



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2020년05월11일

(11) 등록번호 10-2109219

(24) 등록일자 2020년05월04일

(51) 국제특허분류(Int. Cl.)

G06F 3/023 (2006.01)

(21) 출원번호 10-2014-7023559

(22) 출원일자(국제) 2013년01월28일

심사청구일자 2018년01월26일

(85) 번역문제출일자 2014년08월22일

(65) 공개번호 10-2014-0119763

(43) 공개일자 2014년10월10일

(86) 국제출원번호 PCT/GB2013/050182

(87) 국제공개번호 WO 2013/110955

국제공개일자 2013년08월01일

(30) 우선권주장

1201425.4 2012년01월27일 영국(GB)

1205598.4 2012년03월29일 영국(GB)

(56) 선행기술조사문헌

US20110197128 A1\*

US20070055933 A1

US20040201607 A1

WO2011107751 A2

\*는 심사관에 의하여 인용된 문헌

(73) 특허권자

터치타입 리미티드

영국 런던 에스이1 0에이엑스 서더크 브릿지 로드  
91-95

(72) 발명자

메드록 벤저민

영국 런던 에스이3 9엘엘 블랙히스 론 테러스 21

오스본 조셉 하임 베네딕트

영국 애쉬포드 켄트 티엔23 3지지 오키드 코트 4

(74) 대리인

김태홍, 김진희

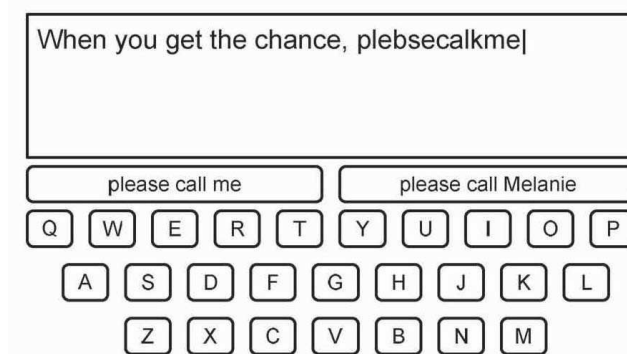
전체 청구항 수 : 총 33 항

심사관 : 박인화

(54) 발명의 명칭 사용자 데이터 입력 예측

**(57) 요약**

본 발명에 따르면, 전자 디바이스에 텍스트를 입력하기 위한 시스템이 제공된다. 상기 시스템은 입력 시퀀스(20)로부터 하나 이상의 후보들을 발생하도록 구성된 후보 발생기(2)를 포함한다. 입력 시퀀스(20)는 연속 문자 시퀀스를 포함한다. 각 후보는 하나 이상의 용어 경계들에 의해 분리된 2 이상의 용어들을 포함한다. 후보 발생기(2)는 문맥언어모델에서 상기 후보의 하나 이상의 용어들을 탐색하고, 문맥언어모델로부터 상기 후보의 하나 이상의 용어들에 해당하는 확률을 상기 후보에 할당함으로써 제 1 확률평가를 각 후보에 할당하며, 문맥언어모델은 용어 시퀀스들을 포함하며, 용어들 시퀀스 각각은 해당 발생확률을 갖는다. 후보 발생기(2)는 해당 제 1 확률평가를 기초로 하나 이상의 후보들을 폐기하도록 구성된다. 사용자 입력 시퀀스에 있는 용어 경계를 추론하는 해당 방법이 제공된다.

**대표도** - 도2a

## 명세서

### 청구범위

#### 청구항 1

전자 디바이스에 텍스트를 입력하기 위한 시스템으로서,

연속(contiguous) 문자 시퀀스를 포함하는 입력 시퀀스로 사용자 입력 신호를 변환하도록 구성된 입력 시퀀스 생성기 - 상기 입력 시퀀스 생성기는, 각 세트 내의 각 문자와 연관된 확률값이 있도록 각 문자 세트가 상기 세트 내의 문자들에 걸쳐서 확률 분포를 갖는 문자 세트의 시퀀스를 생성함으로써 상기 사용자 입력 신호를 상기 입력 시퀀스로 변환함 - ; 및

후보 생성기

를 포함하고,

상기 후보 생성기는:

상기 입력 시퀀스를 하나 이상의 경로 - 상기 하나 이상의 경로는 하나 이상의 후보에 대응하고, 각 후보는 하나 이상의 용어 경계에 의해 분리된 둘 이상의 용어를 포함함 - 를 포함하는 확률적 제약(probabilistic constrained) 시퀀스 그래프 - 상기 확률적 제약 시퀀스 그래프는 노드 및 각각 하나의 노드를 다른 하나의 노드로 연결하는 방향성 에지(directed edge)의 집단(collection)을 포함하는 방향성 비사이클 그래프(directed acyclic graph)의 변형(variant)이며, 상기 확률적 제약 시퀀스 그래프에서 각 문자 세트의 각 문자에는 노드가 할당되고, 각 노드에 대한 인커밍(incoming) 에지는 연관된 문자에 대한 확률에 대응함 - 로 변환함으로써 상기 입력 시퀀스로부터 하나 이상의 후보를 생성하고;

각 용어 시퀀스가 대응하는 발생 확률을 갖는 용어 시퀀스들을 포함하는 문맥 언어 모델에서 후보의 하나 이상의 용어를 검색하고, 상기 문맥 언어 모델로부터의 상기 후보의 하나 이상의 용어에 대응하는 상기 확률을 상기 후보에 할당함으로써, 제1 확률 추정치를 각 후보에 할당하며;

상기 제1 확률 추정치에 기초하여 하나 이상의 후보를 폐기하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 2

제1항에 있어서,

상기 하나 이상의 용어는 상기 후보를 포함하는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 3

제2항에 있어서,

각 후보에 대해, 상기 후보 생성기는 상기 문맥 언어 모델에서 상기 후보와 결합하여 문맥 시퀀스를 포함하는 시퀀스를 검색하도록 구성되고, 상기 문맥 시퀀스는 상기 입력 시퀀스에 선행하는 사용자 입력 텍스트이며, 하나 이상의 용어 경계에 의해 분리된 하나 이상의 용어를 포함하는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 4

삭제

#### 청구항 5

삭제

#### 청구항 6

삭제

#### 청구항 7

삭제

#### 청구항 8

제1항에 있어서,

상기 확률적 제약 시퀀스 그래프의 각 경로는 동일한 길이(length)가 되도록 제약되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 9

제1항에 있어서,

상기 후보 생성기는 상기 확률적 제약 시퀀스 그래프에 하나 이상의 용어 경계 노드를 삽입하도록 구성되며, 상기 하나 이상의 용어 경계 노드 각각은 상기 입력 시퀀스에서 발생 확률(t)을 갖는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 10

제9항에 있어서,

상기 후보 생성기는 상기 입력 시퀀스의 임의의 2개의 이웃하는 문자 세트들에 대응하는 임의의 2개의 이웃하는 노드들 사이에 용어 경계 노드를 삽입하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 11

제9항에 있어서,

상기 후보 생성기는 상기 입력 시퀀스의 문자 세트를 나타내는 임의의 노드에 대한 대안으로서 용어 경계 노드를 삽입하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 12

제1항에 있어서,

상기 후보 생성기는 상기 확률적 제약 시퀀스 그래프에 하나 이상의 와일드카드 노드를 삽입하도록 구성되며, 상기 하나 이상의 와일드카드 노드 각각은 상기 입력 시퀀스의 문자 세트들에 추가로 상기 확률적 제약 시퀀스 그래프에 임의의 문자가 삽입될 수 있게 하는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 13

제1항에 있어서,

상기 후보 생성기는 상기 후보를 나타내는 상기 확률적 제약 시퀀스 그래프를 통한 상기 경로의 누적 확률을 결정함으로써 상기 후보에 대한 제2 확률 추정치를 생성하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 14

제13항에 있어서,

상기 후보 생성기는 각 후보에 대해 제1 및 제2 확률을 결합하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 15

제14항에 있어서,

상기 후보 생성기는:

가장 가능성 있는 후보에 대해 상기 결합된 확률을 결정하는 것;

대상 후보에 대해 상기 결합된 확률을 결정하는 것; 및

상기 가장 가능성 있는 후보와 상기 대상 후보에 대한 결합된 확률의 비가 임계값(t) 미만인 경우 후보를 폐기하는 것에 의해, 후보를 폐기하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 16

제15항에 있어서,

상기 후보 생성기는 폐기되지 않은 임의의 후보에 대한 제3 확률을 결정하도록 구성되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 17

제16항에 있어서,

상기 제3 확률은 상기 후보와 결합하여, 상기 문맥 언어 모델에서, 문맥 시퀀스와 그것의 유효한 정자법적(orthographic)이고 어휘적(lexical)인 변형들을 포함하는 시퀀스를 검색함으로써 결정되며, 상기 문맥 시퀀스는 상기 입력 시퀀스에 선행하는 사용자 입력 텍스트이고, 하나 이상의 용어 경계에 의해 분리된 하나 이상의 용어를 포함하는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 18

제17항에 있어서,

후보의 전체 확률은 상기 제2 확률과 상기 제3 확률의 곱인 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 19

제18항에 있어서,

상기 가장 가능성 있는 후보 중 하나 이상의 후보는 상기 후보 생성기로부터 사용자에게 제시하기 위한 후보 디스플레이로 출력되는 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 20

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 용어 경계는 스페이스(space) 문자인 것인 전자 디바이스에 텍스트를 입력하기 위한 시스템.

#### 청구항 21

연속(contiguous) 문자 시퀀스를 포함하는 입력 시퀀스에서 용어 경계를 추론하는 방법으로서,

각 세트 내의 각 문자와 연관된 확률값이 있도록 각 문자 세트가 상기 세트 내의 문자들에 걸쳐서 확률 분포를 갖는 문자 세트의 시퀀스를 생성하는 단계를 포함하는, 입력 시퀀스 생성기를 사용하여 사용자 입력 신호를 상기 입력 시퀀스로 변환하는 단계;

상기 입력 시퀀스를 하나 이상의 경로 - 상기 하나 이상의 경로는 하나 이상의 후보에 대응하고, 각 후보는 하나 이상의 용어 경계에 의해 분리된 2개의 용어를 포함함 - 를 포함하는 확률적 제약(probabilistic constrained) 시퀀스 그래프 - 상기 확률적 제약 시퀀스 그래프는 노드 및 각각 하나의 노드를 다른 하나의 노드로 연결하는 방향성 에지(directed edge)의 집단(collection)을 포함하는 방향성 비사이클 그래프(directed acyclic graph)의 변형(variant)이며, 상기 확률적 제약 시퀀스 그래프에서 각 문자 세트의 각 문자에는 노드가 할당되고, 각 노드에 대한 인커밍(incoming) 에지는 연관된 문자에 대한 확률에 대응함 - 로 변환함으로써 후보 생성기를 사용하여 상기 입력 시퀀스로부터 하나 이상의 후보를 생성하는 단계;

각 용어 시퀀스가 대응하는 발생 확률을 갖는 용어 시퀀스들을 포함하는 문맥 언어 모델에서 후보의 하나 이상의 용어를 검색하고, 상기 문맥 언어 모델로부터 상기 후보의 하나 이상의 용어에 대응하는 상기 확률을 상기

후보에 할당함으로써, 제1 확률 추정치를 각 후보에 할당하는 단계; 및  
상기 제1 확률 추정치에 기초하여 하나 이상의 후보를 폐기하는 단계  
를 포함하는 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 22

제21항에 있어서,  
상기 하나 이상의 용어는 상기 후보를 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 23

제22항에 있어서,  
각 후보에 대해, 상기 문맥 언어 모델에서 상기 하나 이상의 용어를 검색하는 단계는 상기 후보와 결합하여 문맥 시퀀스를 포함하는 시퀀스를 검색하는 단계를 포함하고, 상기 문맥 시퀀스는 상기 입력 시퀀스에 선행하는 사용자 입력 텍스트이며 상기 하나 이상의 용어 경계에 의해 분리된 하나 이상의 용어를 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 24

삭제

#### 청구항 25

삭제

#### 청구항 26

삭제

#### 청구항 27

삭제

#### 청구항 28

제21항에 있어서,  
상기 확률적 제약 시퀀스 그래프의 각 경로는 동일한 길이(length)가 되도록 제약되는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 29

제21항에 있어서,  
상기 방법은, 상기 후보 생성기를 사용하여 상기 확률적 제약 시퀀스 그래프에 하나 이상의 용어 경계 노드를 삽입하는 단계를 더 포함하고, 상기 하나 이상의 용어 경계 노드 각각은 상기 입력 시퀀스에서 발생 확률(t)을 갖는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 30

제29항에 있어서,  
상기 용어 경계 노드는 상기 입력 시퀀스의 임의의 이웃하는 2개의 문자 세트들에 대응하는 임의의 2개의 이웃하는 노드들 사이에 삽입되는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 31

제29항에 있어서,  
상기 용어 경계 노드는 상기 입력 시퀀스의 문자 세트를 나타내는 임의의 노드에 대한 대안으로 삽입되는 것인

입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 32

제21항에 있어서,

상기 후보 생성기를 사용하여, 상기 확률적 제약 시퀀스 그래프에 하나 이상의 와일드카드 노드를 삽입하는 단계를 더 포함하며, 상기 하나 이상의 와일드카드 노드 각각은 상기 입력 시퀀스의 문자 세트들에 추가로 상기 확률적 제약 시퀀스 그래프에 임의의 문자가 삽입될 수 있게 하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 33

제21항에 있어서,

상기 후보 생성기를 사용하여, 상기 후보를 나타내는 상기 확률적 제약 시퀀스 그래프를 통한 상기 경로의 누적 확률을 결정함으로써 상기 후보에 대한 제2 확률 추정치를 생성하는 단계를 더 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 34

제33항에 있어서,

상기 후보 생성기를 사용하여, 각 후보에 대한 제1 및 제2 확률의 값을 결정하는 단계를 더 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 35

제34항에 있어서,

상기 후보를 폐기하는 단계는:

상기 후보 생성기를 사용하여, 가장 가능성 있는 후보에 대한 결합된 확률을 결정하는 단계;

상기 후보 생성기를 사용하여, 대상 후보에 대한 결합된 확률을 결정하는 단계; 및

상기 후보 생성기를 사용하여, 상기 가장 가능성 있는 후보와 상기 대상 후보에 대한 결합된 확률의 비가 임계값( $t$ ) 미만인 경우 후보를 폐기하는 단계를 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 36

제35항에 있어서,

상기 후보 생성기를 사용하여, 폐기되지 않은 임의의 후보에 대한 제3 확률을 결정하는 단계를 더 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 37

제36항에 있어서,

상기 제3 확률은 상기 후보와 결합하여, 상기 문맥 언어 모델에서, 문맥 시퀀스 및 그것의 유효한 정자법적(orthographic)이고 어휘적(lexical)인 변형들을 포함하는 시퀀스를 검색함으로써 결정되며, 상기 문맥 시퀀스는 상기 입력 시퀀스에 선행하는 사용자 입력 텍스트이고, 하나 이상의 용어 경계에 의해 분리된 하나 이상의 용어를 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 38

제 37 항에 있어서,

후보의 전체 확률은 상기 제2 확률에 상기 제3 확률을 곱함으로써 결정되는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

### 청구항 39

제38항에 있어서,

상기 후보 생성기로부터 상기 가장 가능성 있는 후보 중 하나 이상의 후보를 사용자에게 제시하기 위한 후보 디스플레이로 출력하는 단계를 더 포함하는 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 40

제21항 내지 제23항 중 어느 한 항에 있어서,

상기 용어 경계는 스페이스(space) 문자인 것인 입력 시퀀스에서 용어 경계를 추론하는 방법.

#### 청구항 41

프로세서가 제21항 내지 제23항 중 어느 한 항에 따른 방법을 수행하게 하는 컴퓨터 프로그램 수단이 저장된 컴퓨터 판독가능 기록 매체.

#### 청구항 42

삭제

#### 청구항 43

삭제

#### 청구항 44

삭제

#### 청구항 45

삭제

#### 청구항 46

삭제

### 발명의 설명

#### 기술 분야

[0001] 본 발명은 사용자 입력 텍스트로부터 다수의 용어들을 예측하는 시스템 및 방법에 관한 것으로, 보다 상세하게는 사용자 입력 텍스트에서 하나 이상의 용어 경계들을 예측하는 시스템 및 방법에 관한 것이다.

#### 배경 기술

[0002] 사용자 입력 텍스트에서 용어 경계를 정의하는 몇몇 형태, 가령 Google<sup>TM</sup>의 Android<sup>TM</sup> ICS 키보드 또는 Apple<sup>TM</sup>의 iOS 키보드를 갖는 시스템들이 공지되어 있다. 그러나, 이들 시스템들은 용어 경계를 검출하는데 한계가 있다.

[0003] Google<sup>TM</sup>의 Android<sup>TM</sup> ICS 키보드를 예를 들면, 사용자가 분리 공백 또는 공백과는 다른 문자로 이격시키지 않고 기본적인 어휘로부터 2개의 유효 단어를 입력할 경우, 시스템은 관련 있다면 틀린 문자를 대체하며 공백에 의해 분리된 2 단어들을 제공할 것이다. 그러나, 이 시스템은 많은 제한들이 있는데, 특히

[0004] · 어느 한 단어가 오타 또는 오자인 경우 작동하지 않고,

[0005] · 자동 아포스트로피(') 삽입과 같은 다른 입력조작 기능들과 작동하지 않으며,

[0006] · 2개의 단어들이 일단 분석되게 한다. 즉 단일어 경계만 식별할 수 있다.

[0007] Apple<sup>TM</sup>의 iOS 키보드는 제한된 오타 또는 오자를 보정할 수 있다는 점에서 Google<sup>TM</sup>의 Android<sup>TM</sup> 버전보다 더 이점적이다. 그러나, iOS 키보드는 또한 단일어 경계를 식별하는데 제한된다.

## 발명의 내용

### 해결하려는 과제

[0008] 본 발명의 목적은 이들 하나 이상의 제한들을 극복해, 용어 경계들의 명백한 삽입없이 사용자가 구 또는 심지어 전체 메시지들을 입력하게 하는 것이다.

### 과제의 해결 수단

- [0009] 본 발명의 제 1 태양으로,
- [0010] 연속 문자 시퀀스를 포함한 입력 시퀀스로부터 하나 이상의 용어 경계들에 의해 분리된 2 이상의 용어들을 각각 포함하는 하나 이상의 후보들을 발생하고,
- [0011] 해당 발생확률을 각각 갖는 용어들의 시퀀스를 포함한 문맥언어모델에서 상기 후보의 하나 이상의 용어들을 탐색하고, 문맥언어모델로부터 상기 후보의 하나 이상의 용어들에 해당하는 확률을 상기 후보에 할당함으로써 제 1 확률평가를 각 후보에 할당하며,
- [0012] 해당 제 1 확률평가를 기초로 하나 이상의 후보들을 폐기하도록 구성된 후보 발생기를 포함하는 전자 디바이스에 텍스트를 입력하기 위한 시스템이 있다.
- [0013] 일실시예로, 하나 이상의 용어들은 후보를 포함한다. 바람직하기로, 각 후보에 대해, 후보 발생기는 상기 후보와 조합한 문맥 시퀀스를 포함한 시퀀스에 대한 문맥언어모델에서 탐색하도록 구성되고, 문맥 시퀀스는 입력 시퀀스보다 앞선 사용자 입력 텍스트이며 하나 이상의 용어 경계들에 의해 분리된 하나 이상의 용어들을 포함한다.
- [0014] 시스템은 사용자 입력 신호들을 입력 시퀀스와 문맥 시퀀스로 변환시키도록 구성된 입력 시퀀스 발생기를 더 구비할 수 있다. 입력 시퀀스 발생기는 바람직하게는 각 세트내 각 문자와 관련된 확률값이 있도록 일련의 문자 세트들을 발생함으로써 사용자 입력 신호를 입력 시퀀스로 변환시키며, 각 문자 세트는 세트내 문자들에 대한 확률 분포를 갖는다.
- [0015] 후보 발생기는 입력 시퀀스를 하나 이상의 경로들을 포함한 확률적 제약 시퀀스 그래프로 변환시킴으로써 입력 시퀀스로부터 하나 이상의 후보들을 발생하도록 구성되고, 하나 이상의 경로들은 하나 이상의 후보들에 해당한다. 확률적 제약 시퀀스 그래프는 바람직하게는 노드 및 한 노드를 또 다른 노드에 각각 연결하는 방향성 에지의 수집을 포함한 방향성 비사이클 그래프의 변형이며, 상기 확률적 제약 시퀀스 그래프에서, 각 문자 세트의 각 문자가 노드에 할당되고, 각 노드에 대한 인커밍 에지는 관련된 문자에 대한 확률에 해당한다. 확률 그래프의 각 경로는 바람직하게는 길이가 같도록 제한된다.
- [0016] 후보 발생기는 바람직하게는 하나 이상의 용어 경계 노드들을 확률적 제약 시퀀스 그래프에 삽입하도록 구성되고, 하나 이상의 용어 경계 노드들 각각은 입력 시퀀스에서 발생하는 확률(t)을 갖는다. 바람직하기로, 후보 발생기는 입력 시퀀스의 문자들의 임의의 2개 이웃 세트들에 해당하는 임의의 2개 이웃 노드들 사이에 용어경계 노드를 삽입하도록 구성된다. 더욱이, 후보 발생기는 입력 시퀀스의 문자 세트를 나타내는 임의의 노드에 대한 대안으로서 용어경계 노드를 삽입하도록 구성될 수 있다.
- [0017] 후보 발생기는 바람직하게는 하나 이상의 와일드카드(wildcard) 노드들을 확률적 제약 시퀀스 그래프에 삽입하도록 구성되고, 하나 이상의 와일드카드 노드들 각각은 임의의 문자가 입력 시퀀스의 문자 세트들 이외에 확률적 제약 시퀀스 그래프에 삽입되게 한다.
- [0018] 후보 발생기는 후보를 나타내는 확률적 제약 시퀀스 그래프를 통해 경로의 누적 확률을 판단함으로써 후보에 대한 제 2 확률평가를 발생하도록 구성될 수 있다. 후보 발생기는 바람직하게는 각 후보에 대한 제 1 및 제 2 확률을 조합하도록 구성된다.
- [0019] 바람직하기로, 후보 발생기는:
- [0020] 가장 가능성 있는 후보에 대한 조합된 확률을 판단하는 단계;
- [0021] 해당 후보에 대한 조합된 확률을 판단하는 단계; 및
- [0022] 가장 가능성 있는 후보와 해당 후보에 대한 조합된 확률의 비가 임계치(t) 미만인 경우 후보를 폐기하는 단계에 의해 후보를 폐기하도록 구성된다.



- [0023] 후보 발생기는 폐기되지 않은 임의 후보에 대한 제 3 확률을 판단하도록 구성될 수 있다. 제 3 확률은 문맥 시퀀스, 유효한 정자법 및 그 어휘 변화를 포함한 시퀀스에 대한 문맥언어모델을 탐색함으로써 결정될 수 있으며, 문맥 시퀀스는 입력 시퀀스에 앞선 사용자 입력 텍스트이고 하나 이상의 용어 경계들에 의해 분리된 하나 이상의 용어들을 포함한다. 후보의 전체 확률은 제 2 및 제 3 확률의 곱일 수 있다.
- [0024] 하나 이상의 가장 가능성 있는 후보들은 사용자에게 나타내기 위해 후보 발생기로부터 후보 디스플레이로 출력될 수 있다.
- [0025] 일실시예에서, 용어 경계는 공백문자이다.
- [0026] 본 발명의 제 2 태양으로,
- [0027] 후보 발생기를 이용해 입력 시퀀스로부터 하나 이상의 용어 경계들에 의해 분리된 2 이상의 용어들을 각각 포함하는 하나 이상의 후보들을 발생하는 단계,
- [0028] 해당 발생확률을 각각 갖는 용어들의 시퀀스를 포함한 문맥언어모델에서 상기 후보의 하나 이상의 용어들을 탐색하고, 문맥언어모델로부터 상기 후보의 하나 이상의 용어들에 해당하는 확률을 상기 후보에 할당함으로써 제 1 확률평가를 각 후보에 할당하는 단계,; 및
- [0029] 해당 제 1 확률평가를 기초로 하나 이상의 후보들을 폐기하는 단계를 포함하는, 연속 문자 시퀀스를 포함한 입력 시퀀스에서 용어 경계를 추론하는 방법이 있다.
- [0030] 일실시예에서, 하나 이상의 용어들은 후보를 포함한다. 각 후보에 대해, 문맥언어모델에서 후보를 탐색하는 단계는 바람직하게는 후보와 조합한 문맥 시퀀스를 포함한 시퀀스에 대해 탐색하는 단계를 포함하고, 문맥 시퀀스는 입력 시퀀스보다 앞선 사용자 입력 텍스트이며 하나 이상의 용어 경계들에 의해 분리된 하나 이상의 용어들을 포함한다.
- [0031] 상기 방법은 바람직하게는 입력 시퀀스 발생기를 이용해, 사용자 입력 신호들을 입력 시퀀스와 문맥 시퀀스로 변환시키는 단계를 포함한다. 사용자 입력 신호를 입력 시퀀스로 변환시키는 단계는 바람직하게는 각 세트내 각 문자와 관련된 확률값이 있도록 일련의 문자 세트들을 발생하는 단계를 포함하며, 각 문자 세트는 세트내 문자들에 대한 확률 분포를 갖는다.
- [0032] 입력 시퀀스로부터 하나 이상의 후보들을 발생하는 단계는 입력 시퀀스를 하나 이상의 경로들을 포함한 확률적 제약 시퀀스 그래프로 변환시키는 단계를 포함할 수 있고, 하나 이상의 경로들은 하나 이상의 후보들에 해당한다. 확률적 제약 시퀀스 그래프는 바람직하게는 노드 및 한 노드를 또 다른 노드에 각각 연결하는 방향성 에지의 수집을 포함한 방향성 비사이클 그래프의 변형이며, 상기 확률적 제약 시퀀스 그래프에서, 각 문자 세트의 각 문자가 노드에 할당되고, 각 노드에 대한 인커밍 에지는 관련된 문자에 대한 확률에 해당한다. 확률 그래프의 각 경로는 바람직하게는 길이가 같도록 제한된다.
- [0033] 상기 방법은 바람직하게는 하나 이상의 용어 경계 노드들을 확률적 제약 시퀀스 그래프에 삽입하는 단계를 더 포함하고, 하나 이상의 용어 경계 노드들 각각은 입력 시퀀스에서 발생하는 확률(t)을 갖는다. 입력 시퀀스의 문자들의 임의의 2개 이웃 세트들에 해당하는 임의의 2개 이웃 노드들 사이에 용어경계 노드가 바람직하게 삽입된다.
- [0034] 입력 시퀀스의 문자 세트를 나타내는 임의 노드에 대한 대안으로서 용어경계 노드가 삽입될 수 있다.
- [0035] 상기 방법은 바람직하게는 후보 발생기를 이용해 하나 이상의 와일드카드 노드들을 확률적 제약 시퀀스 그래프에 삽입하는 단계를 더 포함하고, 하나 이상의 와일드카드 노드들 각각은 임의 문자가 입력 시퀀스의 문자 세트들 이외에 확률적 제약 시퀀스 그래프에 삽입되게 한다.
- [0036] 상기 방법은 바람직하게는 후보를 나타내는 확률적 제약 시퀀스 그래프를 통해 경로의 누적 확률을 판단함으로써 후보 발생기를 이용해 후보에 대한 제 2 확률평가를 발생하는 단계를 더 포함한다. 후보 발생기는 각 후보에 대한 제 1 및 제 2 확률의 곱을 판단하는데 사용될 수 있다.
- [0037] 바람직하기로, 후보를 폐기하는 단계는:
- [0038] 후보 발생기를 이용해 가장 가능성 있는 후보에 대한 조합된 확률을 판단하는 단계;
- [0039] 후보 발생기를 이용해 해당 후보에 대한 조합된 확률을 판단하는 단계; 및

- [0040] 후보 발생기를 이용해 가장 가능성 있는 후보와 해당 후보에 대한 조합된 확률의 비가 임계치(t) 미만인 경우 후보를 폐기하는 단계에 의해 후보를 폐기하는 단계를 포함하는 연속 문자 시퀀스를 포함한다.
- [0041] 바람직하기로, 상기 방법은 후보 발생기를 이용해 폐기되지 않은 임의 후보에 대한 제 3 확률을 판단하는 단계를 더 포함한다. 제 3 확률은 후보 발생기를 이용해 상기 후보와 조합해 문맥 시퀀스, 유효한 정자법 및 그 어휘 변화를 포함한 시퀀스에 대한 문맥언어모델을 탐색함으로써 결정될 수 있으며, 문맥 시퀀스는 입력 시퀀스에 앞선 사용자 입력 텍스트이고 하나 이상의 용어 경계들에 의해 분리된 하나 이상의 용어들을 포함한다. 후보의 전체 확률은 제 2 확률에 제 3 확률을 곱함으로써 결정될 수 있다.
- [0042] 상기 방법은 후보 발생기로부터 하나 이상의 가장 가능성 있는 후보들을 사용자에게 나타내기 위해 후보 디스플레이로 출력하는 단계를 더 포함할 수 있다.
- [0043] 일실시예에서, 용어 경계는 공백문자이다.
- [0044] 본 발명의 제 3 태양으로, 프로세서가 상술한 바와 같은 방법을 수행하게 하는 컴퓨터 프로그램 수단을 저장한 컴퓨터 판독가능 매체를 구비하는 컴퓨터 프로그램 제품이 있다.
- [0045] 본 발명의 제 4 태양으로,
- [0046] 현재 사용자가 기입하고 있는 텍스트를 디스플레이하도록 구성된 텍스트판;
- [0047] 사용자가 현재 기입하고 있는 텍스트가 나타내고자 의도한 것의 예측인 후보를 디스플레이하기 위한 예측판;
- [0048] 사용자가 기입한 텍스트를 받아들이기 위한 것으로, 스페이스바를 나타내는 키를 포함하지 않는 가상 키보드를 구비한 사용자 인터페이스가 있다.
- [0049] 바람직하기로, 예측판은 사용자 입력 시퀀스를 디스플레이하는 제 1 예측판과 사용자 입력 시퀀스에 대한 예측을 디스플레이하는 제 2 예측판의 2개의 예측 판을 포함한다.

### 발명의 효과

- [0050] 본 발명의 내용에 포함됨.

### 도면의 간단한 설명

- [0051] 첨부도면을 참조로 본 발명을 더 상세히 설명한다:
- 도 1은 본 발명에 따른 고수준의 예측구조의 개략도이다.
- 도 2a는 본 발명에 따른 예시적인 인터페이스를 도시한 것이다.
- 도 2b는 본 발명에 따른 예시적인 인터페이스를 도시한 것이다.
- 도 3은 본 발명에 따른 시스템의 사용 작동 예이다.
- 도 4는 본 발명에 따른 방법의 흐름도이다.

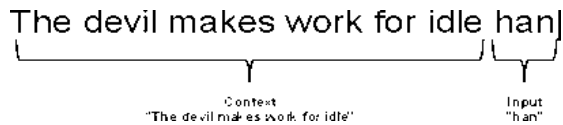
### 발명을 실시하기 위한 구체적인 내용

- [0052] 본 발명은 휴대폰 또는 태블릿과 같은 전자 디바이스에 텍스트 기입을 위한 시스템 및 방법을 제공함으로써, 텍스트 입력시 용어 경계들의 가능한 위치를 추론하기 위한 확률문맥모델이 이용된다.
- [0053] 본 시스템은 문맥언어 데이터를 기초로 이들 결과들을 필터링하는 확률방법으로 기본 어휘로부터 다용어 경계추론 결과들을 발생하는 방법을 포함한다.
- [0054] 이 시스템은 바람직하게는 (본 명세서에 전체가 참조로 함체된) 국제특허출원번호 PCT/GB2011/001419에 기술된 바와 같은 추가 확률문자조작과 결합되어, 동시에 오타 및 오자를 수정할 수 있고 또한 (영어의 자동 아포스트로피 삽입과 같은) 확률적으로 도출된 다른 문자 스트림을 조작할 수 있는 시스템을 만든다. 따라서, 시스템은 높은 확률의 다용어 텍스트 예측을 발생할(및 사용자 또는 시스템에 제공할) 수 있다.
- [0055] 따라서, 본 시스템 및 방법은 사용자가 수동으로 단어들 이격시킬 필요성을 제거함으로써 입력속도를 높이는 수단, 및 이렇게 하기 위해 시도하나 (사용자가 이격 대신 가짜 문자를 삽입했거나, 사용자가 가령 임의의 문자 또는 분리를 삽입하는데 실패한 경우) 실패한 사용자에게 정확한 피드백을 모두 제공한다.

- [0056] 다음 예들은 예시적인 많은 사용자 텍스트 입력들이 주어진 경우 본 발명의 시스템 및 방법의 출력(하나 이상의 용어들의 텍스트 예측)을 설명한다. 예들은 용어 경계들이 전혀 없는 사용자 입력 텍스트, 경계 대신에 가짜 문자들을 가진 텍스트 입력 등과 같은 다른 많은 텍스트 입력 시나리오들에 대한 시스템/방법의 능력을 설명한다.
- [0057] 용어 경계가 없는 사용자에게 의한 텍스트 입력의 예:
- [0058] 사용자 입력 "seeyoulater"에 대해, 시스템/방법은 "see you later"를 예측/출력한다.
- [0059] 경계 대신 가짜 문자로 사용자에게 의한 텍스트 입력의 예:
- [0060] 사용자 입력 "seecyoublater"에 대해, 시스템/방법은 "see you later"를 예측/출력한다.
- [0061] 개시된 시스템 및 방법은 또한 실시간으로 접두사 매치들을 해결할 수 있다. 가령:
- [0062] 사용자 입력 "seeyoula"에 대해, 시스템/방법은 "see you later"를 예측/출력한다.
- [0063] 개시된 시스템/방법은 동시에 단어 경계 및 접두사를 추론하면서 오타/오자를 보정할 수 있다. 가령:
- [0064] 사용자 입력 "seeyoulayer"에 대해, 시스템/방법은 "see you later"를 예측/출력한다.
- [0065] 사용자 입력 "whatstheti"에 대해, 시스템/방법은 "what's the time"를 예측/출력한다.
- [0066] 시스템/방법은 사용자에게 표현하기 위해 확률적으로 랭킹된 하나 이상의 예측 세트를 만들도록 구성될 수 있다. 예컨대, 사용자 입력 "thedrmin"으로부터, 시스템/방법은 다음 랭킹 예측을 출력할 수 있다:
- [0067] "these mints
- [0068] "the seminar
- [0069] <etc>
- [0070] 일반적으로, 하지만 비배타적으로, 본 시스템의 개요가 도 1에 도시되어 있다.
- [0071] 시스템은 바람직하게는 본 명세서에 전체가 참조로 합체된 국제특허출원번호 PCT/GB2011/001419에 기술된 바와 같은 기존 확률 텍스트 기입 시스템을 강화하도록 설계된다. 그러나, 이는 가령 본 출원의 배경부분에 언급된 바와 같은 기존 텍스트 기입 시스템을 단지 강화하기 위해서만 이용될 수 있다.
- [0072] 도 1에 도시된 바와 같이, 시스템은 후보 발생기(2)를 포함한다. 바람직하기로, 시스템은 또한 시퀀스 발생기(1) 및 후보 디스플레이(3)를 포함한다. 후보 디스플레이(3)는 하나 이상의 후보(40)를 사용자에게 디스플레이하도록 구성된다.
- [0073] 입력 시퀀스 발생기(1)는 사용자 입력으로부터 텍스트 시퀀스를 발생하도록 구성된다. 사용자 입력은 가령 다른 타입의 텍스트 입력 인터페이스가 주어진다면 일련의 다른 전자 디바이스들 범위와 함께 다른 타입의 인터랙션에 해당하는 일련의 사용자 신호들(10)에 의해 발생할 수 있다:
- [0074] · 해당기술분야에 공지된 바와 같은 QWERTY(또는 다른 레이아웃)
- [0075] · 해당기술분야에 공지된 바와 같은 확률적 가상 QWERTY(또는 다른 레이아웃)
- [0076] · 보다 최신의 키프레스 모델(가령, GB 출원 1108200.5, "User Input Prediction"로, 상기 참조문헌은 본 명세서에 참조로 합체되어 있음)
- [0077] · 확률연속 스트로크 인터페이스(가령, GB 출원 1200643.3, "System and method for inputting text"로, 상기 참조문헌은 본 명세서에 참조로 합체되어 있음)
- [0078] 입력 시퀀스 발생기(1)는 사용자에게 의해 발생한 입력 신호(10)를 받고 소정의 언어로 구조화된 문자 시퀀스를 되돌려 보낸다. 입력 시퀀스 발생기(1)는 2개의 구조화된 문자 시퀀스, 즉, 입력 시퀀스(20) 및 문맥 시퀀스(30)를 출력한다.
- [0079] 문맥 시퀀스(30)는 고정된 문자 시퀀스를 포함한다. 입력 시퀀스(20)는 본 명세서에서 확률 스트링이라고 하는 형태이다. 확률 스트링은 문자요소들에 대한 확률분포를 각각 갖는 가지며 일련의 문자 세트를 포함하므로, 각 세트에서 각 문자는 이와 관련된 확률값을 갖는다. 확률 스트링이 합이 1이 되는 해당 확률을 갖는 2 이상의 문자들을 각각 갖는 3개 문자 세트들( {-}로 표시됨)에 대한 시퀀스를 포함하는 예가 아래에 도시되어 있다:

[0080]  $\{\{ \{a, 0.8\}, \{s, 0.1\}, \{z, 0.1\} \}, \{ \{r, 0.9\}, \{t, 0.05\}, \{f, 0.05\} \}, \{ \{e, 0.8\}, \{w, 0.2\} \} \}$

[0081] 예로써, 사용자는 시퀀스 "The devil makes work for idle han"를 입력하면, 커서는 "han" 바로 다음에 위치된다. 이 예에서, 문맥 시퀀스(30)는 단어 "idle"까지 포함한 문자들일 테고, 입력 시퀀스(20)는 글자 'h' 'a' 및 'n'으로 의도했을 수 있는 문자들에 대한 분포일 것이다.



[0082]

[0083] 일반적으로, 문맥 시퀀스(30)는 사용자가 이미 적어둔 텍스트로 간주될 수 있다고, 입력 시퀀스(20)는 사용자가 현재 입력하려고 하는 텍스트인 것으로 간주될 수 있다. 입력 시퀀스(20)는 연속 문자 시퀀스이다. 연속 문자 시퀀스는 명백한 용어 경계들이 없는 연속 문자 시퀀스이다. 후보 발생기(2)는 공백이 있는 여러 문자, 하나 이상의 용어들의 교정, 및/또는 구두점 삽입의 교체를 포함할 수 있는 2 이상의 용어들 간에 하나 이상의 추론된 용어 경계들을 갖는 2 이상의 용어들의 예측으로 전체 입력 시퀀스(20)에 대한 대체를 제공하도록 구성된다.

[0084] 예컨대, 사용자가 이미 시스템에 어법 "The devil makes"를 적고 현재 "workforidlehan"을 입력하고 있다면, 문맥 시퀀스(30)는 "The devil makes"인 반면, 입력 시퀀스(20)는 "workforidlehan"에 해당한다.

[0085] 따라서, 입력 시퀀스 발생기(1)는 (예컨대, 후보 예측(40)의 이전 사용자 선택(50)에 의해) 사용자가 시스템에 적은 텍스트에 해당하는 문맥 시퀀스(30) 및 연속 문자 시퀀스의 형태로 사용자가 현재 입력하고 있으나 시스템이 이제 처리해야하는 텍스트에 해당하는 입력 시퀀스(20)를 발생하도록 구성된다.

[0086] 문맥 및 입력 시퀀스(30,20)는 입력 시퀀스 발생기(1)로부터 출력되고 후보 발생기(2)에 의해 입력으로서 받아들여진다.

[0087] 후보 발생기(2)는 용어 경계들이 필요한 경우 추론되는 다용어 시퀀스들에 각각 해당하는 하나 이상의 후보예측(40)을 발생하도록 구성된다. 이를 달성하기 위해, 후보 발생기(2)는 사용자가 이들을 의도했거나 의도한 가능성에 의해 시퀀스를 랭크하도록 구성된다.

[0088] 이는 하기의 식에 의해 지배되는 세트로 시퀀스에 대한 랭킹과 같은 것으로 나타낸다:

### 수학식 1

[0089] 
$$P(s \in S \mid e, M)$$

[0090] 여기서 e는 관찰된 증거이고, M은 확률 추론을 하는데 사용될 수 있는 훈련모델 세트이다. 다시 말하면, 시스템은 예측이 도출될 수 있는 모든 시퀀스 세트에 대해 증거가 주어진 시퀀스의 조건부 확률을 평가할 것이다. 수학식(1)은 베이즈 정리(Bayes Theorem)를 이용해 다음과 같이 나타나도록 재배열된다:

### 수학식 2

[0091] 
$$\frac{P(e \mid s, M)P(s \mid M)}{P(e \mid M)}$$

[0092] 분모에서 타겟 시퀀스에 대한 주변화(Marginalising)는 다음과 같이 나타낸다:

### 수학식 3

$$\frac{P(e | s, M)P(s | M)}{\sum_{j=1}^{|S|} P(e | s_j, M)P(s_j | M)}$$

[0093]

[0094]  $P(e | s, M)$ 을 계산하기 위해, 타겟 시퀀스가 주어지면 관련된 모델 하에서 몇몇 분포와는 무관하게 발생하는 분해 세트  $[e_1, \dots, e_N]$ 로 분리될 수 있다고 가정된다. 따라서, 다음의 가정이 이루어진다:

[0095] 가정 1: 각 세트에 있는 증거는 타겟 시퀀스가 주어지면 모든 다른 세트들에 조건상 독립적이도록 별개 세트들로 나누어질 수 있다.

[0096] 각  $e_i$ 는 이와 관련된 모델( $M_i$ )을 갖는다.

[0097] 이 독립 가정은 다음과 같이 수학적으로 표현된다:

### 수학식 4

$$P(e | s, M) = \prod_{i=1}^N [P(e_i | s, M_i \in M)]$$

[0098]

[0099] 이는 적절한 모델을 각각 갖는 많은 증거 소스들이 결합되게 하는 프레임 워크의 구성이 조합되게 한다.

[0100] 바람직하기로, 모델  $R \in M$ 은 타겟 시퀀스 우선순위와 관련있다. 이것 및 상기 가정이 주어지면, 수학식 3은 다음으로 재진술될 수 있다:

### 수학식 5

$$\frac{P(s | R) \prod_{i=1}^N P(e_i | s, M_i)}{\sum_{j=1}^{|S|} P(s_j | R) \prod_{i=1}^N P(e_i | s_j, M_i)}$$

[0101]

[0102] 수학식 5에서 분모는  $s$ 에 대해 상수이고, 따라서 랭킹에 전혀 영향을 주지 않으며 단지 최종 발생한 확률값에 대한 정규화 인수로 역할한다. 개시된 시스템의 구현시, 상기 값은 정규화 확률값을 제공하도록 하기의 수학식 11을 참조로 기술된 바와 같은 근사로 평가될 수 있다. 그러나, 랭킹에 영향을 주지 않기 때문에, 본 발명의 시스템은 텍스트 예측의 랭킹이 충분할 수 있기에 반드시 분모를 평가할 필요가 없다.

[0103] 상기 수학식(5)를 가정하면, 타겟 시퀀스 우선순위  $P(s | R)$ 는 각 증거 소스( $e_i$ ) 및 관련된 모델( $M_i$ )에 대한 각각의 증거 유사도  $P(e_i | s, M_i)$ 과 함께 계산되어야 한다.

[0104] 개시된 시스템은 증거 후보 해석에 대한 주변화함으로써 확률 프레임워크내 증거 가능성을 계산하는 2가지 방법; 후보 모델 1 및 후보 모델 2를 제공한다.

[0105] **후보 모델 1**

[0106] 하나의 증거소스로부터 증거의 유사도 평가 형성시, 이는 '사용자가 의도한' 시퀀스와 관찰된 증거 사이 중간

단계들인 '후보들'에 대해 모델을 표현하는데 종종 도움이 된다. 유사도는 다음과 같이 재작성될 수 있다:

### 수학식 6

$$P(e|s, M) = \sum_{j=1}^K P(e|c_j, s, M_{candidate}) P(c_j|s, M_{sequence})$$

[0107]

[0108] 여기서,  $c_j$ 는 단일 후보이고, 주어진 증거소스에 대해  $M$ 의 2개 서브모델들, 즉, 후보모델  $M_{candidate}$  및 시퀀스 모델  $M_{sequence}$ 을 갖는다. 여기서 핵심 가정은 다음과 같다:

[0109] 가정 2: 해당 모델 하에서 유사도는 후보들에 대한 주변화로 표현될 수 있고, 후보가 주어지면, 증거는 타겟 시퀀스와 조건상 별개이다.

[0110] 이 가정을 적용하면,  $s$ 에 대한 의존성은 증거 항에서 떨어질 수 있다:

### 수학식 7

$$P(e|s, M) = \sum_{j=1}^K P(e|c_j, M_{candidate}) P(c_j|s, M_{sequence})$$

[0111]

[0112] 후보 모델 2

[0113] 후보모델의 또 다른 변형은 먼저 베이즈 규칙을 이용해 증거 유사도를 변환한다:

### 수학식 8

$$P(e|s, M) = \frac{P(s|e, M) P(e|M)}{P(s|M)}$$

[0114]

[0115] 증거 조건 시퀀스 확률은 다음과 같이 재작성될 수 있다:

### 수학식 9

$$P(s|e, M) = \sum_{j=1}^K P(s|c_j, e, M_{sequence}) P(c_j|e, M_{candidate})$$

[0116]

[0117] 여기서,  $c_j$ 는 단일 후보이고, 주어진 증거소스에 대해  $M$ 의 2개 서브모델들, 즉, 후보모델  $M_{candidate}$  및 시퀀스 모델  $M_{sequence}$ 을 갖는다. 이 경우, 핵심 가정은 다음과 같다:

[0118] 가정 3: 해당 모델 하에서 유사도는 후보에 대한 주변화로 표현될 수 있으며, 타겟 시퀀스는 후보가 주어진다면 조건상 증거에 무관하다.

[0119] 이 가정을 적용하면,  $s$ 에 대한 의존성이 증거 항으로부터 제거될 수 있다:



## 수학식 10

$$P(s|e, M) = \sum_{j=1}^K P(s|c_j, M_{sequence}) P(c_j|e, M_{candidate})$$

[0120]

[0121] 후보 발생기(2)는 입력 시퀀스 발생기(1)에 의해 출력된 문맥 시퀀스(30) 및 입력 시퀀스(20)에 해당하는 문맥 및 입력의 2개의 소스들로부터의 증거를 이용한다.

[0122] 상술한 바와 같이, 문맥은 사용자가 이미 기입한 것에 대한 관찰된 증거를 나타내고, 문맥 시퀀스가 다수의 단어들을 포함한 경우 용어 경계들과 함께 일련의 문자 형태로 있다. 입력은 사용자가 현재 기입하고 있는 것에 대한 관찰된 증거를 나타내고, 확률 스트링의 형태이다.

[0123] 이들 2개의 증거 소스들과 함께 예시적인 수학식(5)은 다음을 나타낸다:

## 수학식 11

$$\frac{P(s|R)P(context|s, M_{context})P(input|s, M_{input})}{Z}$$

[0124]

[0125] 이 표현에서, 랭킹에 영향을 주지 않기 때문에 선택적인 Z는 정규화 인수를 나타내고, 이 정규화 인수의 평가는 본 명세서에 전체가 참조로 함체되어 있는 PCT/GB2011/001419에 더 상세히 설명되어 있다. 특히, 정규화 상수 (Z)는 다음으로 근사된다:

$$\sum_{j=1}^{|S|} P(s_j|R)P(context|s_j, M_{context})P(input|s_j, M_{input})$$

[0126]

[0127] 이 근사는 하기와 같이 시스템에 구현될 수 있다. 함수(z)는 시퀀스 세트(T)에 대해 고려되므로, 다음과 같다:

$$z(T) = \sum_{j=1}^{|T|} P(s_j|R)P(context|s_j, M_{context})P(input|s_j, M_{input})$$

[0128]

[0129] Z는 다음과 같이 계산된다:

$$Z = z(T) + z(\{u\}) * k$$

[0130]

[0131] 여기서 u는 "미지의" 시퀀스를 나타내고, k는  $|S| - |T|$ 의 평가이며,  $|S|$ 는 모든 가능한 타겟 시퀀스 세트에서 시퀀스 개수이고,  $|T|$ 는 기본 증거 모델들 중 적어도 하나가 "기지의" 평가를 갖는 시퀀스 개수이다. 각 개개의 증거 조건모델(M)은  $P(e|u, M)$ 의 평가, 즉, "미지의" 시퀀스가 주어진다면 증거 관찰들에 대한 분포를 반환한다. 본질적으로, 이는 각 개개의 증거 조건모델이 자신의 분포상 평활화를 책임지나, "미지의" 시퀀스의 전체 평가 개수에 비례하는 k와 관련되어야 하는 것을 의미한다. 실제로, 각 모델은  $S' \subset S$ 인 시퀀스( $S'$ ) 세트에 대해 "알" 것이고,  $P(e|s, M)$ 의 평가는 상수이고 모든  $s \notin S'$ 에 대한  $P(e|u, M)$ 와 같다. 이 성향의 평활화는 시스템이 각 증거 소스와 관련된 모델에서 변하는 신뢰 수준을 고려하게 한다.

[0132] 상술한 모델 프레임워크내에서, 개시된 시스템은 3개 모델을 실행한다:

[0133] 타겟 시퀀스 우선순위 모델(R)은

[0134] · 유니그램 모델(Unigram model)은 원자 단위로 각 시퀀스를 내부적으로 다루는 것을 고려함이 없이 언어의 시퀀스에 대한 분포를 구현한다. 가령, 유니그램 모델은 크기 1의 n-그램이다.

[0135] 입력증거 모델( $M_{input}$ )은 2가지 모델로 구성된다:

- [0136] · 후보모델:  $M_{input-candidate}$ 는 특정 후보해석이 주어지면 입력 관찰에 대한 조건 분포를 실행한다.
- [0137] · 시퀀스 모델:  $M_{input-sequence}$ 는 의도한 타겟 시퀀스가 주어지면 후보에 대한 조건 분포를 실행한다.
- [0138] 문맥증거 모델( $M_{context}$ )은 2개 모델로 구성된다:
- [0139] · 후보모델:  $M_{context-candidate}$ 는 특정 후보해석이 주어지면 문맥 관찰에 대한 조건 분포를 실행한다.
- [0140] · 시퀀스 모델:  $M_{context-sequence}$ 는 특정 문맥이 주어지면 언어 또는 언어 세트의 시퀀스에 대한 조건 분포를 실행한다.
- [0141] 후보 발생기(2)는 어떤 후보(40)가 가장 높은 랭킹 확률을 갖지 판단하기 위해 수학적식(11)의 분자에서 개개의 평가들을 계산하도록 구성되며, 이들 후보들 중 하나 이상이 텍스트 예측으로서 후보 디스플레이(3)에 나타나진다.
- [0142] 아래에서, 수학적식(11)에서 관련된 개개의 확률들 각각이 어떻게 계산되는지에 대한 논의가 제공된다. 따라서, 각각의 하기의 항들, 즉, 입력 유사도  $P(input | s, M_{input})$ ; 문맥 유사도  $P(context | s, M_{context})$  및 타겟 시퀀스 우선순위 평가  $P(s | R)$ 의 계산이 논의된다. Z 계산은 선택이므로 상세히 언급하지 않는다.

[0143] **타겟 시퀀스 우선순위**

[0144] 타겟 시퀀스 우선순위는 바람직하게는 다음과 같이 계산된다:

[0145] 
$$P(s|R) = \begin{cases} P(s|R_{unigram}) & \text{if } (s \in V) \\ P(s|R_{character}) & \text{otherwise} \end{cases}$$

[0146] 여기서, V는  $R_{unigram}$ 에 포함된 시퀀스 세트이고, 모델의 실행은 평탄화 주파수기반의 유니그램 언어 모델 및 평탄화된 마코브 체인 문자모델을 구성하기 위한 공지의 기술들에 따른다. 이들 모델을 실행하기 위한 많은 적용가능한 기술들이 아래에 나열되어 있다. 그러나, 나열되지 않은 다른 적절한 기술들도 또한 적용된다:

- [0147] · (해당분야에 공지된) 평탄화된 n-그램 항 또는 문자 모델
- [0148] · 참조문헌 영국특허출원 0917753.6에 기술된 바와 같은 적응식 다언어 모델
- [0149] · 가령 <참조문헌: Scheffler 2008>에 기술된 바와 같은 PPM(부분적 매칭에 의한 예측) 언어 모델
- [0150] · 구성하는 어휘 구성요소들로부터 확률적으로 시퀀스를 발생하도록 구성된 형태분석엔진.

[0151] 타겟 시퀀스 우선순위 모델(R)을 포함함으로써, 시스템은 의도된 시퀀스 예측의 향상된 정확도를 제공한다. 더욱이, 타겟 시퀀스 우선순위 모델(R)은 눈에 보이지 않는 타겟 시퀀스들의 문자기반 추론을 가능하게 한다. 즉, 시스템은 모든 가능한 타겟 시퀀스들에 대해 근사해 미지의 타겟 시퀀스들을 더 잘 추론할 수 있다.

[0152] **입력 유사도  $P(input | s, M_{input})$**

[0153] 입력 유사도  $P(input | s, M_{input})$ 는 후보모델(1) 및 특히 수학적식(7)에 의해 평가된다:

**수학적식 12**

[0154] 
$$P(input|s, M_{input}) = \sum_{j=1}^K P(input|c_j, M_{input-candidate}) P(c_j|s, M_{input-sequence})$$

[0155] 수학적식(7)의 증거소스(e)가 입력이면, 입력 시퀀스(20) 및 관련된 모델은  $M_{input-candidate}$  및  $M_{input-sequence}$ 을 포함한 입력모델( $M_{input}$ )이다.

[0156] 따라서, 입력 유사도  $P(input | s, M_{input})$ 를 결정하기 위해, 입력후보 평가  $P(input | c_j, M_{input-candidate})$  및 입력 시퀀스 평가  $P(c_j | s, M_{input-sequence})$ 를 판단할 필요가 있으며, 각각이 하기에 상세히 언급되어 있다.



[0157] 입력 후보 평가  $P(\text{input} \mid c_j, \mathbf{M}_{\text{input-candidate}})$

[0158] 본 명세서에서 전체가 참조로 함체되어 있는 PCT/GB2011/001419는 각 노드가 하나 이상의 문자들로 구성된 서브시퀀스를 포함하는 DAG(directed acyclic graph)의 변형인 PCSG(probabilistic constrained sequence graph)를 기술한다. 하나 이상의 문자들로 구성된 서브시퀀스는 문자 세트 또는 입력 시퀀스(20)의 2 이상의 문자 세트들에 대한 시퀀스로부터 개개의 문자에 해당한다. 입력 시퀀스(20)의 문자 세트(들)에 있는 문자들에 대한 확률분포(들)를 기초로 노드들에 대한 예지들 각각에 확률이 할당된다. PCSG는 각 경로가 또한 동일한 길이로 제한되는 속성을 갖는다.

[0159] 공식적으로, PCSG는 노드 세트(N), 루트 노드(r), 방향성 에지 세트(E), 및 파라미터(확률) 세트( $\Theta$ )를 포함해 4-tuple로 구성된다:

[0160]  $G = (N, r, E, \Theta)$  (17)

[0161] 2개의 노드들(n 및 n') 간의 에지는  $(n \rightarrow n')$ 로 표시되고, 에지를 따라 n에서 n'까지 이동하는 확률은  $P(n \mid n')$ 로 표시된다. G를 통과하는 경로는 노드(r)에서 시작해 아웃고잉 에지를 전혀 포함하지 않는 노드에 이를 때까지 각 방문된 노드로부터 하나의 아웃고잉 에지를 정확히 따른다. 다음 속성들이 G에 적용된다:

[0162] 1) G는 방향성 비순환 그래프(DAG)이다

[0163] 2)  $\forall n \in N. \exists m. (m \rightarrow n) \in E \Rightarrow n = r$ , 즉, 루트를 제외한 모든 노드들은 적어도 하나의 인커밍 에지를 가져야 한다

[0164] 3)  $\exists m, k \in N. \forall n \in N. (m \rightarrow n) \in E \Rightarrow (n \rightarrow k) \in E$ , 즉, 주어진 노드로부터 분기한 모든 경로들은 연속한 공통 노드에서 바로 재결합된다. 이 속성은 그래프의 구조를 심히 구속하고 모든 경로들이 같은 길이를 갖는 것을 의미하며, 경로확률 계산에 정규화 요건을 완화시킨다.

[0165] 입력후보모델 함수는 하기와 같이 (입력후보 평가와 일치하는) 주어진 경로에 대한 확률을 계산한다:

[0166]  $P(\text{input} \mid c_j, \text{Minput} - \text{candidate}) = f_{\text{input-candidate}}(c_j) = P(p_j \mid G)$

[0167] 여기서  $P(p_j \mid G)$ 는 경로에서 에지들 각각의 곱으로 계산된 경로 유사도이다:

[0168]  $P(p_j \mid G) = P(n_1 \mid r) \prod_{k=2}^K P(n_k \mid n_{k-1})$

[0169] 여기서 K는 경로에서 에지 개수이다. 이 바람직한 공식은 노드들 간에 내재적 독립 가정과 같음이 주목될 수 있다. 이 내재적 독립은 예시적인 PCSG의 설명 후에 하기에 기술된다.

[0170] 따라서, 다음 속성은 에지 상에 확률에 대해 남아 있다:

[0171]  $\forall n \in N. \sum_{(n \rightarrow m) \in E} P(m \mid n) = 1$

[0172] 다시 말하면, 주어진 노드(n)로부터 모든 아웃고잉 에지들에 대한 확률은 합해서 1이 되어야 한다. 이는 또한 다음의  $\sum_i P(p_i \mid G)=1$ 이 보유되는 것을 의미한다. 즉, PCSG에서 모든 경로들의 확률에 대한 합은 1과 같다.

[0173] 개시된 시스템에서, 입력 후보 세트는 확장형 PCSG(EPCSG)로 표현된다. 개개의 입력 후보는 EPCSG를 지나는 고유경로로 표현된다. 각 입력 후보에 대해, 대표적인 경로의 정규화 확률이 계산될 수 있다.

[0174] PCSG는 (입력 후보 세트로 맵핑될 수 있는) 입력 시퀀스(20)로부터 먼저 발생된다. 그런 후 PCSG는 더 표현적이거나 여전히 구속된 후보들을 가능하게 하는 추가 구조들로 증강된다.

[0175] 아래에 입력 시퀀스가 있는 시나리오에 대한 PCSG의 예가 제공된다:

[0176]

[0177]

$$\left[ \begin{array}{l} \{(H \rightarrow 0.5), (h \rightarrow 0.3), (g \rightarrow 0.1), (j \rightarrow 0.1)\} \\ \{(e \rightarrow 0.8), (w \rightarrow 0.1), (r \rightarrow 0.1)\} \end{array} \right]$$

이는 입력 시퀀스 발생기(1)는 사용자가 기입하려고 했던, 가령, 문자 'H' 다음에 문자 'e'를 평가해, 관찰된 입력 이벤트들이 0.5 및 0.8의 각각의 확률을 갖도록 예측될 수 있는 시나리오를 인코딩한다. 그러나, 입력 시퀀스 발생기는 또한 각 주어진 세트의 문자를 가로지르는 확률분포로 2개 문자 세트를 포함한 확률 스트링을 발생함으로써 사용자가 "H" 및 "e"와는 다른 문자를 기입하려고 했을 수 있음을 고려한다.

[0178]

이들 확률분포가 발생하는 방법이 본 개시의 주제는 아니다. 오히려 가령

[0179]

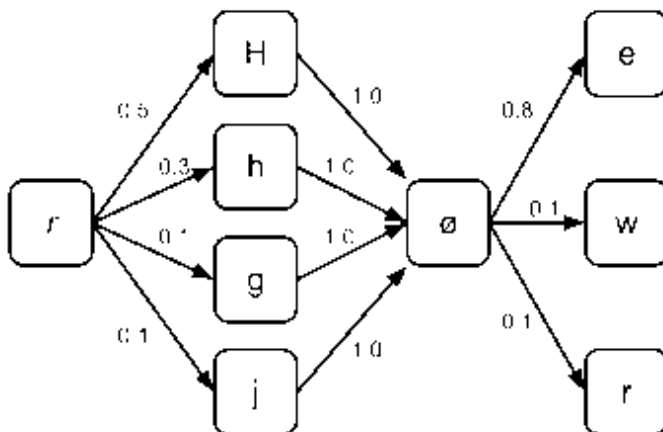
- 분포는 사용자가 "H" 키에 해당하는 영역을 두드릴 경우, 문자 "G" 및 "J"가 어떤 확률로 입력 시퀀스에 포함될 수도 있는 특정 키보드 레이아웃, 가령 QWERTY 키보드에서 주어진 타겟 키를 둘러싼 문자들을 기초로 발생될 수 있다.

[0180]

- 분포는 (터치 스크린 가상 키보드 상의) 터치 좌표 및 지정된 키들의 좌표 사이 거리(또는 제곱 등과 같은 어떤 거리의 함수)를 기초로 발생될 수 있는 것을 포함한 기술 범위가 적용될 수 있는 것이 강조된다.

[0181]

입력 시퀀스에서 각각의 문자 세트들에서 각 문자에 PCSG의 자신의 노드를 할당하는 후보 발생기(2)에 의해 입력 시퀀스(20)로부터 발생된 비증강 PCSG는 다음과 같이 보일 수 있다:



[0182]

[0183]

문자의 각 세트는, 다음 동기 노드 세트 전, 바로 다음의 널 노드에 재결합되는 그래프에서의 동기 노드 세트로서 추가된다. 각 동기 노드에 대한 인커밍 에지 확률은 세트내 문자의 확률값에 해당한다.

[0184]

상술한 바와 같이, 경로 유사도에 대한 바람직한 공식,  $P(p_j | G)$ 은 노드들 간에 내포된 독립 가정과 같다. 본 발명의 경우, 이는 다른 문자 세트들에서 문자들에 대한 확률분포와 별개인 문자 세트에 대한 확률분포를 갖는 각 문자 세트에 의해 생긴다. 따라서, 문자 세트로부터 발생된 동기 노드들의 각 세트는 동기 노드들의 다른 세트들과는 구분되고 별개인 확률분포를 갖는다. PCSG의 구성으로부터, 각 동기노드 세트로부터 단 하나의 이런 노드만이 주어진 경로에 사용될 수 있다.

[0185]

본 명세에 전체가 참조로 합체된 PCT/GB2011/001419에 기술된 바와 같이, PCSG는:

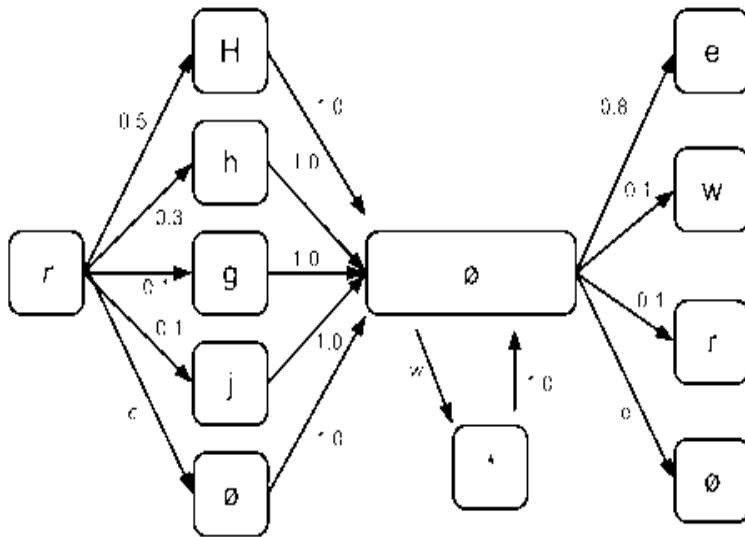
[0186]

· 잘못된 문자들로 입력 시퀀스를 고려한 공노드 서브-경로

[0187]

· 사용자가 타겟 시퀀스로부터 하나 이상의 문자들을 생략한 예들을 고려한 '와일드카드(wildcard)' 구조를 포함한 EPCSG를 생성하기 위해 2개의 추가 구조들로 증강될 수 있다.

[0188] 이런 식으로 (EPCSG를 제공하기 위해) 증강된 동일한 PCSG가 아래에 도시되어 있다:



[0189]

[0190] 기정의된 확률 페널티(e)로, 시스템은 사용자로부터의 잘못된 입력을 다루기 위해 입력 시퀀스에 해당하는 문자 대신 널(null) 문자가 삽입되게 한다. 이를 위해, 후보 발생기(2)는 상술한 동기 노드들의 각 세트에 추가 노드로서 ("Ø"로 표현된) 널 문자를 삽입하도록 구성된다. 따라서, 하나 이상의 널 문자들을 포함한 경로들에 잇따른 하나 이상의 문자들을 스킵하는 EPCSG를 지나는 경로들이 있다.

[0191]

또한, 사전 정의된 페널티(w)로, 시스템은 입력시 전체적으로 생략된 타겟 시퀀스에 있는 문자를 고려한다. (\*)로 표시된 이 노드는 문자 세트(가령 '와일드카드')에서 모든 가능한 문자를 포함한 가지에 대한 속기이다. 상기 도시된 EPCSG에 따르면, 후보 발생기(2)는 각각의 동기 문자 노드 세트 간에 와일드카드 구조(가령, 와일드카드를 나타내는 노드)를 삽입하도록 구성된다. 이 구조는 입력만을 기초로 예측되지 못한 후보에 생략된 문자가 고려되게 한다.

[0192]

본 시스템의 후보 발생기(2)는 용어 경계들 형태의 다른 구조들을 EPCSG에 삽입한다. 용어 경계는 공간과 같은 문자를 고려할 필요가 없다. 몇몇 언어에서, 단어들 간에 공백이 없으나, 시스템은 한 단어가 끝나고 다른 단어가 시작되는 곳을 예측한다. 즉, 용어 경계를 식별한다. 이를 위해, 후보 발생기(2)는 용어 경계, 가령 공백이라기보다는 문자를 입력하는 사용자에게 대해 수정하도록 상술한 각각의 동기 노드들 세트에 추가 노드로서 ('\_'로 라벨붙은) 용어 경계를 삽입하도록 구성된다. 더욱이, 후보 발생기는 동기 문자 노드들의 각 세트 간에 다른 용어 경계 노드를 삽입한다. 용어 경계가 와일드카드 구조의 임의 사용들 간에 고려되어야 하므로 동기 와일드카드 구조는 2개의 동일한 구조들로 분할되어야 하고 용어 경계 구조가 그 내에 삽입되어야 한다. 용어 경계 구조는 간단히 용어 경계 노드 및 용어 경계를 선택적으로 하는 널 문자 노드로 구성된다.



## 수학식 14

$$P(t_k | t_{k-1}, t_{k-2}) * P(t_{k-1} | t_{k-2}, t_{k-3}) * \dots * P(t_4 | t_3, t_2) * P(t_3 | t_2, t_1) * P(t_2, t_1) * P(t_1)$$

[0205]

[0206]

예컨대, 문맥 시퀀스(30)가 "have you ever"이고 입력 시퀀스(20)(즉, EPCSG를 지나는 경로)에 대한 후보가 "been to London"이면, 후보 발생기(2)는 3개 트라이그램들: "you ever been", "ever been to" 및 "been to London"을 발생할 것이다. "have you ever" 트라이그램은 문맥 시퀀스(30)로부터 전체적으로 구성된다. 입력되어진 임의의 이런 트라이그램은 1의 확률을 갖고, 따라서 결과를 바꿀 수 없기 때문에 고려되지 않는다. 이 예에 대한 문맥확률평가는 3개 트라이그램들 각각의 확률들을 함께 곱함으로써 발생된다.

[0207]

이는 이 예의 k 길이 시퀀스가 {you,ever,been,to,London}를 포함한 길이 5가 되는 상기 수학식(14)을 반영한다.

[0208]

문맥 시퀀스가 없으나 동일한 입력 시퀀스가 있는 경우, 후보 발생기(2)는 단일 트라이그램 "been to London"의 문맥 확률 평가에 대한 후보를 가지치기할 것이다. 더 긴 입력, 가령 "been to London to see"으로, 후보는 3개 트라이그램 "been to London", "to London to" 및 "London to see"의 문맥확률 평가에 대해 가지쳐 질 수 있다.

[0209]

임의의 적절한 문맥언어모델이 후보용 확률평가를 제공하는데 사용될 수 있다.: 본 시스템은 예시용으로 사용되는데 트라이그램 모델의 사용에 국한되지 않는다.

[0210]

따라서, 본 발명은 입력 시퀀스의 전체 내용 및 문맥 시퀀스(30)가 있는 문맥 시퀀스의 용어들을 기초로 하는 가지치기를 제공한다.

[0211]

이런 식으로 훈련된 문맥 데이터의 사용으로 의도하거나 하지 않았던 간에 폐기될 모델에 의해 이전에 보이지 않았던 구들이 초래될 수 있다.

[0212]

EPCSG의 경로를 가지치기하기 위해, 후보 발생기는 2개의 확률의 곱을 판단함으로써 주어진 후보에 대한 문맥 확률평가를 누적된 PCSG 경로확률과 조합하도록 구성된다. 후보 발생기(2)는 임계치(t)를 통해 현재 가장 가능성 있는 시퀀스와 덜 가능한 시퀀스 간의 비, 즉, 가장 가능성 있는 경로와 가지치기할 것으로 고려되는 경로 간의 확률비에 대한 하한을 설정한다. 따라서, 경로( $n_1, \dots, n_L$ )는 다음이 확보된다면 가지쳐진다:

## 수학식 15

$$\frac{P(t_k^n, t_{k-1}^n, t_{k-2}^n, \dots, t_3^n, t_2^n, t_1^n) P(n_1 | r) \prod_{j=2}^L P(n_j | n_{j-1})}{\operatorname{argmax}_m [P(t_k^m, t_{k-1}^m, t_{k-2}^m, \dots, t_3^m, t_2^m, t_1^m) P(m_1 | r) \prod_{j=2}^L P(m_j | m_{j-1})]} < t$$

[0213]

[0214]

그러나, 가지치기 기술은 수학식 14 및 15에 대해 상술한 것과는 달리 사용될 수 있다. '확률평가'라는 말은 진정 수학적 확률 또는 랭킹 또는 스코어링과 같은 확률을 평가하는 임의의 다른 형태를 의미할 수 있다. 후보에 할당된 확률평가 또는 '스코어'는 수학식 14에서 항들을 가중화함으로써 후보에서 초기 항들에 대해 바이어스될 수 있어, EPCSG의 큰 폭이 가능한 초기에 폐기될 수 있게 하므로, 상기 방법은 모든 경로에 대한 전체 EPCSG를 가로질러 이동할 필요가 없기 때문에, 계산효율을 증가시킨다. 더욱이, 검색된 용어들의 시퀀스는 후보를 구성하는 용어들의 전체 시퀀스보다 미만일 수 있다. 가령, 후보의 처음 한두 항들을 포함할 수 있다. 여하튼, 후보에 할당된 스코어/확률평가는 여전히 문맥모델(Mcontext)에서 후보의 일련의 항들을 검색함으로써 결정될 것이다. 게다가, 수학식 15에 대한 대안으로 임계치 한계를 설정하는 다른 방식들도 사용될 수 있다.

[0215]

바람직한 실시예에서, 문맥 유사도, 및 특히 문맥 시퀀스 평가를 결정하는데 사용된 동일한 트라이그램들을 평가함으로써 이력 후보들을 가지칠 수 있기 때문에, 후보 발생기(2)는 수학식(14)의 확률평가를 계산하도록 구성된다. 이는 가지쳐지지 않은 후보들에 대한 트라이그램 결과들이 아래의 수학식(17)에 대해 언급된 바와 같이 후보 발생기(2)에 의해 캐시 및 재사용될 수 (및 가지쳐진 후보들에 대해, 트라이그램 결과들이 폐기될 수) 있

기 때문에 필연적으로 계산회수를 줄인다.

[0216] 수학식 12 및 16에서 평가하기 위한 몇몇 후보들이 있기 때문에, 경로들을 가지치는 것은 또한 (수학식 11을 통해) 후보 발생기(2)에 의한 예측 계산을 줄인다.

[0217] 경로들은 EPCSG에서 가능할 수 있으나 문맥확률평가를 기초로 후보 발생기(2)에 의해 가지쳐질 수 있다. 그러나, 이 가지치기가 여기서 행해지지 않았다면, 동일한 후보들은 (이들이 가지 쳐졌을 정확히 동일한 이유로) 수학식(11)에서의 문맥모델로부터 극히 낮은 스코어를 받도록 진행될 것이고, 이는 극히 낮은 확률을 줄 것이며, 대신 나중 단계에서 이들은 폐기되어 진다.

[0218] **입력 시퀀스 평가  $P(c_j, | s, M_{input-sequence})$**

[0219] 입력 시퀀스 평가  $P(c_j, | s, M_{input-sequence})$ 는 타겟 시퀀스가 주어지면 후보 시퀀스에 대해 분포되고 정규화 표시자 함수로서 평가될 수 있다:

### 수학식 16

$$P(c_j | s, M_{input-sequence}) = \frac{\delta(s, c_j)}{Z}$$

[0220]

[0221] 여기서  $t'$ 이  $t$ 의 접두사이면  $\delta(t, t')=1$ 이고, 그렇지 않으면 0이며,  $Z=\sum_k \delta(s, c_k)$ 이다. 즉, 모든 후보들의 합이다.

[0222] 후보들의 고유성이 가정되고 후보 세트가 모든 가능한 시퀀스들을 포함하도록 허용되면, 길이(n)의 타겟 시퀀스에 대해, 정확히 n개의 매칭 접두사들, 및 이에 따라 후보들이 있을 것이기 때문에 정규화 인수  $Z=length(s)$ 는 고쳐 쓸 수 있다.

[0223] **문맥 유사도  $P(context | s, M_{context})$**

[0224] 문맥 유사도,  $P(context | s, M_{context})$ 는 후보모델(2) 및 특히 증거소스(e)가 문맥인 수학식(8 및 10)에 의해 평가된다:

### 수학식 17

$$P(context | s, M_{context}) = \frac{\sum_{j=1}^K P(s | c_j, M_{context-sequence}) P(c_j | context, M_{context-candidate}) P(context | M_{context})}{P(s | M_{context})}$$

[0225]

[0226] 수학식 17에서 j에 대한 합은 문맥의 어휘 및 정자법 변화들에 대한 합에 대한 것이다. 입력 후보들이 가지치기 단계에서 폐기되었다면, 이 용어 시퀀스들에 대한 이런 변화들은 평가될 필요가 없다.

[0227] 문맥 유사도를 계산하기 위해 수학식 17에 도시된 바와 같이, 후보 발생기(2)는 각각이 하기에 더 상세히 기술된 다음 평가들을 계산하기 위해 구성된다: 문맥 시퀀스 평가  $P(s | c_j, M_{context-sequence})$ ; 문맥 후보 평가  $P(c_j | context, M_{context-candidate})$ ; 문맥 우선순위 평가  $P(context | M_{context})$ ; 및 타겟 시퀀스 우선순위 평가  $P(s | M_{context})$ .

[0228] **문맥 시퀀스 평가  $P(s | c_j, M_{context-sequence})$**

[0229] 문맥 시퀀스 평가  $P(s | c_j, M_{context-sequence})$ 는 문맥 시퀀스 모델 하에서 특정 후보 시퀀스( $c_j$ )가 주어진 경우 타겟 시퀀스(s)의 확률이다. 형식적으로 문맥 시퀀스 모델은 문맥 시퀀스가 주어진 경우 타겟 시퀀스의 확률, 즉  $f_s(t_{target}, t_{context})=P(t_{target} | t_{context}, \theta_s)$ 을 되돌려 보내며,  $\theta_s$ 는 모델의 파라미터이다. 따라서, 문맥 시퀀스 확률은 다음과 같이 계산된다:  $P(s | c_i, S)=f_s(s, c_i)$ .

[0230] 다음과 같이 별개로 또는 조합으로 사용되는 적용가능한 많은 기술들이 있다:

- [0231] · (해당기술분야에 공지된) n그램 언어 모델
- [0232] · 본 명세서에 전체가 참조로 합체된 PCT/GB2010/001898에 기술된 바와 같이 적응식 다언어 모델
- [0233] · 해당기술분야에 공지된 PPM(부분 매칭에 의한 예측) 언어 모델
- [0234] · 해당기술분야에 공지된 바와 같이 발생적 HMM(히든 마코브 모델) 확률적 언어 일부의 태거
- [0235] · 적절히 구성된 신경 네트워크

[0236] **문맥 후보 평가  $P(c_j | \text{context}, M_{\text{context-candidate}})$**

[0237] 문맥 후보 평가  $P(c_j | \text{context}, M_{\text{context-candidate}})$ 는  $f_{\text{context-candidate}}(t, \text{context}) = P(t | \text{context}, \Theta_{\text{context-candidate}})$  형태의 함수이고, 여기서  $t$ 는 임의 시퀀스이고,  $\Theta_{\text{context-candidate}}$ 는 모델의 파라미터이다. 따라서, 문맥 후보의 조건 평가는 다음과 같이 계산된다:

[0238]  $P(c_j | \text{context}, M_{\text{context-candidate}}) = f_{\text{context-candidate}}(c_j, \text{context})$ .

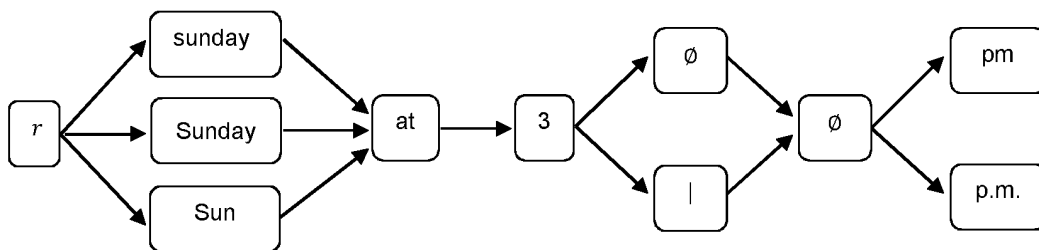
[0239] 문맥 시퀀스에서 이런 변화를 고려할 경우, 입력 후보 평가의 방식과 유사한 방식으로 EPCSG를 이용한 기술이 사용되며, 그래프는 문맥 용어들의 정자법 변화(가령, 대분자 쓰기, 축약, 두문자)에 대해 분기한다.

[0240] 다음 12개 문맥 후보 시퀀스들에 대한 예가 제공된다:

“Sunday at 3pm”	“sunday at 3pm”	“Sun at 3pm”
“Sunday at 3 pm”	“sunday at 3 pm”	“Sun at 3 pm”
“Sunday at 3p.m.”	“sunday at 3p.m.”	“Sun at 3p.m.”
“Sunday at 3 p.m.”	“sunday at 3 p.m.”	“Sun at 3 p.m.”

[0241]

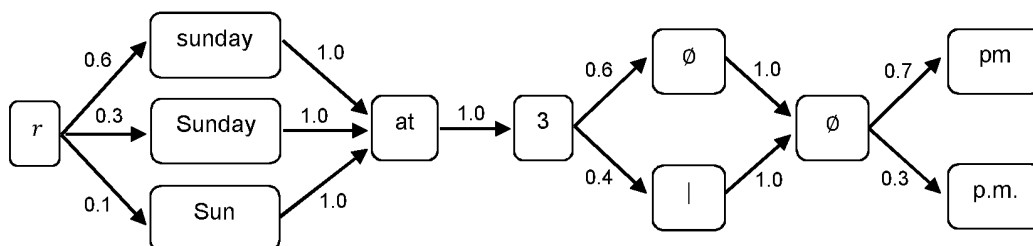
[0242] 이들은 다음 PCSG(' | '로 표시된 명백한 단어 경계 및 공문자 'Ø'에 의한 빈 시퀀스)에 의해 표현될 것이다:



[0243]

[0244] 가령 수식(19)에 따른 문맥후보모델에 따라 확률이 에지에 할당된다:

### 수식 19



[0245]



[0246] 12개 시퀀스에 대한 후보확률들이 하기와 같이 (간략히 하기 위해 3개 예만 나타냄) PCSG로부터 발생된다:

$$P(\text{"sunday at 3pm"}|\text{"sunday at 3pm"}, C) = 0.6 * 1.0 * 1.0 * 0.6 * 1.0 * 0.7 = 0.252$$

$$P(\text{"Sunday at 3 pm"}|\text{"sunday at 3pm"}, C) = 0.3 * 1.0 * 1.0 * 0.4 * 1.0 * 0.7 = 0.084$$

$$P(\text{"Sun at 3 p.m."}|\text{"sunday at 3pm"}, C) = 0.1 * 1.0 * 1.0 * 0.4 * 1.0 * 0.3 = 0.012$$

[0247]

[0248] PCSG를 구성하고 확률을 노드에 할당하는데 사용된 모델의 사양들은 시스템의 특정 예에 따라 변할 것이다. 상기 구성은 3개의 일반 변화들의 예를 인코딩한다:

[0249]

단어 경계에서 (잠정적으로 명백한) 분기

[0250]

격변화에서 분기

[0251]

어휘 변화에서 분기

[0252]

임의 타입의 변화가 이 프레임워크에서 인코딩될 수 있음이 이해될 것이다. 또 다른 예는 이전 제안들에 대해 분기될 수 있다. 가령, 시스템이 "on" 및 "in" 모두를 예측했고 사용자가 "in"을 선택했다면, 이는 "in"에 할당된 확률 가중치를 가지나 사용자가 우발적으로 잘못된 제안을 받아들인 확률을 나타내기 위해 "on"에 할당된 작은 확률을 갖는 분기로 인코딩될 수 있다. 상기의 경우, 다음 원리들이 인코딩된다:

[0253]

소문자 첫 's'를 갖는 'sunday'는 완전한 변형 'Sunday'보다 그 자체로 확률이 덜한 축약형 'Sun'보다 가능성이 덜하다;

[0254]

"pm"이 숫자 "3"과 떨어져 있는 토큰경계인식(tokenisation)은 아닌 경우보다 약간 가능성이 덜하다;

[0255]

마침표 변화 "p.m."는 마침표가 없는 형태 "pm"보다 약간 가능성이 덜하다.

[0256]

문맥 후보 PCSG의 특정 예는 바람직하게는 다음 방식으로 초기 시퀀스(s)로부터 알고리즘적으로 구성된다.

[0257]

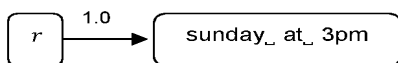
루트에 연결된 단일노드  $n^s$  내에 캡슐화함으로써 s를 PCSG로 바꾼다:

[0258]

변화점에서 분기 노드를 도입함으로써 반복적으로  $n^s$ 를 해체한다.

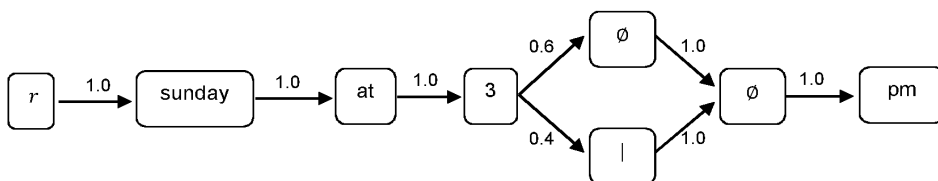
[0259]

예컨대, 본래 시퀀스 "sunday at 3pm"에 작용하는 PCSG 구성 알고리즘이 고려된다. 먼저 단계 1:



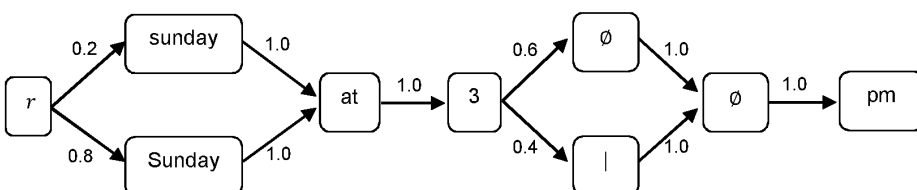
[0260]

[0261] 시스템은 확률 토크나이저(Tokeniser)를 전개해 다음과 같이 된다:



[0262]

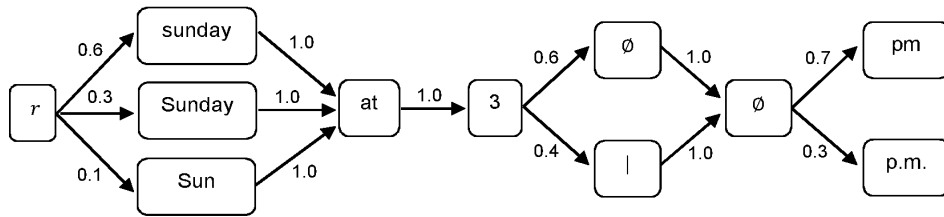
[0263] 상기 PCSG 속성 3으로 인한 변형은 전체 경로 확률들에 영향을 주지 않기 때문에 연이은 처리에 편리한 한 노드의 분기인 특별한 경우에 따라 항상 분기 및 재결합 구조적 삽입의 형태를 취하는 것에 유의하라. 예지 확률들이 하기에 더 상세히 설명된 모델에 따라 추가된다. 알고리즘을 계속하면, 격변화 분석기가 전개된다:



[0264]



[0265] 그리고 마지막으로 어휘변형분석이 전개된다:



[0266]

[0267] PCSG 속성 3으로 인해, 분기들은 재분기 전으로 수렴되어야 함에 유의하라. 이는 2개 분기점들이 접촉해 발생하면 몇몇 경우 공 노드가 삽입되어야 하는 것을 의미한다.

[0268] 예지 확률들이 바람직하게 PCSG에 할당된다. 예지 확률들의 할당이 바람직하게는 문맥 후보 모델의 파라미터들에 대해 수행된다. 이들 확률들의 직관적 해석은 두 부분이 있다:

[0269] (1) 이들은 사용자가 특정 가지에 할당한 시퀀스를 의도한 유사도의 평가를 나타낸다. 예컨대, 사용자가 "Dear ben"을 기입하면, 몇가지 확률로 이들이 실제로는 "Dear Ben"을 기입하도록 의도했음을 인정하고자 할 수도 있다.

[0270] (2) 이들은 특정 분기가 관찰된 시퀀스의 유효한 정자법 변형인 "백오프" 가능성을 나타낸다. 예컨대, 사용자가 "See you on Thur"를 기입하면, "Thur"의 다른 정자법 형태는 "Thurs"일 수 있다.

[0271] 특정 예지에 할당된 확률은 또한 어떤 배경모델 정보가 주어지면 정자법 변화의 평가된 유사도에 의해 영향받을 수 있다. 예컨대, 문맥 시퀀스 모델(S)은 실제로 분기 확률들을 산출하기 위해 다른 확률측정들과 조합해 사용될 수 있는 다른 정자법 변화들에 대한 유사도 평가를 얻기 위해 재사용될 수 있다. 이런 식으로 문맥 시퀀스 모델을 이용하는 것은 문맥 후보 모델(C)이 실제로 문맥 시퀀스 모델(S)의 예를 포함하는 것을 의미하며, 이는 후보 및 시퀀스 모델 간에 독립 가정에 대한 명백한 위반이다(상기 속성 7); 그러나, 이 가정은 실제로는 결코 문맥 경우에 요구되지 않으므로, 상대적으로 안전하다.

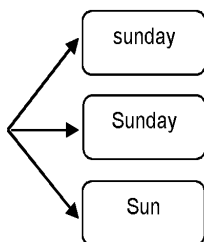
[0272] 일예가 명확히 하는데 도움될 것이다. 바람직한 실시예에서, 문맥후보모델은 다음 알고리즘을 이용해 확률을 할당하는 것으로 추정된다:

[0273] 관찰된 시퀀스는 0.8의 확률을 받는다; 다른 것들은 균일하게 밸런스를 받는다.

[0274] 문맥 시퀀스 모델 평가에 의해 값들이 스케일된다.

[0275] 상기 PCSG 속성(19)에 맞게 값들이 정규화된다.

[0276] 상기 PCSG 예로부터, 다음 분기가 고려될 수 있다:



[0277]

[0278] "study"는 본래 관찰이므로, 상기 알고리즘 중 1단계에 의해 0.8의 확률이 할당될 초기에 것이며, 다른 예지들은 0.1이 각각 할당될 것이다. 상기 예로, 문맥 시퀀스 모델에 의해 반환된 평가들은 다음과 같다:

$$P(\text{"sunday"}|C^S) = 0.01$$

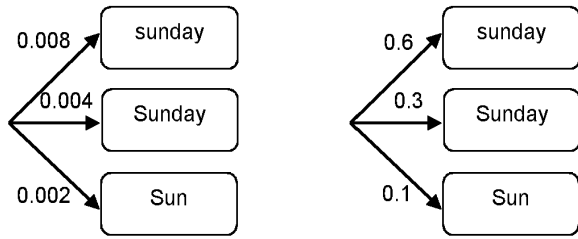
$$P(\text{"Sunday"}|C^S) = 0.04$$

$$P(\text{"Sun"}|C^S) = 0.02$$

[0279]

[0280] 여기서  $C^S$ 는 문맥후보모델이 이 경우 문맥 시퀀스 모델을 이용하는 것을 나타낸다. 따라서, 이 예에서 각 예지에

할당된 비정규화 및 정규화(반올림) 확률은 (각각) 다음과 같다:



**문맥 우선순위 평가  $P(\text{context} | M_{\text{context}})$**

문맥 우선순위 평가  $P(\text{context} | M_{\text{context}})$ 는 문맥과 관련된 본래 시퀀스(t)의 주파수를 정규화함으로써 근사될 수 있다.

**수학식 21**

$$P(\text{context} | M_{\text{context}}) \cong \frac{\text{freq}(t)}{\sum_{t'} \text{freq}(t')}$$

freq(t)는 트레이닝 데이터에서 시퀀스(t)의 주파수이고, 분모는 트레이닝 데이터에서 모든 시퀀스들의 주파수 합이다. 시퀀스 "t"는 문맥 시퀀스(30)이다. 문맥 우선순위는 예측이 도출된 해당 모델  $M_{\text{context}}$ 이 소정의 문맥 시퀀스(30)를 포함하는 확률에 따라 문맥 후보들의 확률값들을 가중화한다. 이렇게 하기 위해, 문맥 우선순위는 수식의 평가에 따라 예측 값들을 가중화한다.

실제로, 평가는 가령 보이지 않는 시퀀스들의 발생 주파수를 평가함으로써 평탄해질 것이다.

**타겟 시퀀스 우선순위 평가  $P(s | M_{\text{context}})$**

유사하게, 타겟 시퀀스 우선순위 평가  $P(s | M_{\text{context}})$ , 즉, 수학식(17)의 분모는 트레이닝 데이터에 대한 평탄화 주파수 분석을 이용해 평가될 수 있다. 가령, 타겟 시퀀스 우선순위는 문맥 트레이닝 데이터에서 모든 시퀀스들에 대한 타겟 시퀀스의 주파수를 정규화함으로써 근사될 수 있다

$$P(s | M_{\text{context}}) \cong \frac{\text{freq}(s)}{\sum_{s'} \text{freq}(s')}$$

여기서 f(s)는 트레이닝 데이터에서 타겟 시퀀스의 주파수이고 분모는 트레이닝 데이터에서 모든 타겟 시퀀스들의 모든 주파수들의 합이다. 분모는 트레이닝 데이터에 있는 항목의 전체 개수(중복 카운트)와 똑같이 근사적으로 결정될 수 있다.

**후보 디스플레이(3)**

후보 발생기(2)는 텍스트 예측으로서 가장 높은 랭킹 후보 또는 하나 이상의 텍스트 예측들로서 가장 높은 랭킹 후보들 중 하나 이상을 사용자에게 표현을 위한 후보 디스플레이(3)에 출력하도록 구성된다. 가장 높은 랭킹 후보들은 Z의 계산이 바람직하지만 선택적인 수학식(11)에 의해 결정된 가장 높은 확률을 갖는 후보들이다:

$$\frac{P(s|R)P(\text{context}|s, M_{\text{context}})P(\text{input}|s, M_{\text{input}})}{Z}$$

일실시예에서, 후보(40)는 사용자에게 의한 선택을 위해 가상 키보드를 따라 후보 디스플레이(3)에 나타난다. 후

보가 선택되면 후보가 발생된 입력 시퀀스(20)에 해당하는 입력 텍스트의 일부를 대체한다.

[0295] 다른 실시예에서, 사용자는 후보(40)를 선택하지 않고도 계속 타이프를 친다. 사용자에게 다용어 후보를 나타낸 후에, 계속 입력을 제공하며, 이는 가장 높은 랭킹 후보의 제 1 용어를 변경하지 않고, 예측으로부터 제 1 용어가 추출될 수 있고, 작성 텍스트에 "적어질 수 있다"(따라서, 문맥의 일부가 될 수 있다). 일예가 도 2a 및 도 2b에 도시되어 있다.

[0296] 도 2a에 도시된 바와 같이, 후보 발생기(2)는 후보 디스플레이(3)에 2개의 가장 높은 랭킹 후보들(40)을 출력하도록 구성될 수 있으며, 이 경우 "please call me" 및 "please call Melanie"이다.

[0297] 대안으로, 도 2b에 도시된 바와 같이, 후보 발생기(2)는 축어적 입력과 함께 가장 높은 랭킹 후보들(40)을 후보 디스플레이(3)에 출력하도록 구성될 수 있으며, 이 경우 "call me back" 및 "callmebac"이다.

[0298] 도 2a 및 도 2b의 예시적인 사용자 인터페이스들은 종래 "스페이스바" 키를 포함하지 않는 것에 유의하라. 후보 디스플레이와 선택 인터페이스와 조합해 개시된 시스템은 스페이스바 키를 필요로 하지 않는 텍스트 기입 인터페이스를 고려한다. 스페이스바 키가 없는 키보드가 이점적인데 이는 유효한 스크린 공간을 절감해 다른 키들이 더 넓은 공간을 차지하게 하거나 스크린이 크기가 줄어들게 하기 때문이다.

#### [0299] 시스템의 예시적인 사용

[0300] 사용자가 "When you get this, send me a reply" 메시지를 타이핑하는 상황을 생각하자. 몇몇 단계에서, 입력 디스플레이 필드가 나타난 상태에 있다고 또한 생각하자:

When you get this, send me are|

[0301] 이 경우, 입력 시퀀스 발생기(1)는 다음의 문맥(30) 및 입력(20) 시퀀스들을 출력할 수 있다:

[0303] · 문맥 시퀀스(30); "When you get this, send me"

[0304] · 입력 시퀀스(20):  $\{\{a, 0.8\}, \{s, 0.1\}, \{z, 0.1\}\},$

[0305]  $\{\{r, 0.9\}, \{t, 0.05\}, \{f, 0.05\}\},$

[0306]  $\{\{e, 0.8\}, \{w, 0.2\}\}$

[0307] 도 3에 도시된 바와 같이, 후보 발생기(2)는 입력 모델 평가  $P(\text{input} | s, M_{\text{input}})$ 를 계산해 관련된 확률로 다음 후보들을 생성한다:

[0308] { are, 0.576 }

[0309] { arena, 0.576 }

[0310] { ate, 0.032 }

[0311] { a text, 0.032 }

[0312] { a reply, 0.576 }

[0313] { a repot, 0.576 }

[0314] 그래프 평가의 일부분의 가상 예가 평가의 각 단계에서 각 후보의 확률을 강조하며 도 3에 도시되어 있다.

[0315] 도 3에서 보이는 바와 같이, 일단 다용어 "a report"("a reply" 및 "a text"에 비해, 아주 낮은 확률 바이그램)가 형성되고나면, 낮은 문맥 확률로 인해 후보 발생기(2)에 의해 가지쳐진다.

[0316] 후보 발생의 마지막에서, 의도된 시퀀스가 아닌 톱-스코어 후보들이 많이 있다.

[0317] 그러나, 동일한 후보와 문맥 시퀀스(3)가 주어지면 문맥 모델  $M_{\text{context}}$ (이 경우 트라이그램 모델)를 이용해, 후보 발생기(2)는 관련된 확률로 다음 후보들을 생성할 수 있다:

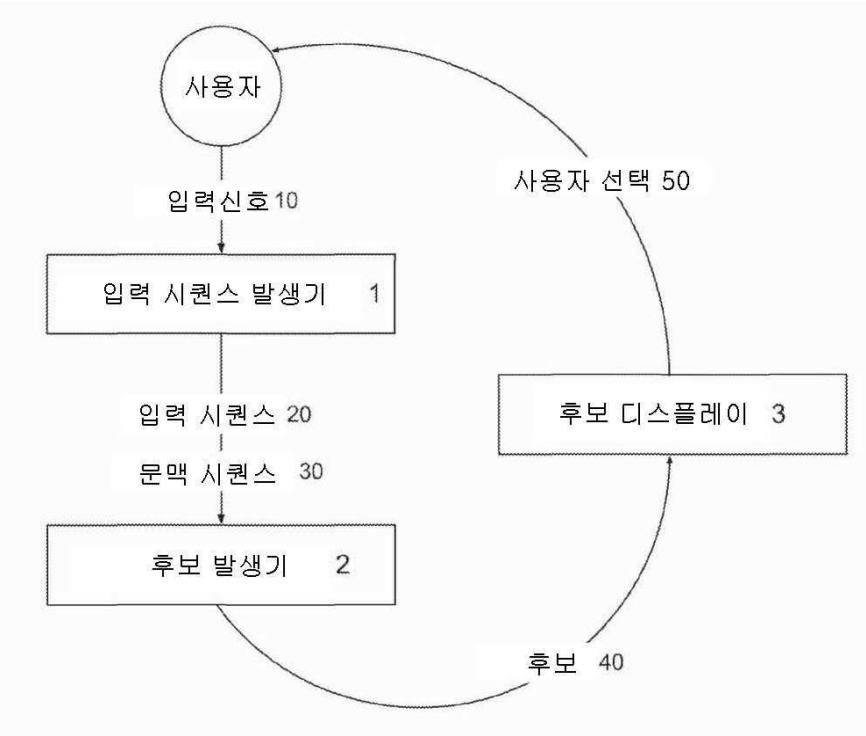
[0318] { [send me] are, 0.0000023 }

[0319] { [send me] arena, 0.000000002 }

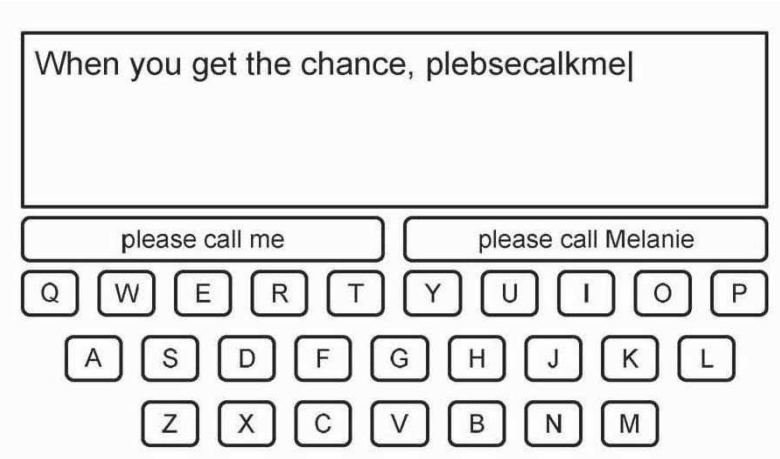
- [0320] { [send me] ate, 0.00000007 }
- [0321] { [send me] a text, 0.007 }
- [0322] { [send me] a reply, 0.005 }
- [0323] 문맥과 관련해 가장 선호될 수 있는 후보는 실제로 "text"이다.
- [0324] 마지막으로, 2개 모델로부터의 확률평가는 후보 발생기(2)에 의해 조합된다. 이들은 물론 바람직하다면, 상대적인 랭킹에 영향을 주지 않으며 정규화될 수 있으므로, 간략히 하기 위해 생략된다.
- [0325] 따라서, 최종 예시적 평가는 다음과 같다:
- [0326] { are, 0.00000132 }
- [0327] { arena, 0.00000000159 }
- [0328] { ate, 0.00000000224 }
- [0329] { a text, 0.000224 }
- [0330] { a reply, 0.00288 }
- [0331] 따라서, 이 기본적인 예에서, 모델들의 조합된 세트는 제 1 예측으로서 정확히 "a reply", 그 상황에서 명백히 가치 있는 후보인 제 2 예측으로 "a text"를 선택하는데, 이는 사용자가 단순히 T 키 대신 R 키를 쳤기 때문이다.
- [0332] 거꾸로 해도 맞고 R 키 대신 T 키를 사용자가 쳤다면, (특히 제 2 문자에서) 입력 확률 스트링의 분포는 T 문자를 지지하여 역전됐을 것이고, 상위의 2 결과들도 단순히 역전될 가능성이 있다. 어느 한 방식이든, 의도된 결과는 아주 높이 랭크되어 후보 디스플레이(3)를 통해 사용자에게 나타내도록 준비된다.
- [0333] 시스템과 관련해 언급되고 도 4에 도시된 바와 같이, 일반적인 방법은 입력 시퀀스(20)로부터 후보 발생기(2)를 이용해 하나 이상의 후보들을 발생하는 단계(100); 제 1 확률평가를 각 후보에 할당하는 단계(200)를 포함하고, 각 후보는 하나 이상의 용어 경계들에 의해 분리된 2 이상의 용어들을 포함한다. 문맥언어모델에서 상기 후보의 하나 이상의 용어들을 탐색함으로써 각 후보에 제 1 확률평가가 할당되며, 문맥언어모델은 용어들의 시퀀스들을 포함하고, 용어들의 각 시퀀스는 해당 발생확률을 갖고 문맥언어모델로부터 후보 중 하나 이상의 용어들에 해당하는 확률을 후보에 할당한다. 상기 방법에서 마지막 단계는 해당 제 1 확률평가를 기초로 하나 이상의 후보들을 폐기하는 단계(300)를 포함한다.
- [0334] 본 방법의 다른 태양들은 상술한 시스템과 유사하다. 가령, 본 방법의 바람직한 실시예에서, 하나 이상의 후보들을 발생하는 것은 후보 발생기(2)를 이용해 입력 시퀀스로부터 확률적 제약 시퀀스 그래프('PCSG')를 발생하는 단계와 용어경계 노드를 삽입하는 단계를 포함한다.
- [0335] 문맥 후보 평가를 결정하기 위해 PCSG에서 일반화 구조(및 특히 와일드카드 분기)를 실행하는 시스템과 관련해 상술한 바와 같이, 본 방법의 바람직한 실시예에서, 시퀀스 예측 세트들 중 적어도 하나는 사용자에게 의해 사용자 인터페이스에 텍스트 입력의 조절 또는 정정된 형태에 해당한다.
- [0336] 본 설명은 단지 예이며, 특허청구범위에 정의된 바와 같이 본 발명의 범위로부터 벗어남이 없이 변형 및 변경이 이루어질 수 있음이 명백할 것이다.

도면

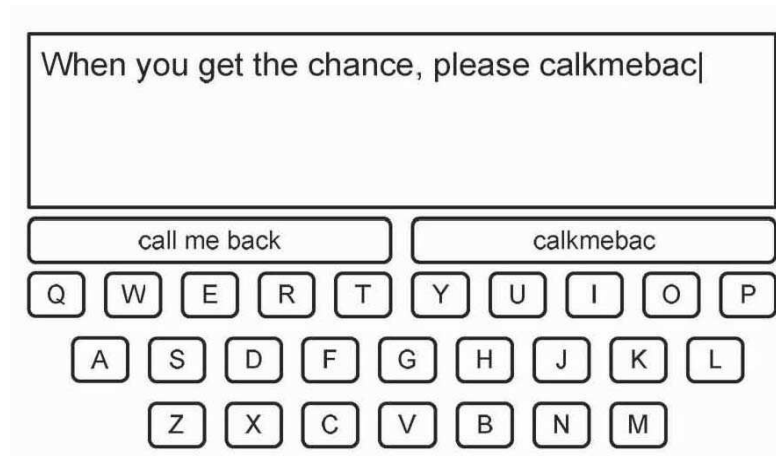
도면1



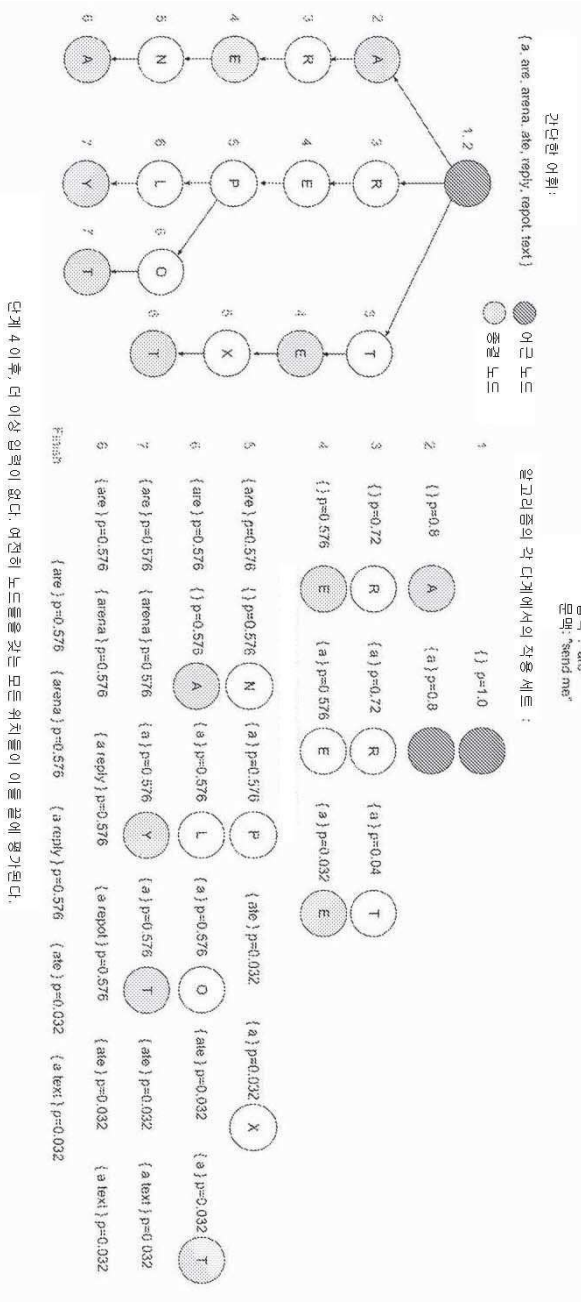
도면2a



도면2b



예제 그래프 평가



도면3

도면4

