



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0000980
(43) 공개일자 2018년01월04일

(51) 국제특허분류(Int. Cl.)
G10L 15/02 (2006.01) G10L 15/05 (2013.01)
G10L 15/183 (2013.01) G10L 15/19 (2013.01)
(52) CPC특허분류
G10L 15/02 (2013.01)
G10L 15/05 (2013.01)
(21) 출원번호 10-2016-0079411
(22) 출원일자 2016년06월24일
심사청구일자 2016년06월24일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
김지환
서울특별시 노원구 노해로 508, 604동 801호 (상계동, 상계주공6단지아파트)
이동현
서울특별시 서대문구 거북골로 154, 104동 510호 (북가좌동, 삼호아파트)
박재현
서울특별시 서초구 바우뫼로4길 12
(74) 대리인
특허법인충현

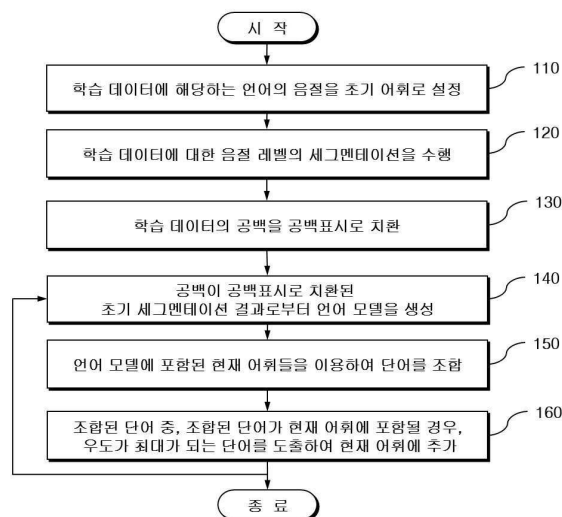
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 최대 우도에 기반한 자동 어휘 생성

(57) 요약

본 발명은 자동 어휘 생성 방법에 관한 것으로서, 학습 데이터에 해당하는 언어의 음절을 초기 어휘로 설정하는 단계, 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 수행하는 단계, 상기 학습 데이터의 공백을 공백표시로 치환하는 단계, 상기 공백이 공백표시로 치환된 초기 세그멘테이션 결과로부터 언어 모델을 생성하는 단계, 상기 언어 모델에 포함된 현재 어휘들을 이용하여 단어를 조합하는 단계, 및 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여, 현재 어휘에 추가하는 단계를 포함하고, 상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하는 단계 및 상기 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 단계를 반복하는 것을 특징으로 함으로써, 기 설정된 수만큼의 단어를 가지는 어휘를 자동으로 생성할 수 있고, 생성된 어휘로부터 언어 모델을 성공적으로 생성할 수 있다.

대표도 - 도1



(52) CPC특허분류

G10L 15/183 (2013.01)

G10L 15/19 (2013.01)

명세서

청구범위

청구항 1

학습 데이터에 해당하는 언어의 음절을 초기 어휘로 설정하는 단계;

상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 수행하는 단계;

상기 학습 데이터의 공백을 공백표시로 치환하는 단계;

상기 공백이 공백표시로 치환된 초기 세그멘테이션 결과로부터 언어 모델을 생성하는 단계;

상기 언어 모델에 포함된 현재 어휘들을 이용하여 단어를 조합하는 단계; 및

상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여, 현재 어휘에 추가하는 단계를 포함하고,

상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하는 단계 및 우도가 최대가 되는 단어를 도출하여, 현재 어휘에 추가하는 단계를 반복하는 것을 특징으로 하는 자동 어휘 생성 방법.

청구항 2

제 1 항에 있어서,

상기 음절을 초기 어휘로 설정하는 단계는,

상기 학습 데이터에 해당하는 언어의 각 음절을 공백표시가 없는 음절, 좌측에 공백표시가 있는 음절, 우측에 공백표시가 있는 음절, 및 좌우측에 공백표시가 있는 음절로 복제하는 단계를 포함하는 것을 특징으로 하는 자동 어휘 생성 방법.

청구항 3

제 1 항에 있어서,

상기 학습 데이터의 공백을 공백표시로 치환하는 단계는,

상기 학습 데이터로부터 세그멘테이션 된 음절의 좌측 또는 우측에 공백이 있는 음절을 공백에 대응되도록 상기 좌측 또는 우측에 공백표시가 있는 음절로 복제하여 대체하는 것을 특징으로 하는 자동 어휘 생성 방법.

청구항 4

제 1 항에 있어서,

상기 언어 모델을 생성하는 단계는,

현재 세그멘테이션 결과로부터 음절 단위 카운트를 계산하는 단계; 및

상기 음절 단위 카운트를 이용하여 현재 어휘에 대한 언어 모델을 생성하는 단계를 포함하는 자동 어휘 생성 방법.

청구항 5

제 4 항에 있어서,

상기 음절 단위 카운트는,

바이그램(bi-gram) 카운트 또는 트라이그램(tri-gram) 카운트 중 어느 하나 이상을 포함하는 것을 특징으로 하는 자동 어휘 생성 방법.

청구항 6

제 1 항에 있어서,

상기 도출된 단어를 현재 어휘에 추가하는 단계는,

현재 세그멘테이션 결과에 상기 도출된 단어를 추가하여 세그멘테이션 결과를 갱신하는 단계; 및

상기 갱신된 세그멘테이션 결과로부터 언어 모델을 갱신하는 단계를 포함하는 것을 특징으로 하는 자동 어휘 생성 방법.

청구항 7

제 1 항 내지 제 6 항 중에 어느 한 항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 8

학습 데이터에 해당하는 언어의 음절과 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 이용하여 설정된 현재 어휘들을 이용하여 단어를 조합하고, 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 처리부; 및

상기 학습 데이터 및 상기 현재 어휘들로 구성되는 언어 모델을 저장하는 저장부를 포함하고,

상기 처리부는,

상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하고 상기 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 과정을 반복하는 것을 특징으로 하는 자동 어휘 생성 장치.

청구항 9

제 8 항에 있어서,

상기 처리부는,

상기 학습 데이터로부터 세그멘테이션 된 음절의 좌측 또는 우측에 공백이 있는 음절을 공백에 대응되도록 상기 좌측 또는 우측에 공백표시가 있는 음절로 복제하여 대체하는 것을 특징으로 하는 자동 어휘 생성 장치.

청구항 10

제 8 항에 있어서,

상기 처리부는,

현재 세그멘테이션 결과로부터 음절 단위 카운트를 계산한 결과를 이용하여 현재 어휘에 대한 언어 모델을 생성하는 것을 특징으로 하는 자동 어휘 생성 장치.

청구항 11

제 8 항에 있어서,

상기 처리부는,

현재 세그멘테이션 결과에 상기 도출된 단어를 추가하여 세그멘테이션 결과를 갱신하는 것을 특징으로 하는 자동 어휘 생성 장치.

발명의 설명

기술 분야

본 발명은 자동 어휘 생성 방법에 관한 것으로서, 더욱 상세하게는 최대 우도를 이용하여 자동으로 어휘를 생성하는 자동 어휘 생성 방법 및 그 장치에 관한 것이다.

[0001]

배경 기술

- [0002] 무제한 어휘 음성 인식을 위해선, 어휘내의 단어들을 조합하여 임의의 문장을 생성할 수 있어야 하는데, 기존의 어휘 생성방법들은 out-of-lexicon 단어들이 어휘에 포함되지 않는 문제점과, 형태소 분석 오류 및 세그멘테이션(Segmentation) 에러가 음성인식의 에러로 전파된다는 문제점이 있다.

선행기술문헌

특허문헌

- [0003] (특허문헌 0001) 한국공개특허공보 제10-2000-0019150호 "음성 모델의 생성 방법"

발명의 내용

해결하려는 과제

- [0004] 본 발명이 해결하고자 하는 첫 번째 과제는 최대 우도를 이용하여 자동으로 어휘를 생성하는 자동 어휘 생성 방법을 제공하는 것이다.
- [0005] 본 발명이 해결하고자 하는 두 번째 과제는 최대 우도를 이용하여 자동으로 어휘를 생성하는 자동 어휘 생성 장치를 제공하는 것이다.

과제의 해결 수단

- [0006] 본 발명은 상기 첫 번째 과제를 달성하기 위하여, 학습 데이터에 해당하는 언어의 음절을 초기 어휘로 설정하는 단계; 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 수행하는 단계; 상기 학습 데이터의 공백을 공백표시로 치환하는 단계; 상기 공백이 공백표시로 치환된 초기 세그멘테이션 결과로부터 언어 모델을 생성하는 단계; 상기 언어 모델에 포함된 현재 어휘들을 이용하여 단어를 조합하는 단계; 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여, 현재 어휘에 추가하는 단계를 포함하고, 상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하는 단계 및 상기 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 단계를 반복하는 것을 특징으로 하는 자동 어휘 생성 방법을 제공한다.
- [0007] 본 발명의 실시예에 의하면, 상기 음절을 초기 어휘로 설정하는 단계는, 상기 학습 데이터에 해당하는 언어의 각 음절을 공백표시가 없는 음절, 좌측에 공백표시가 있는 음절, 우측에 공백표시가 있는 음절, 및 좌우측에 공백표시가 있는 음절로 복제하는 단계를 포함하는 것을 특징으로 하는 자동 어휘 생성 방법일 수 있다.
- [0008] 본 발명의 실시예에 의하면, 상기 학습 데이터의 공백을 공백표시로 치환하는 단계는, 상기 학습 데이터로부터 세그멘테이션 된 음절의 좌측 또는 우측에 공백이 있는 음절을 공백에 대응되도록 상기 좌측 또는 우측에 공백표시가 있는 음절로 복제하여 대체하는 것을 특징으로 하는 자동 어휘 생성 방법일 수 있다.
- [0009] 본 발명의 실시예에 의하면, 상기 언어 모델을 생성하는 단계는, 현재 세그멘테이션 결과로부터 음절 단위 카운트를 계산하는 단계; 및 상기 음절 단위 카운트를 이용하여 현재 어휘에 대한 언어 모델을 생성하는 단계를 포함하는 자동 어휘 생성 방법일 수 있다.
- [0010] 본 발명의 실시예에 의하면, 상기 음절 단위 카운트는, 바이그램(bi-gram) 또는 트라이그램(tri-gram) 중 어느 하나 이상을 포함하는 것을 특징으로 하는 자동 어휘 생성 방법일 수 있다.
- [0011] 본 발명의 실시예에 의하면, 상기 도출된 단어를 현재 어휘에 추가하는 단계는, 현재 세그멘테이션 결과에 상기 도출된 단어를 추가하여 세그멘테이션 결과를 갱신하는 단계; 및 상기 갱신된 세그멘테이션 결과로부터 언어 모델을 갱신하는 단계를 포함하는 것을 특징으로 하는 자동 어휘 생성 방법일 수 있다.
- [0012] 본 발명은 상기 두 번째 과제를 달성하기 위하여, 학습 데이터에 해당하는 언어의 음절과 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 이용하여 설정된 현재 어휘들을 이용하여 단어를 조합하고, 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 처리부; 및 상기 학습 데이터 및 상기 현재 어휘들로 구성되는 언어 모델을 저장하는 저장부를 포함하고, 상기 처리부는, 상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하고 상기 우도가 최대가 되는 단어를 도출하

여 현재 어휘에 추가하는 과정을 반복하는 것을 특징으로 하는 자동 어휘 생성 장치를 제공한다.

발명의 효과

[0013] 본 발명에 따르면, 최대 우도를 이용하여 어휘를 생성하는바, 자동으로 어휘를 생성하여, 이 어휘내의 단어를 조합하여 임의의 문장을 생성 할 수 있다. 또한, out-of-lexicon 단어들이 어휘에 포함되지 않는 문제점과 형태소 분석과 공백 세그멘테이션 에러가 음성인터페이스 에러로 전파된다는 문제점을 해결한다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시예에 따른 자동 어휘 생성 방법의 흐름도이다.

도 2 내지 5는 본 발명의 실시예에 따른 자동 어휘 생성 방법의 흐름도이다.

도 6은 본 발명의 일 실시예에 따른 자동 어휘 생성 장치의 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0015] 본 발명에 관한 구체적인 내용의 설명에 앞서 이해의 편의를 위해 본 발명이 해결하고자 하는 과제의 해결 방안의 개요 혹은 기술적 사상의 핵심을 우선 제시한다.

[0016] 본 발명의 일 실시예에 따른 자동 어휘 생성 방법은 학습 데이터에 해당하는 언어의 음절을 초기 어휘로 설정하는 단계, 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 수행하는 단계, 상기 학습 데이터의 공백을 공백표시로 치환하는 단계, 상기 공백이 공백표시로 치환된 초기 세그멘테이션 결과로부터 언어 모델을 생성하는 단계, 상기 언어 모델에 포함된 현재 어휘들을 이용하여 단어를 조합하는 단계, 및 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여, 현재 어휘에 추가하는 단계를 포함하고, 상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하는 단계 및 상기 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 단계를 반복하는 것을 특징으로 한다.

[0017] 이하 첨부된 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명을 용이하게 실시할 수 있는 실시 예를 상세히 설명한다. 그러나 이들 실시예는 본 발명을 보다 구체적으로 설명하기 위한 것으로, 본 발명의 범위가 이에 의하여 제한되지 않는다는 것은 당업계의 통상의 지식을 가진 자에게 자명할 것이다.

[0018] 본 발명이 해결하고자 하는 과제의 해결 방안을 명확하게 하기 위한 발명의 구성을 본 발명의 바람직한 실시예에 근거하여 첨부 도면을 참조하여 상세히 설명하되, 도면의 구성요소들에 참조번호를 부여함에 있어서 동일 구성요소에 대해서는 비록 다른 도면상에 있더라도 동일 참조번호를 부여하였으며 당해 도면에 대한 설명시 필요한 경우 다른 도면의 구성요소를 인용할 수 있음을 미리 밝혀둔다. 아울러 본 발명의 바람직한 실시예에 대한 동작 원리를 상세하게 설명함에 있어 본 발명과 관련된 공지 기능 혹은 구성에 대한 구체적인 설명 그리고 그 이외의 제반 사항이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우, 그 상세한 설명을 생략한다.

[0019] 도 1은 본 발명의 일 실시예에 따른 자동 어휘 생성 방법의 흐름도이고, 도 2 내지 5는 본 발명의 실시예에 따른 자동 어휘 생성 방법의 흐름도이다.

[0020] 본 발명의 일 실시예에 따른 자동 어휘 생성 방법은 최초 해당 언어의 음절로부터 시작하여 최대 우도를 이용하여 단어들을 추가하여 자동으로 어휘를 생성한다. 자동으로 어휘를 생성함에 있어서, 공백을 고려한 어휘를 생성한다. 이하, 자동으로 어휘를 생성하는 구체적인 과정에 대해 도 1 내지 도6을 이용하여 자세히 설명하도록 한다.

[0021] 110 단계는 학습 데이터에 해당하는 언어의 음절을 초기 어휘로 설정하는 단계이다.

[0022] 보다 구체적으로, 자동 어휘를 생성하고자 하는, 즉 학습 데이터에 해당하는 언어의 음절을 초기 어휘로 설정한다. 다양한 언어들에 적용될 수 있으며, 특히 동아시아 언어들의 어휘를 자동으로 생성하는데 이용될 수 있다. 한국어에 대한 어휘를 자동으로 생성하기 위해서, 우선 EUC-KR에 정의된 한국어 음절을 초기 어휘로 설정한다. EUC-KR에 정의된 한국어 음절은 약 3900개로 구성되어 있다.

[0023] 한국어는 공백(띄어쓰기)가 음절 좌측 또는 우측에 위치하는 경우, 공백의 위치에 따라 뜻이 달라질 수 있다. 따라서, 공백에 따른 에러를 제거하기 위하여, 공백표시를 이용할 수 있다.

[0024] 이를 위하여, 도 2의 210 단계와 같이, 상기 학습 데이터에 해당하는 언어의 각 음절을 네 가지 형태, 즉 공백

표시가 없는 음절, 좌측에 공백표시(underscore)가 있는 음절, 우측에 공백표시가 있는 음절, 및 좌우측에 공백표시가 있는 음절로 복제하는 단계를 더 포함할 수 있다. 예를 들어,

- [0025] [가]라는 음절은 공백표시가 없는 음절(오리지널(original) 음절)[가], 좌측에 공백표시가 있는 음절[_가], 우측에 공백표시가 있는 음절[가_], 및 좌우측에 공백표시가 있는 음절[_가_]로 나타낸다. 이와 같이, 공백을 공백표시로 표시하여, 정확한 어휘를 생성하고 해당 음성을 인식하는데 제공할 수 있다.
- [0026] 초기 어휘를 설정한 이후에는 학습 데이터에 대한 120 단계에서 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 수행한다.
- [0027] 보다 구체적으로, 자동으로 학습 데이터로부터 학습을 통해 자동 어휘를 생성하기 위하여, 학습하고자 하는 학습 데이터에 대한 음절 레벨의 세그멘테이션을 수행한다. 음절 레벨의 세그멘테이션을 통해 학습 데이터를 음절 단위로 세분화한다.
- [0028] 학습 데이터에 대한 세그멘테이션을 수행한 후, 130 단계에서 학습 데이터의 공백을 공백표시로 치환한다.
- [0029] 공백을 공백표시로 치환하기 위하여, 도 3과 같이, 310 단계에서 상기 학습 데이터로부터 세그멘테이션 된 음절의 좌측 또는 우측에 공백이 있는 음절을 공백에 대응되도록 상기 좌측 또는 우측에 공백표시가 있는 음절로 복제하여 대체함으로써 공백을 공백표시로 치환한다.
- [0030] 언어의 음절에 공백표시를 적용하여 복제한 방법과 같이, 상기 학습 데이터로부터 세그멘테이션 된 음절의 좌측 또는 우측에 공백이 있는 경우, 해당 공백에 대응되도록 상기 좌측 또는 우측에 공백표시가 있는 음절로 복제하여 대체한다. 단순히 음절을 이용하는 것이 아닌 해당 음절의 좌측 또는 우측에 위치하는 공백을 함께 표시하여 이용할 수 있도록 좌측 또는 우측에 공백표시가 있는 음절로 복제하여 대체한다. 상기 좌측 또는 우측에 공백이 있는 세그멘테이션 된 음절을 모두 복제된 음절로 대체한 이후에는 공백이 모두 공백표시로 표시되어 있는바, 학습 데이터의 공백을 모두 공백표시로 치환한다.
- [0031] 공백을 공백표시로 치환한 이후에는 130 단계에서 상기 공백이 공백표시로 치환된 초기 세그멘테이션 결과로부터 언어 모델을 생성한다.
- [0032] 보다 구체적으로, 학습 데이터에 대한 공백이 공백표시로 치환된 세그멘테이션을 수행하고, 이후 그 세그멘테이션 결과로부터 언어 모델을 생성한다. 상기 언어 모델을 생성함에 있어서, 도 4와 같이, 현재 세그멘테이션 결과로부터 음절 단위 카운트를 계산하는 410 단계 및 상기 음절 단위 카운트를 이용하여 현재 어휘에 대한 언어 모델을 생성하는 420 단계의 과정을 수행할 수 있다.
- [0033] 먼저 공백이 공백표시로 치환된 초기 세그멘테이션 결과에 대해 음절 단위 카운트 (n-gram count)를 계산하고, 현재 음절 단위 카운트에 기초하여, 현재 어휘에 대한 언어 모델을 생성할 수 있다. 상기 음절 단위 카운트는, 바이그램(bi-gram) 카운트 또는 트라이그램(tri-gram) 카운트 중 어느 하나 이상을 포함할 수 있다. 바이그램 카운트는 두 음절씩 카운트하는 것이고, 트라이그램 카운트는 세 음절씩 카운트하는 것으로, 다른 엔그램 카운트를 이용할 수도 있다. 정확한 어휘 생성을 위하여, 바이그램 카운트 및 트라이그램 카운트를 이용할 수 있다.
- [0034] 상기와 같이, 음절 및 학습 데이터에 따른 초기 세그멘테이션 결과를 이용하여 현재 어휘에 대한 언어 모델을 생성할 수 있다. 이렇게 생성된 언어 모델에 대해 자동으로 어휘를 생성하기 위하여, 하기의 과정을 거친다.
- [0035] 140 단계에서 상기 언어 모델에 포함된 현재 어휘들을 이용하여 단어를 조합하고, 150 단계에서 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여, 현재 어휘에 추가한다.
- [0036] 보다 구체적으로, 현재 어휘들을 이용하여 모든 어휘 단어 조합을 생성하고, 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출한다.
- [0037] 최대 우도를 이용하여 단어를 추가하는 과정에 대한 상세한 설명을 위해, 먼저 학습 데이터에 $s_1, s_2, \dots, s_n$ 까지 n개의 문장으로 이루어진 코퍼스가 있다고 가정하자. 해당 코퍼스에서 생성된 언어 모델을 f 라 하자. α 를 현재의 lexicon으로부터 2단어를 조합하여 만들 수 있는 모든 가능한 단어 set이라 하고, θ 는 α 의 원소라 하자. 임의의 θ 가 lexicon에 포함되었을 때 코퍼스로부터 생성되는 언어모델을 $\{f(*|\theta), \theta \in \alpha\}$ 와 같이 정의 할 수 있다. 각 문장이 독립이고 고정되어 있으므로 언어모델을 다음과 같이 표현할 수 있다.

수학식 1

$$f(s_1, s_2, \dots, s_n | \theta)$$

$$= f(s_1 | \theta) \times f(s_2 | \theta) \times \dots \times f(s_n | \theta)$$

여기서 $s_1, s_2, \dots, s_n$ 은 고정이며 θ 는 가변이다. 이에 대한 우도 L은 다음과 같이 표현된다.

수학식 2

$$L(\theta | s_1, \dots, s_n)$$

$$= f(s_1, s_2, \dots, s_n | \theta) = \prod_{i=1}^n f(s_i | \theta)$$

계산에 편의상 로그를 취하여 우도를 다음과 같이 표현한다.

수학식 3

$$\ln L(\theta | s_1, \dots, s_n) = \sum_{i=1}^n \ln f(s_i | \theta)$$

제안된 ML estimation 방법으로 $\hat{\theta}$ 를 다음과 같이 찾는다.

수학식 4

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \alpha} \sum_{i=1}^n \ln f(s_i | \theta)$$

$\hat{\theta}$ 를 찾은 결과 α 와 f 가 업데이트 된다. 이러한 과정을 우도가 증가하는 동안 반복한다.

우도를 tri-gram을 이용하여 estimation 할 수 있다. 문장 s_j 가 J 개의 단어열, $lw_1, \dots, lw_J$, 로 이루어 졌을 때 s_j 에 대한 우도는 다음과 같이 계산된다.

수학식 5

$$f(s_j | \theta) = p(lw_1, lw_2, \dots, lw_J) =$$

$$\prod_{j=1}^J p(lw_j | lw_1, \dots, lw_{j-1}) \approx$$

$$\prod_{j=1}^J p(lw_j | lw_{j-2}, lw_{j-1})$$

$$p(lw_j | lw_{j-2}, lw_{j-1})$$
 는 다음과 같이 계산된다.

수학식 6

$$p(lw_j | lw_{j-2}, lw_{j-1}) = \frac{c(lw_{j-2}, lw_{j-1}, lw_j)}{c(lw_{j-2}, lw_{j-1})}$$

[0053]

[0054]

[0055]

[0056]

[0057]

[0058]

[0059]

[0060]

[0061]

[0062]

[0063]

[0064]

상기와 같이, 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하는 단계 및 상기 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 단계를 반복하여, 원하는 만큼의 어휘를 자동으로 생성할 수 있다. 어휘 단어 수를 제한하여, 필요한 수만큼의 어휘를 생성할 수 있으며, 우도가 증가하는 동안 반복함으로써 어휘 생성의 정확성을 높일 수 있다.

상기 도출된 단어를 현재 어휘에 추가한 이후에는 도 5와 같이, 현재 세그멘테이션 결과에 상기 도출된 단어를 추가하여 세그멘테이션 결과를 갱신하는 510 단계 및 상기 갱신된 세그멘테이션 결과로부터 언어 모델을 갱신하는 520 단계를 더 포함할 수 있다. 새로 추가되는 단어와 중복되는 현재 세그멘테이션을 상기 단어로 대체하여 제한된 단어의 수에 정확한 어휘가 포함될 수 있도록 하고, 이를 통해 세그멘테이션 결과가 갱신되면, 갱신된 세그멘테이션 결과에 대해 음절 단위 카운트를 계산하여 언어 모델을 갱신한다.

위와 같은 과정을 통해 개발자가 지정한 수의 단어를 가지는 어휘를 생성할 수 있고, 생성된 어휘로부터 언어 모델을 성공적으로 생성할 수 있다. 상기 과정을 통해 생성된 어휘와 언어모델을 사용함으로써, 음성인식기는 동아시아 언어에 대해서 무제한 어휘의 문장을 생산할 수 있다. 모든 과정이 자동으로 이루어지며 어떠한 사전 지식(pre-knowledge)이 필요하지 않기 때문에 1) out-of-lexicon 단어들을 lexicon에 포함시키지 못하는 문제 및 2) 형태소 분석과 세그멘테이션 에러가 음성인식 에러로 전파되는 문제를 방지할 수 있다.

상기 과정을 통해 생성된 어휘는 공백표시가 포함되어 있다. 따라서 이에 기반한 음성인식기의 결과에도 공백표시가 포함될 수 있다. 공백표시는 가독성(readability) 저하를 불러올 수 있기 때문에 이에 대한 후처리를 더 수행할 수 있다. 음성 인식 결과에 따른 공백은 좌측 또는 우측에 공백표시가 된 어휘 단어로 생성된다. 중복된 공백표시가 된 경우에는 첫 번째 공백표시는 좌측 어휘 단어에서 생성되고, 두 번째 공백표시는 우측 어휘 단어에서 생성된다. 이러한 문제를 해결하기 위하여, 보수적인 규칙을 적용할 수 있다. 즉, 두 개의 연이은 공백표시는 하나의 공백표시를 제거하고, 하나의 공백표시로 변환할 수 있다. 예를 들어, [_나는_ _친구 가_ _좋다_] 는 [나는 친구가 좋다] 로 최종 인식 결과로 변환할 수 있다.

도 6은 본 발명의 일 실시예에 따른 자동 어휘 생성 장치의 블록도이다.

본 발명의 일 실시예에 따른 자동 어휘 생성 장치(600)는 처리부(610) 및 저장부(620)로 구성되며, 통신부를 더 포함할 수 있다.

처리부(610)는 학습 데이터에 해당하는 언어의 음절과 상기 학습 데이터에 대한 음절 레벨의 세그멘테이션을 이용하여 설정된 현재 어휘들을 이용하여 단어를 조합하고, 상기 조합된 단어 중, 조합된 단어가 현재 어휘에 포함될 경우, 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하며, 상기 어휘 단어 수의 제한된 범위 내에서 상기 단어를 조합하고 상기 우도가 최대가 되는 단어를 도출하여 현재 어휘에 추가하는 과정을 반복한다.

처리부(610)는 상기 학습 데이터로부터 세그멘테이션 된 음절의 좌측 또는 우측에 공백이 있는 음절을 공백에 대응되도록 상기 좌측 또는 우측에 밑줄 표시가 있는 음절로 복제하여 대체할 수 있고, 현재 세그멘테이션 결과로부터 음절 단위 카운트를 계산한 결과를 이용하여 현재 어휘에 대한 언어 모델을 생성할 수 있으며, 현재 세그멘테이션 결과에 상기 도출된 단어를 추가하여 세그멘테이션 결과를 갱신할 수 있다.

처리부(610)가 자동 어휘를 생성하는 과정은 도 1 내지 도 5의 자동 어휘 생성 방법에 대응하는바, 중복되는 설명은 생략하도록 한다.

저장부(620)는 학습 데이터 및 현재 어휘들로 구성되는 언어 모델을 저장할 수 있다. 통신부는 언어의 음절에 대한 데이터, 학습 데이터를 외부로부터 수신할 수 있으며, 생성된 어휘 및 언어 모델을 외부로 송신할 수 있다.

본 발명의 실시예들은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계

되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체 (magnetic media), CD-ROM, DVD와 같은 광기록 매체 (optical media), 플롭티컬 디스크 (floptical disk)와 같은 자기-광 매체 (magneto-optical media), 및 롬 (ROM), 램 (RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0065] 이상과 같이 본 발명에서는 구체적인 구성 요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.

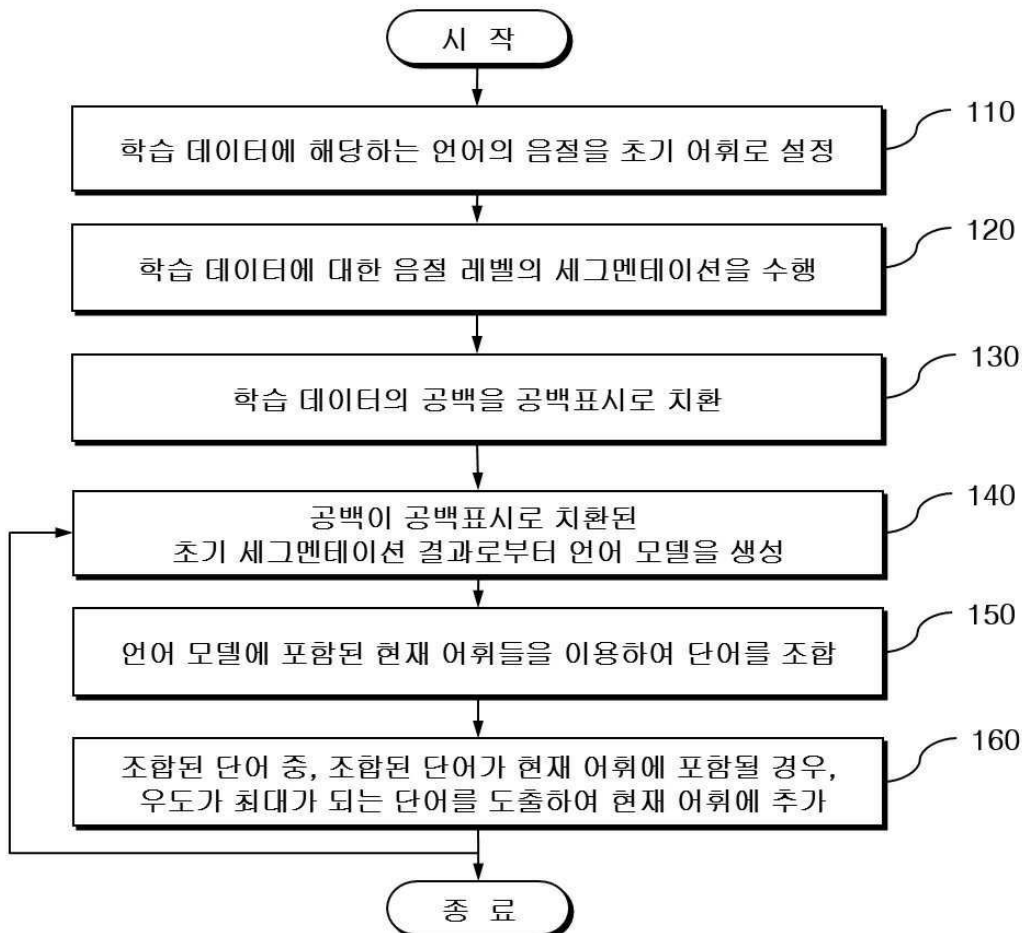
[0066] 따라서, 본 발명의 사상은 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

부호의 설명

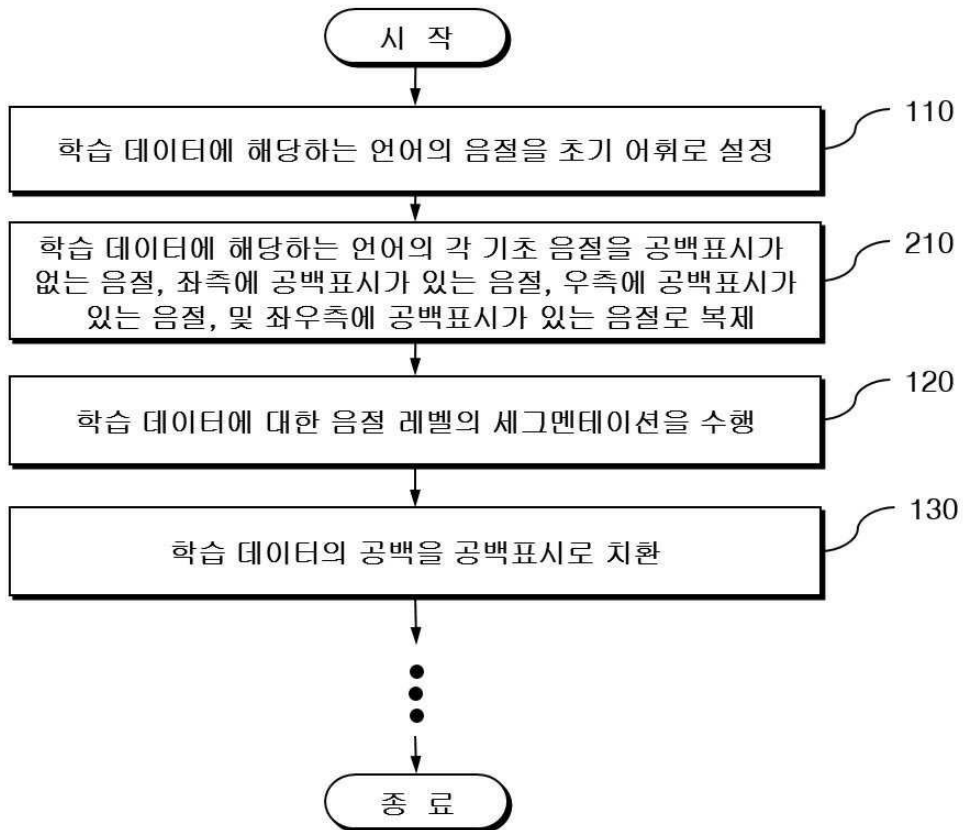
[0067] 600: 자동 어휘 생성 장치
610: 처리부
620: 저장부

도면

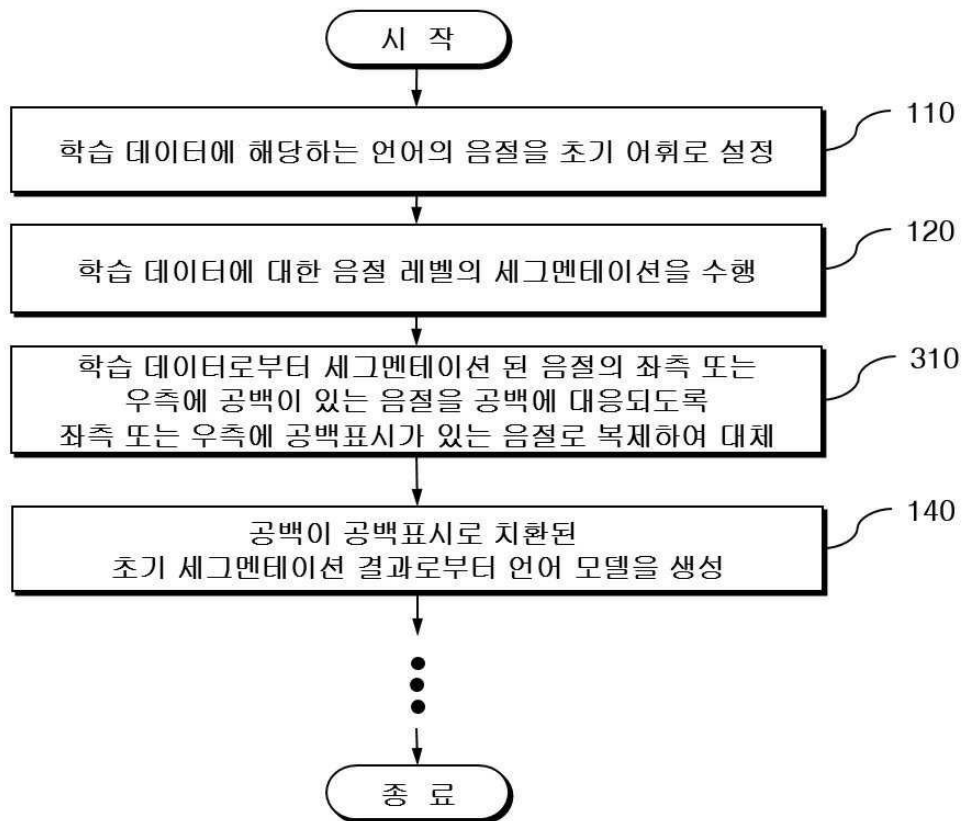
도면1



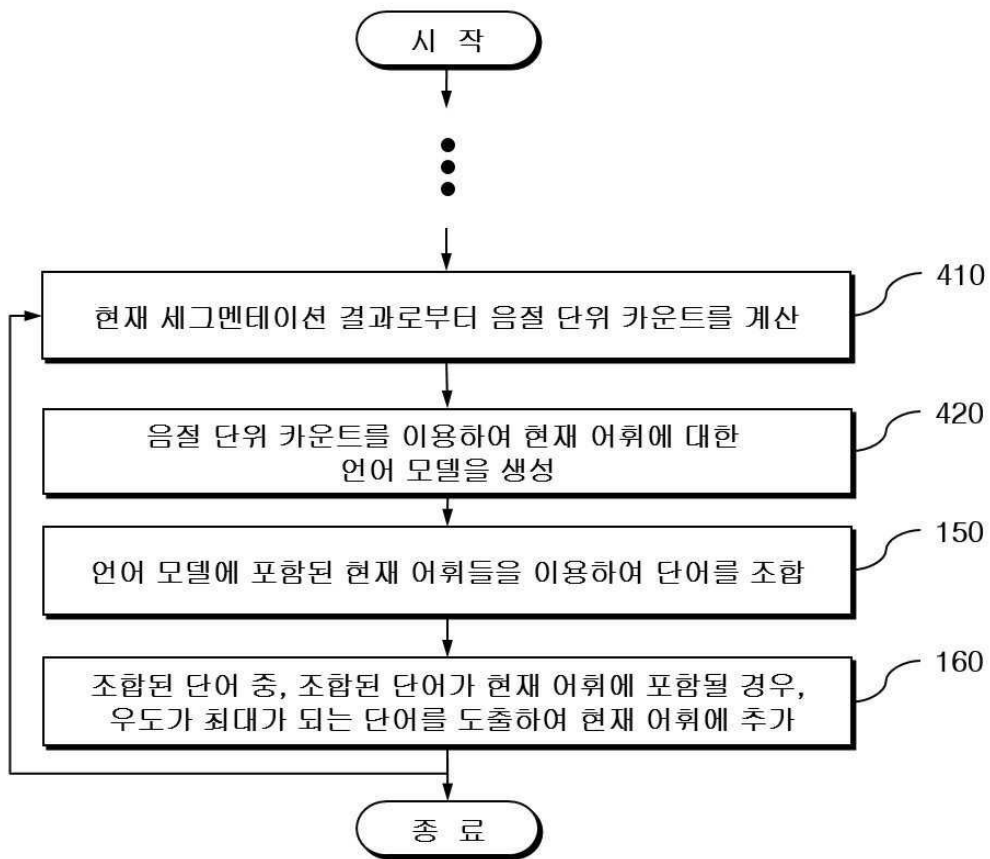
도면2



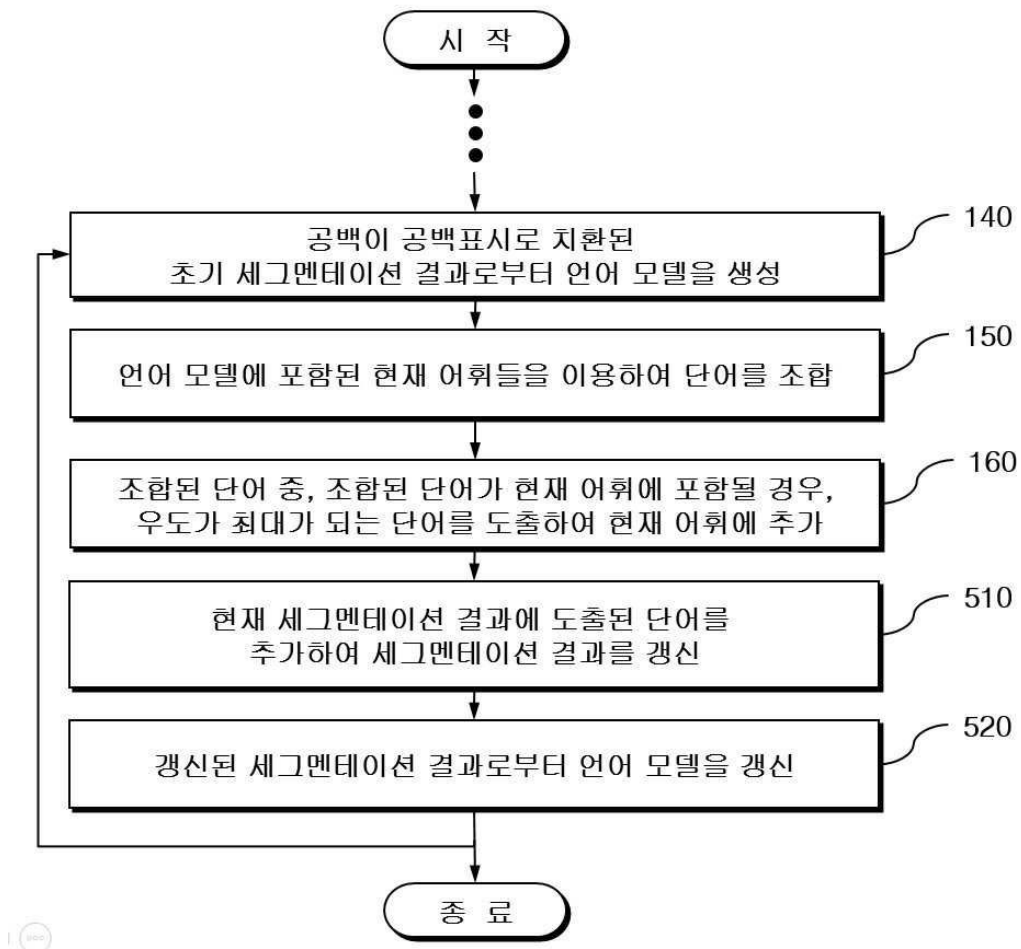
도면3



도면4



도면5



도면6

