을 높일 수 있다.

5

<채팅 데이터 학습 방법>

본 발명의 바람직한 실시예에 따른 채팅 데이터 학습 방법의 구성을 도 2를 참조하여 상세히 설명한다.

상기 채팅 데이터 학습 시스템은 채팅 데이터 학습장치(100)는
10 외부로부터 사용자 말뭉치 데이터가 학습 데이터로 입력되거나 내부에서 생성된 재학습 데이터가 학습 데이터로 재입력되는지를 체크한다(300단계).

상기 채팅 데이터 학습장치(100)는 상기 학습 데이터가 입력되면,
상기 학습 데이터에 포함된 사용자 발화 데이터의 문장정보를 추출하고, 그
문장정보와 시스템 응답 데이터를 매칭하는 색인정보를 구성하여 색인
15 데이터베이스(200)에 저장한다(302단계).

또한 상기 채팅 데이터 학습장치(100)는 상기 학습 데이터 또는
재학습 데이터에 포함된 사용자 발화 데이터와 시스템 응답 데이터를
이루는 형태소들 사이의 관계를 지시하는 상호 정보량을 산출하고, 그 상호
정보량 산출값과 상기 학습 데이터의 색인정보에 대한 식별정보로 상호
20 정보량 정보를 생성하고, 그 상호 정보량 정보를 상호 정보량
데이터베이스(202)에 저장한다(304단계).

또한 상기 채팅 데이터 학습장치(100)는 상기 학습 데이터의 사용자
발화 데이터가 중문인지를 체크하고, 상기 사용자 발화 데이터가 중문이면
종속적 연결어미로 연결된 중문인지 대등적 연결어미로 연결된 중문인지를
25 체크한다(306, 308, 312단계).

상기 사용자 발화 데이터가 종속적 연결어미로 연결된 중문이면, 상기
채팅 데이터 학습장치(100)는 사용자 발화 데이터를 단문들로 분할하고, 그
분할된 단문들 중 시스템 응답 데이터와 상호 정보량이 높은 단문을
검출하여 선택하고, 그 선택된 단문과 시스템 응답 데이터를 결합하여
30 재학습 데이터를 생성하고, 그 재학습 데이터를 다시 학습 데이터로
재입력하여 학습을 이행한다(310단계).

만일 상기 사용자 발화 데이터가 대등적 연결어미로 연결된 중문이면, 상기 채팅 데이터 학습장치(100)는 사용자 발화 데이터를 단문들로 분할하고, 그 분할된 단문들 각각과 시스템 응답 데이터를 결합하여 재학습 데이터를 생성하고, 그 재학습 데이터를 다시 학습 데이터로 재입력하여
5 학습을 이행한다(314단계). 이로서 색인 데이터베이스의 색인정보 및 상호정보량 데이터베이스의 상호정보량을 갱신한다.

<채팅 데이터 서비스 시스템>

이제 상기한 색인정보 및 상호정보량 정보를 이용하여 채팅 데이터
10 서비스를 제공하는 본 발명의 바람직한 실시예에 따른 채팅 데이터 서비스 시스템의 구성을 도 3을 참조하여 상세하게 설명한다.

상기 채팅 데이터 서비스 시스템은 채팅 데이터 서비스 장치(400)와 색인 데이터베이스(200)와 상호 정보량 데이터베이스(202)로 구성된다.

상기 채팅 데이터 서비스 장치(400)는 사용자 발화 데이터를 입력받아
15 색인 데이터베이스(200)와 상호 정보량 데이터베이스(202)에 저장된 색인정보와 상호 정보량 정보를 이용하여 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 검출하고, 그 시스템 응답 데이터를 출력한다.

상기 색인 데이터베이스(202)에는 색인정보를 저장하며, 상기 색인정보에는 시스템 응답 데이터와 그에 대응되는 사용자 발화 데이터에
20 대한 제1 내지 제3색인정보가 포함된다. 상기 제1 내지 제3색인정보에 대해서는 상술되었으므로 상세한 설명을 생략한다.

상기 상호 정보량 데이터베이스(202)는 학습 데이터의 학습과정에서 산출된 시스템 응답 데이터의 형태소들과 그에 대응되는 사용자 발화 데이터의 형태소들 사이의 상호 정보량 산출값과 해당 학습 데이터에 대한
25 상호 정보량 정보가 저장된다.

상기 채팅 데이터 서비스 장치(400)는 문장정보 추출부(402)와 시스템 응답 데이터 검출부(404)와 중문 판별부(406)와 제1중문 처리부(408)와 제2중문 처리부(410)로 구성된다.

상기 문장정보 추출부(402)는 사용자 발화 데이터가 입력되면, 상기
30 사용자 발화 데이터로부터 문장정보를 추출하여, 그 문장정보를 시스템 응답 데이터 검출부(404)로 전송한다.

상기 시스템 응답 데이터 검출부(404)는 사용자 발화 데이터의 문장정보에 대응되는 시스템 응답 데이터를 색인 데이터베이스(200)에서 검출한다.

특히 상기 시스템 응답 데이터 검출부(404)는 축약수준에 따라 제1 내지 제3색인정보 중 어느 한 색인정보를 이용하여 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 검출한다. 상기 제1 내지 제3색인정보의 선택은 사용자에게 의한 정확도 설정에 대응되게 이루어질 수 있다. 즉 사용자가 엄밀한 응답을 원하는 경우에는 제1색인정보를 선택할 수 있고, 그렇지 않을 경우에는 제2색인정보 또는 제3색인정보를 선택할 수 있다.

상기한 바와 같이 시스템 응답 데이터 검출부(404)는 사용자 발화 데이터의 문장정보에 대응되는 시스템 응답 데이터가 색인 데이터베이스(200)에서 검출되면, 이를 출력한다.

그러나 상기 사용자 발화 데이터의 문장정보에 대응되는 시스템 응답 데이터가 존재하지 않으면, 상기 사용자 발화 데이터를 중문 판별부(406)에 제공한다.

상기 중문 판별부(406)는 상기 사용자 발화 데이터가 종속적 연결어미로 연결된 중문인지, 대등적 연결어미로 연결된 중문인지를 판별하고, 중문이 아닌 경우에는 시스템 응답 데이터 없음을 출력한다.

상기 중문 판별부(406)는 상기 사용자 발화 데이터가 종속적 연결어미로 연결된 중문이면 이를 제1중문 처리부(408)로 제공하고, 상기 사용자 발화 데이터가 대등적 연결어미로 연결된 중문이면 이를 제2중문처리부(410)로 제공한다.

상기 제1중문 처리부(408)는 상기 사용자 발화 데이터를 연결어미를 기준으로 단문들로 분할하고, 그 분할된 단문들에 대해 문장정보를 추출한다. 이후 상기 제1중문 처리부(408)는 단문들 각각에 대한 문장정보에 대응되는 시스템 응답 데이터들을 검출한다. 이후 상기 제1중문 처리부(408)는 상기 단문들 각각에 대한 시스템 응답 데이터들 사이의 상호정보량 산출값을 상기 상호정보량 데이터베이스(202)에서 독출하고, 그 상호정보량 산출값이 가장 높은 시스템 응답 데이터를 선택하여 출력한다.

상기 제2중문처리부(410)는 상기 사용자 발화 데이터를 연결어미를 기준으로 단문들로 분할하고, 그 분할된 단문들 각각에 대해 문장정보를

추출하고, 단문들 각각에 대해 단문들 각각의 문장정보에 대응되는 시스템 응답 데이터들을 검출하고, 이 시스템 응답 데이터들을 출력한다. 이와 같이 제2중문처리부(410)는 단문들 각각에 대한 시스템 응답 데이터를 모두 출력하는 OR 검색결과를 출력한다.

5

<채팅 데이터 서비스 방법>

상기한 채팅 데이터 서비스 장치의 동작을 도 4를 참조하여 설명한다.

상기 채팅 데이터 서비스 장치(400)는 사용자 발화 데이터가 입력되면(500단계), 사용자 발화 데이터의 문장정보를 추출함과 아울러
10 사용자 발화 데이터의 문장정보에 대응되는 시스템 응답 데이터를 색인 데이터베이스(200)에서 검출한다(502단계).

상기 채팅 데이터 서비스 장치(400)는 상기 시스템 응답 데이터가 검출되면 이를 출력한다(516단계).

이와 달리 상기 시스템 응답 데이터가 검출되지 않으면, 상기 채팅
15 데이터 서비스 장치(400)는 입력된 사용자 발화 데이터가 중문인지를 체크한다(506단계). 상기 사용자 발화 데이터가 중문이면, 상기 채팅 데이터 서비스 장치(400)는 종속적 연결어미로 연결된 중문인지를 체크한다(508단계). 상기 사용자 발화 데이터가 종속적 연결어미로 연결된 중문이면, 상기 채팅 데이터 서비스 장치(400)는 사용자 발화 데이터를
20 단문들로 분할하고, 분할된 단문들에 대응되는 시스템 응답 데이터들을 검출하고, 그 단문들과 시스템 응답 데이터들 사이의 상호 정보량을 상호 정보량 데이터베이스(202)에서 독출하고, 그 독출한 상호 정보량 산출값이 가장 높은 시스템 응답 데이터를 선택하여 출력한다(510단계).

이와 달리 상기 사용자 발화 데이터가 대등적 연결어미로 연결된
25 중문이면, 상기 채팅 데이터 서비스 장치(400)는 사용자 발화 데이터를 단문들로 분할하고, 분할된 단문들 각각에 대응되는 시스템 응답 데이터들을 검출하고, 그 시스템 응답 데이터들을 출력한다(512, 514단계).

그러나 상기 사용자 발화 데이터에 대응되는 시스템 응답 데이터도
30 검출되지 않고, 그 사용자 발화 데이터가 중문도 아닌 경우에는 시스템 응답 없음을 출력한다(518단계).

상기한 본 발명의 바람직한 실시예에서는 학습 데이터에 포함된 사용자 발화 데이터가 중문인 경우에는 1차적으로 학습을 이행한 후에, 중문을 단문으로 분할하여 2차적으로 재학습하는 것만을 예시하였으나, 학습 데이터에 입력되었을 때에 중문을 가려내어 학습할 수도 있다.

- 5 이를 좀 더 설명하면, 채팅 데이터 학습장치는, 사용자 발화 데이터와 시스템 응답 데이터로 구성되는 학습 데이터를 입력받아, 상기 사용자 발화 데이터에 연결어미가 포함되었는지를 판별하고, 상기 사용자 발화 데이터에 연결어미가 포함되어 있지 않으면, 사용자 발화 데이터에서 형태소를 포함하는 문장정보를 추출하고, 그 문장정보와 학습 데이터에 포함된
- 10 시스템 응답 데이터를 매칭시킨 색인정보를 생성하여 색인 데이터베이스에 저장함과 아울러 사용자 발화 데이터의 형태소들과 시스템 응답 데이터의 형태소들 사이의 상호정보량을 산출하고, 그 상호정보량을 포함하는 상호정보량 정보를 상호정보량 데이터베이스에 저장하는 학습을 이행하고, 상기 사용자 발화 데이터에 연결어미가 포함되어 있고, 그 연결어미가
- 15 종속적 연결어미이면, 상기 종속적 연결어미를 기준으로 중문인 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 형태소들과 시스템 응답 데이터의 형태소들 사이의 상호정보량을 산출하고, 그 상호정보량 산출값이 큰 단문과 시스템 응답 데이터를 결합하여 학습 데이터를 재구성하고, 그 학습 데이터에 대한 색인정보를 생성하여 색인
- 20 데이터베이스에 저장함과 아울러 상호정보량 정보를 생성하여 상호정보량 데이터베이스에 저장하는 학습을 이행하고, 상기 사용자 발화 데이터에 연결어미가 포함되어 있고, 그 연결어미가 대등적 연결어미이면, 상기 대등적 연결어미를 기준으로 중문인 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들과 상기 시스템 응답 데이터를 결합하여 학습
- 25 데이터들을 재구성하고, 그 재구성된 학습 데이터들에 대한 색인정보들을 생성하여 색인 데이터베이스에 저장함과 아울러 상호정보량 정보들을 생성하여 상호정보량 데이터베이스에 저장하는 학습을 이행할 수 있다.

【산업상 이용가능성】

- 30 본 발명은 인터넷이나 컴퓨터 등을 이용한 대화형 서비스 또는 채팅 서비스에 널리 사용될 수 있다.

【청구의 범위】

【청구항 1】

색인 데이터베이스와 상호정보량 데이터베이스와 채팅 데이터
 5 학습장치로 구성되는 채팅 데이터 학습 시스템에서의 채팅 데이터 학습
 장치에 의해 구현되는 채팅 데이터 학습방법에 있어서,

(a) 사용자 발화 데이터와 시스템 응답 데이터로 구성되는 학습
 데이터를 입력받는 단계;

(b) 상기 학습 데이터를 이용하여 학습하여, 사용자 발화 데이터의
 10 문장 정보들과 이에 대응되는 시스템 응답 데이터의 색인 정보를 생성하여
 색인 데이터베이스에 저장하고, 사용자 발화 데이터의 형태소들과 상기
 시스템 응답 데이터의 형태소들 사이의 상호정보량을 산출하여 상호정보량
 데이터베이스에 저장하는 단계;

(c) 상기 학습 데이터의 사용자 발화 데이터가 중문인지 여부를
 15 판단하는 단계;

(d) 상기 사용자 발화 데이터가 중문이면, 사용자 발화 데이터를
 단문으로 분할하고, 분할된 단문들을 이용하여 학습 데이터를 재구성하고,
 재구성된 학습 데이터를 이용하여 재학습하여 색인 데이터베이스 및
 상호정보량 데이터베이스를 갱신하는 단계;

20 를 구비하는 채팅 데이터 학습방법.

【청구항 2】

제1항에 있어서, 상기 (b) 단계는,

(b1) 학습 데이터 중 사용자 발화 데이터에서 형태소를 포함하는
 25 문장정보를 추출하고, 그 문장정보와 상기 시스템 응답 데이터를 매칭시킨
 색인정보를 생성하여 색인 데이터베이스에 저장하는 단계;

(b2) 상기 사용자 발화 데이터의 형태소들과 상기 시스템 응답
 데이터의 형태소들 사이의 상호정보량을 산출하고, 그 상호정보량을
 포함하는 상호정보량 정보를 상호정보량 데이터베이스에 저장하는 단계;

30 를 포함하는 것을 특징으로 하는 채팅 데이터 학습 방법.

【청구항 3】

제1항에 있어서, 상기 (c) 단계는 사용자 발화 데이터에 연결 어미가 있는지 여부를 판별하고, 연결 어미가 있는 경우 중문이라고 판단하는 것을 특징으로 하는 채팅 데이터 학습 방법.

5

【청구항 4】

제1항에 있어서, 상기 (d) 단계는,

(d1) 중문인 사용자 발화 데이터에 포함된 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하는 단계;

10 (d2) 상기 연결어미가 종속적 연결어미이면, 상기 분할된 단문들의 형태소들과 시스템 응답 데이터의 형태소들 사이의 상호정보량들을 산출하고, 상기 산출된 상호정보량이 가장 큰 단문과 시스템 응답 데이터를 결합하여 학습 데이터를 재구성하는 단계;

15 (d3) 상기 연결어미가 대등적 연결어미이면, 상기 분할된 각 단문들과 상기 시스템 응답 데이터를 결합하여 학습 데이터를 재구성하는 단계;

(d4) 상기 재구성된 학습 데이터를 이용하여 학습하여 색인 데이터베이스 및 상호정보량 데이터베이스를 갱신하는 단계;

를 구비함을 특징으로 하는 채팅 데이터 학습방법.

【청구항 5】

제1항에 있어서, 상기 (b) 단계에서의 상기 상호정보량은,

전체 학습 데이터에 대해 사용자 발화 데이터에 속한 제1형태소들과 시스템 응답 데이터에 속한 제2형태소들이 동시에 출현할 확률을,

25 현재 학습 데이터에 대해 사용자 발화 데이터에 속한 제1형태소들이 출현할 확률과, 현재 학습 데이터에 대해 시스템 응답 데이터에 속한 제2형태소들이 출현할 확률을 곱한 값으로 나누어 산출함을 특징으로 하는 채팅 데이터 학습방법.

【청구항 6】

30 제1항에 있어서,
상기 문장정보는,

모든 품사에 대한 형태소들, 체언 및 용언, 부사에 대한 형태소들, 명사류에 대한 형태소들, 문장특성정보, 연결어미정보가 포함되며,

상기 색인정보는 제1 내지 제3색인정보로 구성되며,

제1색인정보는 모든 품사에 대한 형태소들로 구성되고,

5 제2색인정보는 체언 및 용언, 부사에 대한 형태소들과 문장특성정보로 구성되고,

제3색인정보는 명사류에 대한 형태소들과 문장특성정보로 구성됨을 특징으로 하는 채팅 데이터 학습방법.

10 【청구항 7】

채팅 데이터 학습 시스템에 있어서,

사용자 발화 데이터의 문장정보와 그에 대응되는 시스템 응답 데이터를 매칭시킨 색인정보를 저장하는 색인 데이터베이스;

15 사용자 발화 데이터와 시스템 응답 데이터로 구성되는 학습 데이터들 전체에 대해 사용자 발화 데이터의 형태소들과 시스템 응답 데이터의 형태소들 사이의 상호정보량을 포함하는 상호정보량 정보를 저장하는 상호정보량 데이터베이스;

사용자 발화 데이터와 시스템 응답 데이터로 구성되는 학습 데이터를 이용하여 학습하는 채팅 데이터 학습장치;를 구비하고,

20 상기 채팅 데이터 학습 장치는,

사용자 발화 데이터로부터 문장 정보를 추출하고 각 문장 정보와 이에 대응되는 시스템 응답 데이터에 대한 색인 정보를 생성하여 상기 색인 데이터베이스에 저장하는 문장 정보 추출부;

25 사용자 발화 데이터 및 이에 대응되는 시스템 응답 데이터를 구성하는 형태소들 사이의 상호 정보량을 산출하여 상호정보량 데이터베이스에 저장하는 상호정보량 산출부;

사용자 발화 데이터가 중문인지 여부를 판별하는 중문 판별부;

30 상기 중문 판별부에 의해 중문으로 판별된 사용자 발화 데이터들에 대하여 단문으로 분할하고, 각 단문들을 이용하여 학습 데이터를 재구성하는 중문 처리부;를 구비하고,

상기 채팅 데이터 학습 장치는

상기 문장정보추출부 및 상기 상호정보량 산출부를 이용하여 학습 데이터에 대하여 학습하여 색인 정보 및 상호정보량을 각각 상기 색인 데이터베이스 및 상호정보량 데이터베이스에 저장하고,

상기 중문 판별부를 이용하여 중문으로 구성된 사용자 발화 데이터를 판별하고, 상기 중문 처리부를 이용하여 학습 데이터를 재구성하고, 재구성된 학습 데이터에 대하여 다시 학습하여 상기 색인 데이터베이스 및 상호정보량 데이터베이스를 갱신하는 것을 특징으로 하는 채팅 데이터 학습 시스템.

10 【청구항 8】

제7항에 있어서,

상기 중문 판별부는 사용자 발화 데이터에 연결어미가 포함되었는지 여부를 판별하고, 연결어미가 포함된 경우 중문으로 판단하는 것을 특징으로 하는 채팅 데이터 학습 시스템.

15

【청구항 9】

제7항에 있어서,

상기 중문 처리부는 제1 및 제2 중문 처리부로 구성되며,

상기 중문 판별부는 사용자 발화 데이터에 종속적 연결어미가 포함된 경우 제1 중문 처리부를 구동하고, 사용자 발화 데이터에 대등적 연결어미가 포함된 경우 제2 중문 처리부를 구동하는 것을 특징으로 하며,

상기 제1 중문 처리부는, 종속적 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하고, 각 단문들의 형태소들과 시스템 응답 데이터의 형태소들 사이의 상호정보량을 산출하고, 그 상호정보량이 가장 큰 단문과 이에 대응하는 시스템 응답 데이터를 결합하여 학습 데이터를 재구성하는 것을 특징으로 하며,

상기 제2 중문 처리부는, 대등적 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하고, 각 단문들과 상기 시스템 응답 데이터를 결합하여 학습 데이터들을 재구성하는 것을 특징으로 하는 채팅 데이터 학습 시스템.

30

【청구항 10】

제7항에 있어서, 상기 상호정보량 산출부는,

전체 학습 데이터에 대해 사용자 발화 데이터에 속한 제1형태소들과 시스템 응답 데이터에 속한 제2형태소들이 동시에 출현할 확률을,

- 5 현재 학습 데이터에 대해 사용자 발화 데이터에 속한 제1형태소들이 출현할 확률과, 현재 학습 데이터에 대해 시스템 응답 데이터에 속한 제2형태소들이 출현할 확률을 곱한 값으로 나누어 상호정보량을 산출함을 특징으로 하는 채팅 데이터 학습 시스템.

10 【청구항 11】

제7항에 있어서,

상기 문장정보는,

모든 품사에 대한 형태소들, 체언 및 용언, 부사에 대한 형태소들, 명사류에 대한 형태소들, 문장특성정보, 연결어미정보가 포함되며,

- 15 상기 색인정보는 제1 내지 제3색인정보로 구성되며,
제1색인정보는 모든 품사에 대한 형태소들로 구성되고,
제2색인정보는 체언 및 용언, 부사에 대한 형태소들과 문장특성정보로 구성됨을 특징으로 하는 채팅 데이터 학습 시스템.

20 【청구항 12】

색인 데이터베이스와 상호 정보량 데이터베이스와 채팅 데이터 서비스 장치를 포함하는 채팅 데이터 서비스 시스템에서의 채팅 데이터 서비스 장치에 의해 구현되는 채팅 데이터 서비스 방법에 있어서,

- (a) 외부로부터 사용자 발화 데이터를 입력받는 단계;
25 (b) 상기 입력된 사용자 발화 데이터의 문장정보를 추출하고, 그 문장정보에 대응되는 시스템 응답 데이터를 색인 데이터베이스에 저장된 색인정보에서 검출하여 출력하는 단계;
(c) 상기 문장정보에 대응되는 시스템 응답 데이터가 검출되지 않으면, 그 사용자 발화 데이터가 중문인지 여부를 판별하는 단계;
30 (d) 상기 사용자 발화 데이터가 중문이면, 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그

문장정보에 대응되는 시스템 응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 그 시스템 응답 데이터들과 단문들 사이의 상호정보량을 검출하고, 상기 상호정보량을 이용하여 시스템 응답 데이터를 선택하여 출력하는 단계;

5 를 구비하는 것을 특징으로 하는 채팅 데이터 서비스 방법.

【청구항 13】

제12항에 있어서, 상기 (c) 단계는

10 상기 사용자 발화 데이터에 연결어미가 있는지 여부를 판별하고, 연결어미가 있는 경우 중문으로 판단하는 것을 특징으로 하는 채팅 데이터 서비스 방법.

【청구항 14】

제12항에 있어서, 상기 (d) 단계는

15 (d1) 상기 사용자 발화 데이터가 종속적 연결 어미로 연결된 중문인지 대등적 연결어미로 연결된 중문인지 판별하는 단계;

 (d2) 상기 사용자 발화 데이터가 종속적 연결어미로 연결된 중문이면, 상기 종속적 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그 문장정보에
20 대응되는 시스템 응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 그 시스템 응답 데이터들과 단문들 사이의 상호정보량을 검출하고, 그 상호정보량이 가장 큰 시스템 응답 데이터를 선택하여 출력하는 단계;

 (d3) 상기 사용자 발화 데이터가 대등적 연결어미로 연결된 중문이면,
25 상기 대등적 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그 문장정보들에 대응되는 시스템 응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 상기 검출된 시스템 응답 데이터를 출력하는 단계;

30 를 더 구비함을 특징으로 하는 채팅 데이터 서비스 방법.

【청구항 15】

채팅 데이터 서비스 시스템에 있어서,
 사용자 발화 데이터의 문장정보와 그에 대응되는 시스템 응답 데이터를 매칭시킨 색인정보를 저장하는 색인 데이터베이스;

- 5 사용자 발화 데이터와 시스템 응답 데이터로 구성되는 학습 데이터들 전체에 대해 사용자 발화 데이터의 형태소들과 시스템 응답 데이터의 형태소들 사이의 상호정보량을 산출한 상호정보량 정보를 저장하는 상호정보량 데이터베이스;

- 10 외부로부터 사용자 발화 데이터가 입력되면, 색인 데이터베이스 및 상호정보량 데이터베이스를 이용하여, 상기 입력된 사용자 발화 데이터에 대응되는 시스템 응답 데이터를 추출하여 제공하는 채팅 데이터 서비스 장치;

를 구비하고,

상기 채팅 데이터 서비스 장치는,

- 15 외부로부터 입력된 사용자 발화 데이터로부터 문장 정보를 추출하고, 추출된 문장정보에 대응되는 시스템 응답 데이터를 색인 데이터베이스에 저장된 색인정보에서 검출하여 출력하고,

- 상기 문장정보에 대응되는 시스템 응답 데이터가 검출되지 않으면, 그 사용자 발화 데이터가 중문인지를 판별하고, 상기 사용자 발화 데이터가 중문이면, 상기 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그 문장정보에 대응되는 시스템 응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 상기 검출된 시스템 응답 데이터들 중 적어도 하나 이상을 출력하는 것을 특징으로 하는 채팅 데이터 서비스 시스템.

25

【청구항 16】

제15항에 있어서,

상기 채팅 데이터 서비스 장치가,

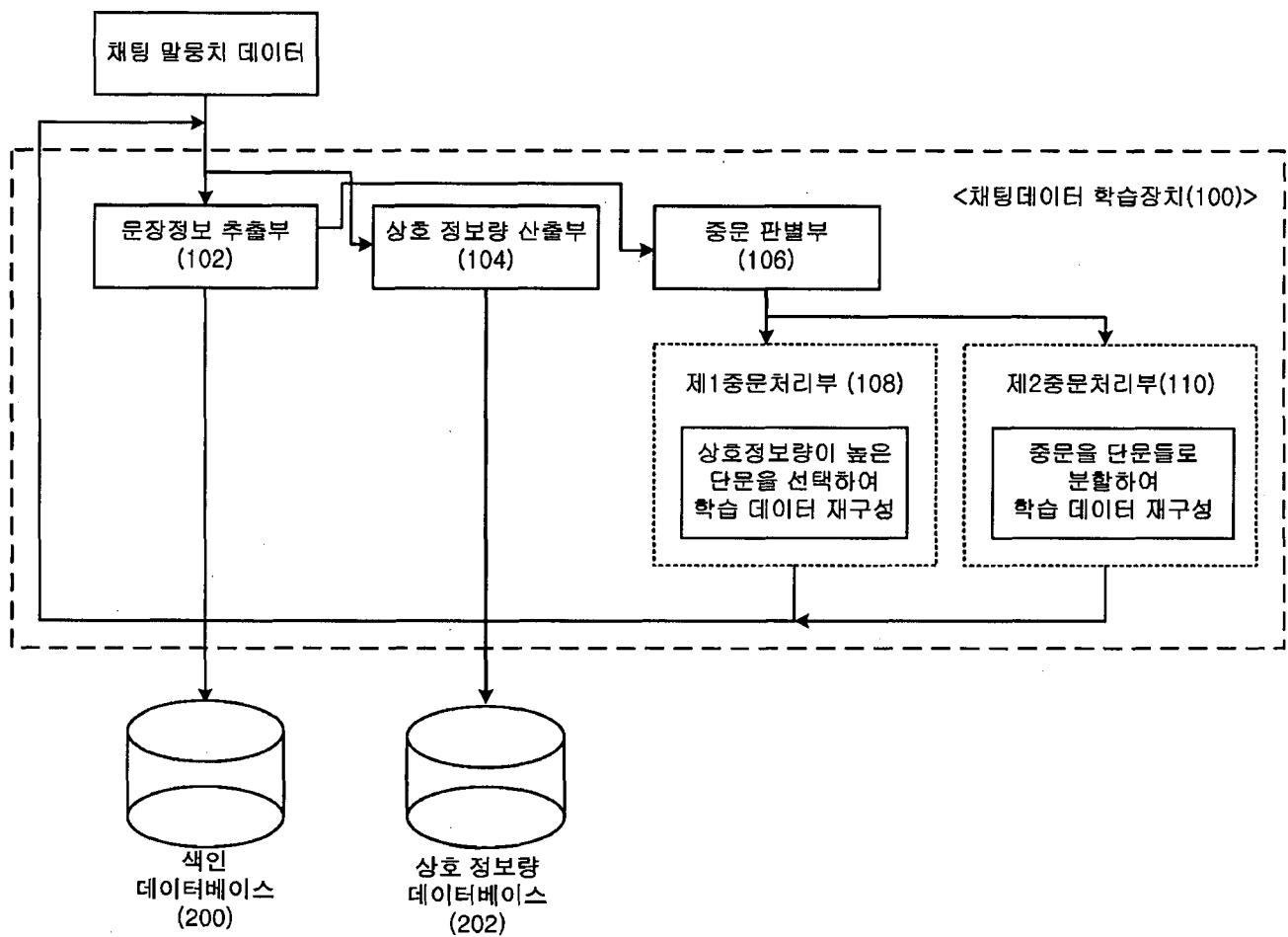
- 30 상기 사용자 발화 데이터가 종속적 연결어미로 연결된 중문이면 상기 종속적 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그 문장정보에 대응되는 시스템

응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 그 시스템 응답 데이터들과 단문들 사이의 상호정보량을 상기 상호정보량 데이터베이스에서 검출하고, 상호정보량이 가장 큰 시스템 응답 데이터를 선택하여 출력하고,

- 5 상기 사용자 발화 데이터가 대등적 연결어미로 연결된 중문이면, 상기 대등적 연결어미를 기준으로 하여 사용자 발화 데이터를 단문들로 분할하고, 각 분할된 단문들의 문장정보를 추출하고, 그 문장정보들에 대응되는 시스템 응답 데이터들을 상기 색인 데이터베이스에 저장된 색인정보에서 검출하고, 그 시스템 응답 데이터들을 출력함을 특징으로 하는 채팅 데이터
- 10 서비스 시스템.

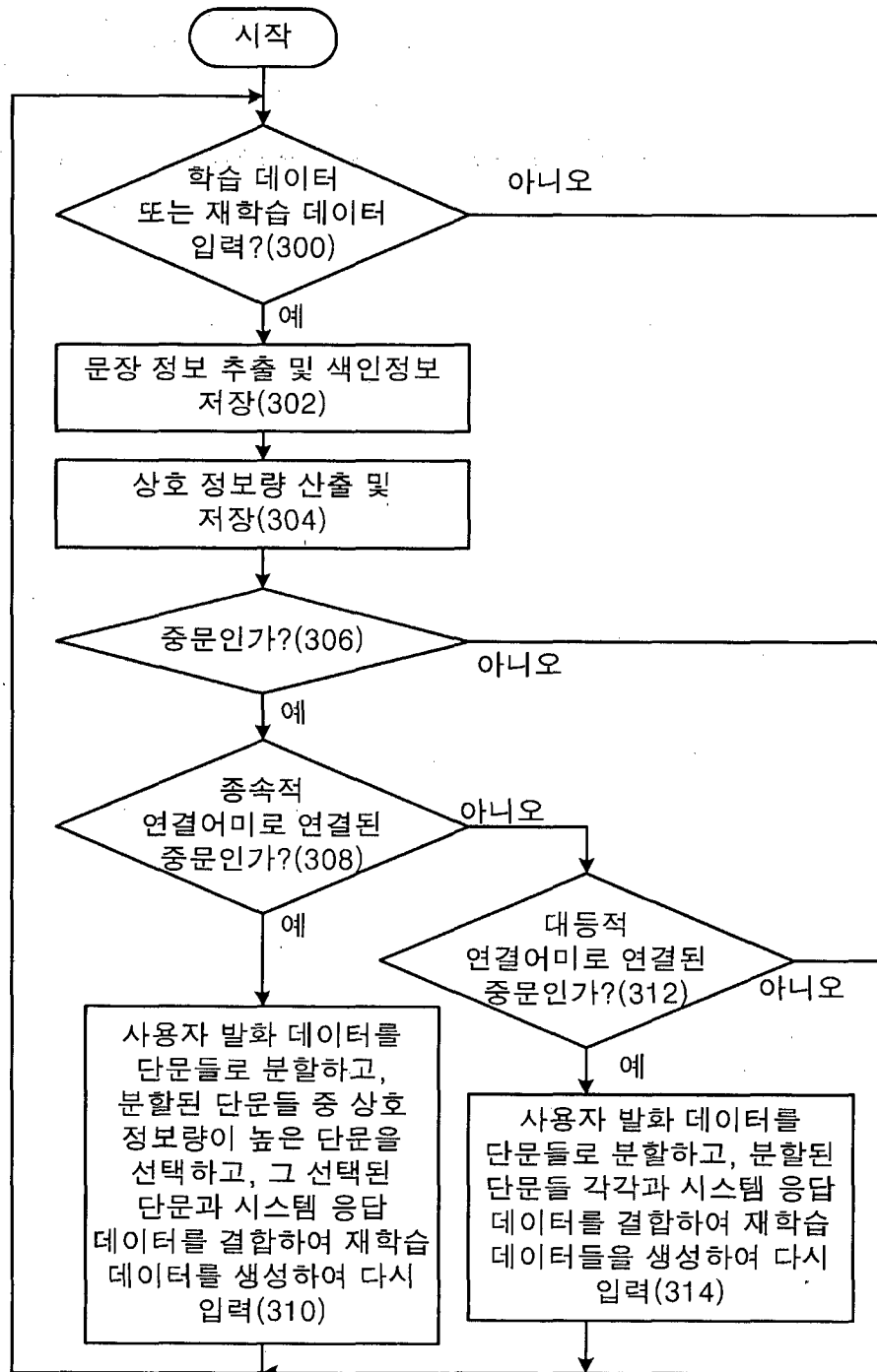
1/4

도 1

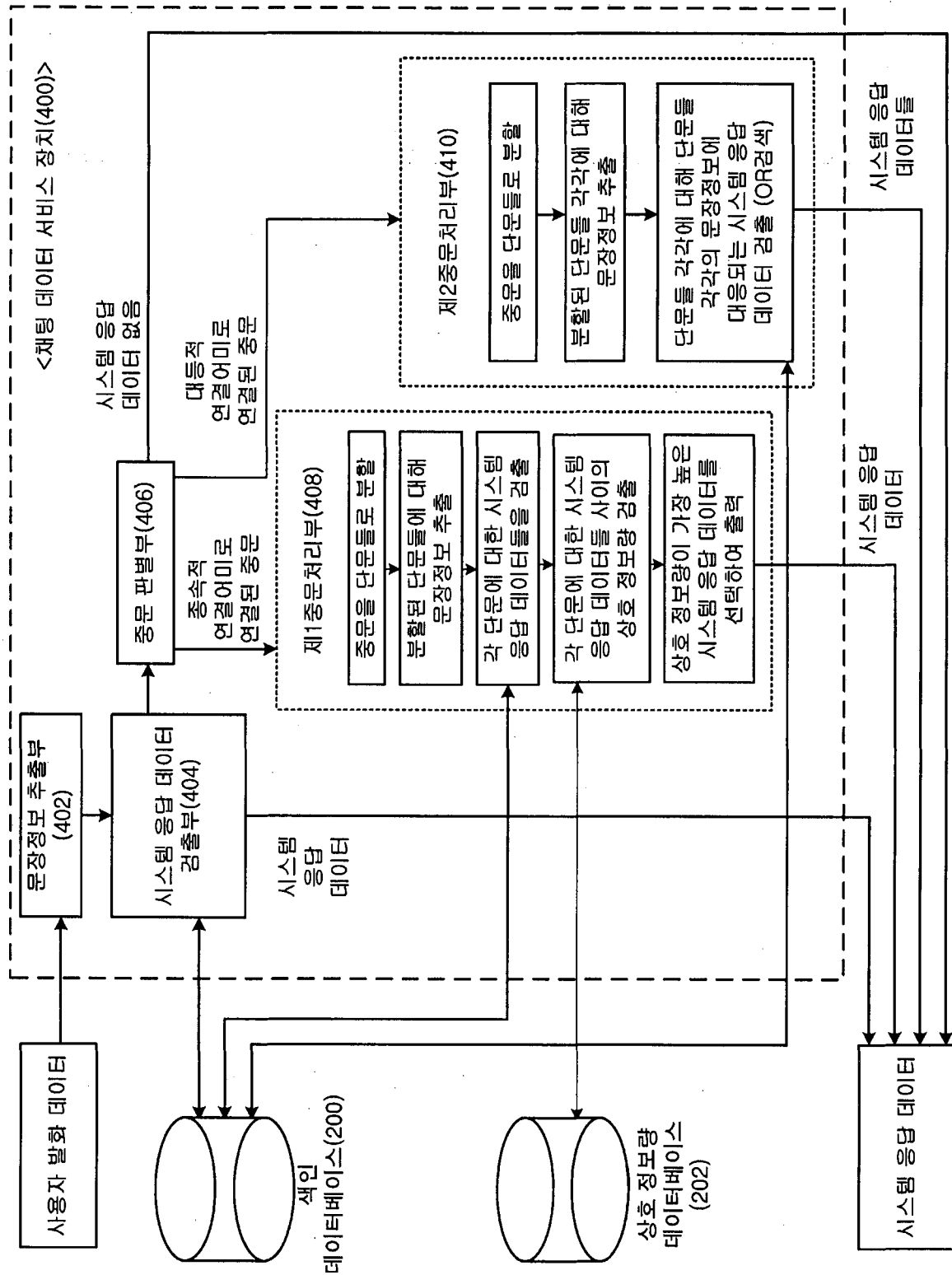


2/4

도 2

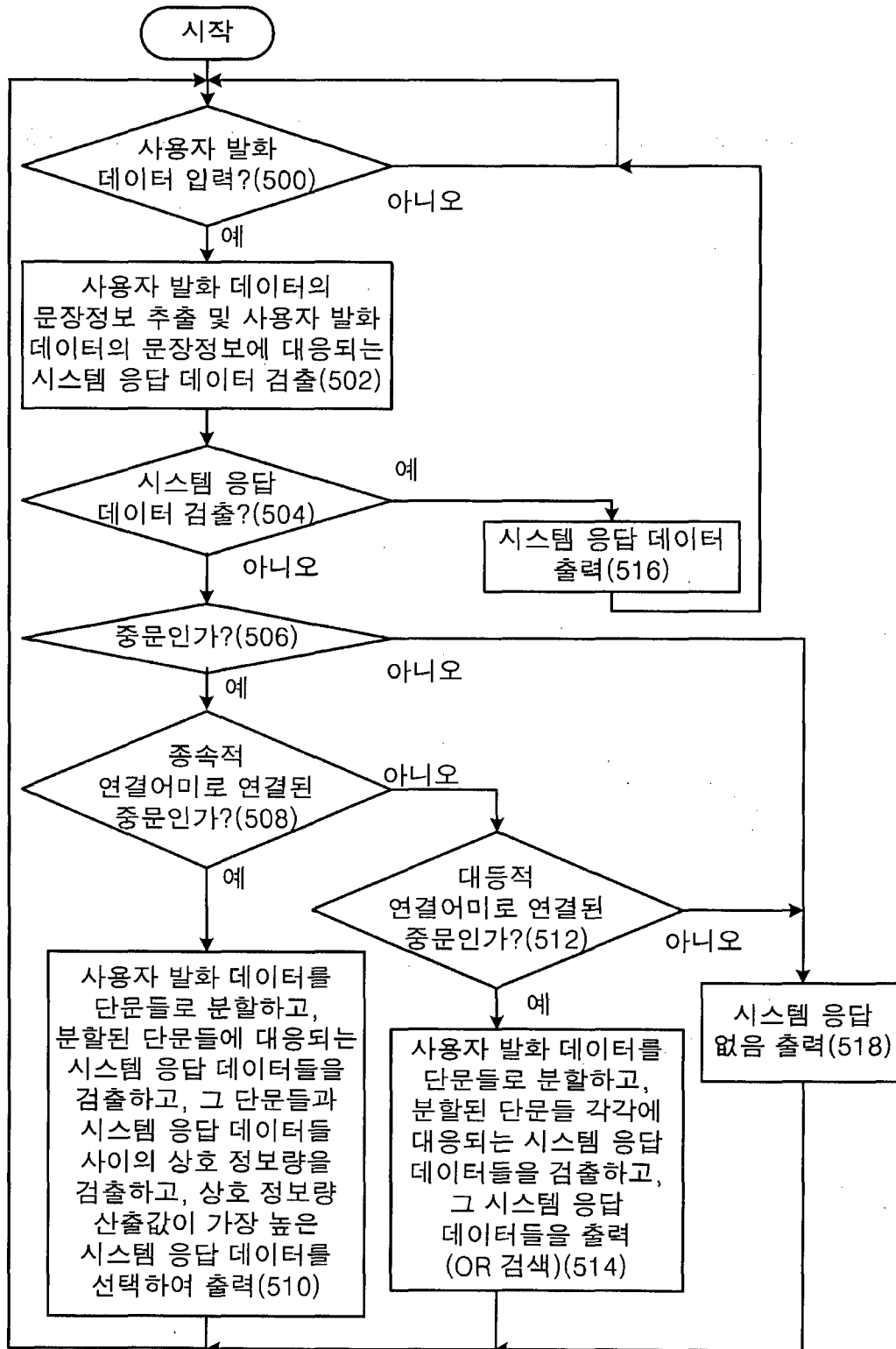


도 3



4/4

도 4



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2014/012741**A. CLASSIFICATION OF SUBJECT MATTER****G06Q 50/20(2012.01)i, G06Q 50/30(2012.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06Q 50/20; G06Q 50/30; G06F 17/30; G06F 17/27

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean Utility models and applications for Utility models: IPC as above

Japanese Utility models and applications for Utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: natural language, machine learning, chatting, conversation, articulation, morpheme, compound sentence, short sentence

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KR 10-0818979 B1 (POSTECH FOUNDATION et al.) 04 April 2008 See claims 1, 3 and paragraphs [0018]-[0024].	1-16
Y	KR 10-2004-0111715 A (NASYPNY VLADIMIR VLADIMIROVICH et al.) 31 December 2004 See claim 1 and pages 5, 16-18.	1-16
Y	KR 10-2009-0046280 A (POSTECH ACADEMY-INDUSTRY FOUNDATION) 11 May 2009 See claims 1, 4, 8.	1-16
A	JP 2000-259628 A (ATR INTERPRETING TELECOMMUNICATIONS RES LAB) 22 September 2000 See paragraphs [0056]-[0063], [0155].	1-16



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"I" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

16 MARCH 2015 (16.03.2015)

Date of mailing of the international search report

17 MARCH 2015 (17.03.2015)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex-Daejeon, 189 Seonsa-ro, Daejeon 302-701,
Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2014/012741

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-0818979 B1	04/04/2008	NONE	
KR 10-2004-0111715 A	31/12/2004	CA 2487739 A1 EP 1508861 A1 JP 2005-535007A US 2005-0071150 A1 WO 03-100659 A1	04/12/2003 23/02/2005 17/11/2005 31/03/2005 04/12/2003
KR 10-2009-0046280 A	11/05/2009	NONE	
JP 2000-259628 A	22/09/2000	JP 3016779 B1 JP 3016779 B2	06/03/2000 06/03/2000

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06Q 50/20(2012.01)i, G06Q 50/30(2012.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06Q 50/20; G06Q 50/30; G06F 17/30; G06F 17/27

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 자연어, 기계학습, 채팅, 대화, 발화, 형태소, 중문, 단문

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	KR 10-0818979 B1 (학교법인 포항공과대학교 외 1명) 2008.04.04 청구항 1, 3 및 단락 [0018]-[0024] 참조.	1-16
Y	KR 10-2004-0111715 A (나씨프니 블라디미르 블라디미로비치 외 1명) 2004.12.31 청구항 1 및 페이지 5, 16-18 참조.	1-16
Y	KR 10-2009-0046280 A (포항공과대학교 산학협력단) 2009.05.11 청구항 1, 4, 8 참조.	1-16
A	JP 2000-259628 A (ATR INTERPRETING TELECOMMUNICATIONS RES LAB) 2000.09.22 단락 [0056-0063], [0155] 참조.	1-16

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

“A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

“T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

“E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌

“X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

“L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

“Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

“O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

“&” 동일한 대응특허문헌에 속하는 문헌

“P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

국제조사의 실제 완료일

2015년 03월 16일 (16.03.2015)

국제조사보고서 발송일

2015년 03월 17일 (17.03.2015)

ISA/KR의 명칭 및 우편주소



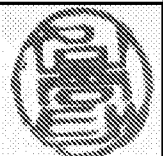
대한민국 특허청
(302-701) 대전광역시 서구 청사로 189,
4동 (둔산동, 정부대전청사)

팩스 번호 ++82 42 472 7140

심사관

김종태

전화번호 +82-42-481-5706



국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-0818979 B1	2008/04/04	없음	
KR 10-2004-0111715 A	2004/12/31	CA 2487739 A1 EP 1508861 A1 JP 2005-535007A US 2005-0071150 A1 WO 03-100659 A1	2003/12/04 2005/02/23 2005/11/17 2005/03/31 2003/12/04
KR 10-2009-0046280 A	2009/05/11	없음	
JP 2000-259628 A	2000/09/22	JP 3016779 B1 JP 3016779 B2	2000/03/06 2000/03/06