(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0036623
(43) 공개일자 2023년03월15일

(51) 국제특허분류(Int. Cl.)
G06N 5/02 (2023.01) *G06F 40/205* (2020.01)
G06F 40/247 (2020.01) *G06F 40/51* (2020.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06N 5/02 (2023.01)
G06F 40/205 (2020.01)
(21) 출원번호 10-2021-0118924
(22) 출원일자 2021년09월07일
심사청구일자 2021년09월07일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스애기능생활관 311호 NLP&AI연구실
문현석
서울특별시 강남구 자곡로 260, 422동 903호
(뒷면에 계속)
(74) 대리인
김홍석

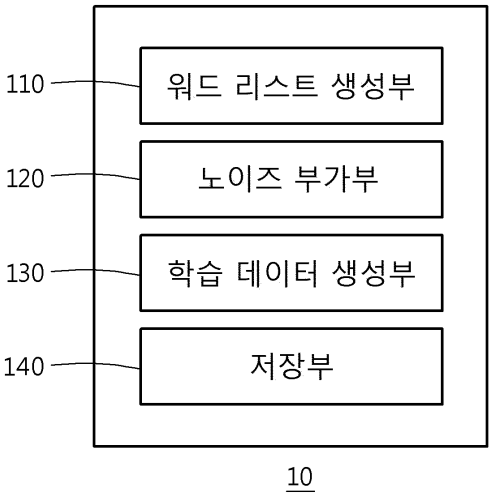
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **기계번역 사후교정을 위한 학습 데이터 생성 장치 및 방법**

(57) 요 약

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는, 기계번역 사후교정 모델 생성을 위한 학습 데이터 생성 방법이 개시된다. 상기 학습 데이터 생성 방법은 병렬 말뭉치를 이용하여 워드 리스트를 생성하는 단계, 병렬 말뭉치에 포함되고 제1 문장과 제2 문장을 포함하는 문장쌍에 대하여, 상기 제2 문장에 노이즈를 부가하여 노이즈가 부가된 문장을 생성하는 단계, 상기 제1 문장을 상기 학습 데이터를 구성하는 소스 문장(Source Sentence, SRC)으로 하고, 상기 노이즈가 부가된 문장을 상기 학습 데이터를 구성하는 기계번역문(Machine Translation Sentence, MT)으로 하고, 상기 제2 문장을 상기 학습 데이터를 구성하는 사후교정문(Post-edited Sentence, PE)으로 하는 상기 학습 데이터를 생성하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 40/247 (2020.01)
G06F 40/51 (2020.01)
G06F 40/58 (2020.01)
G06N 20/00 (2021.08)

(72) 발명자
박찬준
서울특별시 송파구 중대로 24, 306동 1303호 (문정동, 올림픽훼밀리타운)
어수경
경상남도 합천군 삼가면 금리4길 17

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126407
과제번호	2018-0-01405-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는, 기계번역 사후교정 모델 생성을 위한 학습 데이터 생성 방법에 있어서,

병렬 말뭉치를 이용하여 워드 리스트를 생성하는 단계;

병렬 말뭉치에 포함되고 제1 언어로 된 제1 문장과 제2 언어로 된 제2 문장을 포함하는 문장쌍에 대하여, 상기 제2 문장에 노이즈를 부가하여 노이즈가 부가된 문장을 생성하는 단계;

상기 제1 문장을 상기 학습 데이터를 구성하는 소스 문장(Source Sentence, SRC)으로 하고, 상기 노이즈가 부가된 문장을 상기 학습 데이터를 구성하는 기계번역문(Machine Translation Sentence, MT)으로 하고, 상기 제2 문장을 상기 학습 데이터를 구성하는 사후교정문(Post-edited Sentence, PE)으로 하는 상기 학습 데이터를 생성하는 단계를 포함하는 학습 데이터 생성 방법.

청구항 2

제1항에 있어서,

상기 워드 리스트를 생성하는 단계는,

상기 병렬 말뭉치에 포함되는 단어들을 포함하는 상기 워드 리스트를 생성하고,

상기 워드 리스트에 포함된 단어들 각각에는 품사가 태깅되어 있는,

학습 데이터 생성 방법.

청구항 3

제2항에 있어서,

상기 노이즈가 부가된 문장을 생성하는 단계는,

상기 제2 문장에 포함된 단어를 상기 워드 리스트에 포함된 단어로 교체하여 상기 노이즈가 부가된 문장을 생성하는,

학습 데이터 생성 방법.

청구항 4

제2항에 있어서,

상기 노이즈가 부가된 문장을 생성하는 단계는,

상기 제2 문장에 포함된 단어를 상기 워드 리스트에 포함된 단어들 중 상기 제2 문장에 포함된 단어의 품사와 동일한 품사를 갖는 단어로 교체하여 상기 노이즈가 부가된 문장을 생성하는,

학습 데이터 생성 방법.

청구항 5

제2항에 있어서,

상기 노이즈가 부가된 문장을 생성하는 단계는,

상기 제2 문장에 포함된 단어를 상기 제2 문장에 포함된 단어와 동일한 의미를 갖거나 유사한 의미를 갖는 단어로 교체하여 상기 노이즈가 부가된 문장을 생성하는,

학습 데이터 생성 방법.

청구항 6

제5항에 있어서,

상기 동일한 의미를 갖거나 유사한 의미를 갖는 단어는 워드넷(WordNet)에 기초하여 획득되는,

학습 데이터 생성 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 지능 소프트웨어에 관한 것으로, 보다 상세하게는 기계번역 사후교정을 위한 학습 데이터를 생성하는 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 인공 지능 기반 기계번역 시스템을 통한 번역문에는, 사람 수준에서 교정해야 하는 오류들이 다수 포함되어 있다. 이에, 번역문의 오류를 수정하는 사람의 노력을 경감시키고자, 이러한 오류들을 자체적으로 교정하기 위한 시스템을 개발하는 연구 분야인 기계번역 사후교정(Automatic Post Editing, APE) 연구가 세계적으로 활발하게 이루어지고 있는 실정이다. 하지만, 기계번역 사후교정 연구는 다량의 데이터가 존재하는 영어-독일어 등에서 이루어질 뿐, 아직 한국어와 관련한 사후교정 연구 및 사후교정 시스템은 제안된 바 없다.

[0003] 최근 APE 연구 흐름을 살펴보면, 현재 APE 연구들은 데이터 부족 문제를 해결하기 위한 방법에 집중하고 있다고 볼 수 있다. 이에, 본 발명에서는 특히 리소스가 부족한 언어(Low Resource Language)를 위한 APE 학습 데이터를 자동으로 생성하는 방법을 제안한다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 대한민국 공개특허 제2021-0034928호 (2021.03.31. 공개)

(특허문헌 0002) 대한민국 공개특허 제2020-0057277호 (2020.05.26. 공개)

발명의 내용

해결하려는 과제

[0005] 본 발명이 이루고자 하는 기술적인 과제는 리소스가 부족한 언어를 위한 APE(기계번역 사후교정) 학습 데이터를 생성하는 방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0006] 본 발명의 일 실시예에 따른 학습 데이터 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는, 기계번역 사후교정 모델 생성을 위한 학습 데이터 생성 방법으로서, 병렬 말뭉치를 이용하여 워드 리스트를 생성하는 단계, 병렬 말뭉치에 포함되고 제1 문장과 제2 문장을 포함하는 문장쌍에 대하여, 상기 제2 문장에 노이즈를 부가하여 노이즈가 부가된 문장을 생성하는 단계, 상기 제1 문장을 상기 학습 데이터를 구성하는 소스 문장(Source Sentence, SRC)으로 하고, 상기 노이즈가 부가된 문장을 상기 학습 데이터를 구성하는 기계번역문(Machine Translation Sentence, MT)으로 하고, 상기 제2 문장을 상기 학습 데이터를 구성하는 사후교정문(Post-edited Sentence, PE)으로 하는 상기 학습 데이터를 생성하는 단계를 포함한다.

발명의 효과

[0007] 본 발명의 실시예에 의한 기계번역 사후교정을 위한 학습 데이터 생성 장치 및 방법에 의할 경우, 병렬 말뭉치로부터 기계번역 사후교정 모델을 생성하기 위한 학습 데이터를 생성할 수 있다.

도면의 간단한 설명

[0008] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

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

도 2는 도 1에 도시된 학습 데이터 생성 장치에 의해 수행되는 학습 데이터 생성 방법을 설명하기 위한 흐름도이다.

도 3은 도 1에 도시된 노이즈 부가부에 의한 기계번역문 생성 과정을 설명하기 위한 개념도이다.

발명을 실시하기 위한 구체적인 내용

[0009] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0010] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0011] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0012] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0013] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0014] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0015] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0016] 도 1은 본 발명의 일 실시예에 따른 학습 데이터 생성 장치의 기능 블럭도이고, 도 2는 도 1에 도시된 학습 데이터 생성 장치에 의해 수행되는 학습 데이터 생성 방법을 설명하기 위한 흐름도이다.

[0017] 도 1 내지 도 3을 참조하면, 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치로 구현되는 학습 데이터 생성 장치(10)는 워드 리스트 생성부(110), 노이즈 부가부(120), 학습 데이터 생성부(130), 및 저장부(140) 중 적어도 하나 이상을 포함할 수 있다. 학습 데이터 생성 장치(10)는 병렬 말뭉치(병렬 데이터)를 이용하여 기계학습 사후교정을 위한 학습 데이터를 생성할 수 있다. 학습 데이터 생성 방법은 학습

데이터 생성 장치를 구성하는 컴퓨팅 장치에 의해 수행되는 것으로 이해될 수 있다. 즉, 학습 데이터 생성 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치의 프로세서에 의한 동작으로 이해될 수 있다.

[0018] 실시예에 따라, 학습 데이터 생성 장치(10)는 모델 생성부(미도시)를 더 포함할 수 있다. 이 경우, 학습 데이터 생성 장치(10)는 APE(기계번역 사후교정)를 위한 학습 데이터를 생성할 뿐만 아니라 생성된 학습 데이터를 이용하여 소정의 신경망 모델을 학습시킴으로써, 기계번역 사후교정 모델을 생성할 수 있다. 따라서, 학습 데이터 생성 장치(10)는 기계번역 사후교정 모델 생성 장치로 명명될 수도 있다.

[0019] 워드 리스트 생성부(110)는 복수의 단어들에 대한 워드 리스트를 생성할 수 있다. 워드 리스트를 생성하기 위해, 워드 리스트 생성부(110)는 유무선 통신망을 통해 문서들(웹문서들)을 크롤링할 수 있고, 문서들에 포함된 단어들을 추출함으로써 워드 리스트를 생성할 수 있다.

[0020] 실시예에 따라, 워드 리스트 생성부(110)는 저장부(140)에 저장되어 있는 병렬 말뭉치를 이용하여 워드 리스트를 생성할 수도 있다. 즉, 워드 리스트 생성부(110)는 병렬 말뭉치에 포함된 단어들을 추출함으로써 워드 리스트를 생성할 수 있다.

[0021] 실시예에 따라, 워드 리스트는 도메인 별로 특화된 워드 리스트를 의미할 수 있다. 즉, 특정 분야에 특화된 기계번역 사후교정 학습 데이터를 생성하기 위하여, 동일 분야에서 활용되는 단어들로만 이루어진 워드 리스트가 생성될 수 있다.

[0022] 또한, 워드 리스트는 품사별로 구분된 복수의 단어들을 포함하는 워드 리스트를 의미할 수 있다. 예컨대, 도 3을 참조하면, 워드 리스트에는 형용사 단어(JJ)로써 Big, Large, 및 Rough가 포함되고, 명사 단어(NN)로써 Score, Figure, 및 Shape가 포함되고, 동사 단어(VB)로써 Find, Consider, 및 Train이 포함될 수 있다. 워드 리스트 생성부(110)에 의해 생성된 워드 리스트는 저장부(140)에 저장될 수 있다.

[0023] 노이즈 부가부(120)는 저장부(140)에 저장되어 있는 병렬 말뭉치에 포함된 문장쌍에서 적어도 하나의 문장에 노이즈를 부가하여 노이즈가 부가된 문장을 생성할 수 있다.

[0024] 병렬 말뭉치는 복수의 문장쌍들을 포함하고, 문장쌍은 제1 언어(예컨대, 한국어)로 된 제1 문장과 제2 언어(예컨대, 영어)로 된 제2 문장을 포함할 수 있다. 여기서, 노이즈 부가부(120)는 제1 문장 또는 제2 문장에 노이즈를 부가할 수 있다. 구체적으로, 제1 언어로 된 문장을 제2 언어로 기계번역한 결과물에 대한 기계번역 사후교정을 위해서, 노이즈 부가부(120)는 제2 문장에 노이즈를 부가할 수 있다. 마찬가지로, 제2 언어로 된 문장을 제1 언어로 기게번역한 결과물에 대한 기계번역 사후교정을 위해서, 노이즈 부가부(120)는 제1 문장에 노이즈를 부가할 수 있다. 실시예에 따라, 노이즈 부가부(120)는 제1 문장과 제2 문장 모두에 노이즈를 부가할 수도 있다. 기계번역 시스템이 반드시 제1 언어를 제2 언어로 번역할 뿐만 아니라, 제2 언어를 제1 언어로 번역할 수도 있기 때문이다.

[0025] 노이즈 부가부(120)에 의해 생성된 노이즈가 부가된 문장은 기계번역 사후교정을 위한 학습 데이터 중 기계번역문(Machine Translation Sentence, MT)이 될 수 있다. 노이즈를 부가하는 구체적인 기법은 후술하기로 한다.

[0026] 학습 데이터 생성부(130)는 기계번역 사후교정을 위한 학습 데이터, 즉 기계번역 사후교정 모델을 학습시키기 위한 용도로 사용되는 학습 데이터 및/또는 테스트 데이터를 생성할 수 있다. 구체적으로, 학습 데이터 생성부(130)는 소스 문장(Source Sentence, SRC), 기계번역문(Machine Translation Sentence, MT), 및 사후 교정문(Post-edited Sentence, PE)으로 이루어진 기계번역 사후교정 트리플렛(APE triplet) 데이터를 생성할 수 있다. 여기서, SRC는 병렬 말뭉치에 포함된 제1 문장(또는 제2 문장)이고, MT는 노이즈가 부가된 문장이고, PE는 병렬 말뭉치에 포함된 제2 문장(또는 제1 문장)이 될 수 있다.

[0027] 저장부(140)에는 병렬 말뭉치(병령 데이터),워드 리스트 생성부(110)에 의해 생성되 워드 리스트, 노이즈 부가부(120)에 의해 생성된 노이즈가 부가된 기계번역문, 유사어 또는 동일어 검색을 위한 소정의 DB(예컨대, 워드넷(WordNet)), 학습 데이터 생성부(130)에 의해 생성된 학습 데이터, 모델 생성부에 의해 생성된 기계번역 사후교정 모델 등이 저장될 수 있다.

[0028] 모델 생성부는 소스 문장과 기계번역문을 입력으로 받아 사후교정문을 출력하는 기계번역 사후교정 모델을 생성할 수 있다. 이를 위해, 모델 생성부는 생성된 학습 데이터를 이용하여 소정의 신경망, 예컨대 트랜스포머(Transformer)를 학습시킴으로써 기계번역 사후교정 모델을 생성할 수 있다.

[0029] 도 3은 도 1에 도시된 노이즈 부가부에 의한 기계번역문 생성 과정을 설명하기 위한 개념도이다.

[0030] 노이즈 부가부(120)는 세가지 기법 중 적어도 하나의 기법을 이용하여 노이즈가 부가된 문장을 생성할 수 있다. 기술의 편의를 위해 노이즈가 부가되는 문장을 원본 문장(Original Sentence)이라고 한다.

[0031] 우선, 노이즈 부가부(120)는 원본 문장에 포함된 단어들 중 적어도 하나의 단어를 임의의 단어로 교체함으로써 노이즈가 부가된 문장을 생성할 수 있다. 원본 문장에 포함된 단어를 임의의 단어로 교체하는 동작이기 때문에, 추가된 노이즈를 랜덤 노이즈(Random Noise)라 명명할 수 있다. 또한, 노이즈 부가부(120)는 워드 리스트에 포함된 복수의 단어들 중 임의의 단어를 노이즈로 추가함으로써 노이즈가 부가된 문장을 생성할 수 있다.

[0032] 이때, 학습 데이터와는 완전히 동떨어진 단어가 선택되는 것을 방지하기 위해서, 병렬 말뭉치에 포함된 단어들만을 포함하는 워드 리스트를 이용할 수 있다. 워드 리스트는 병렬 말뭉치의 제1 문장들 및/또는 제2 문장들을 공백(white space) 단위로 분절하고, 분정될 단어들을 중복되지 않게 하나의 리스트로 모으는 방법을 통해 구축될 수 있다.

[0033] 도 3의 예시에서, 노이즈 부가부(120)는 원본 문장(Can I help you?)에 포함된 'help'를 워드 리스트에 포함된 'big'으로 교체함으로써 노이즈가 부가된 문장(Can I big you?)을 생성할 수 있다. 랜덤 노이즈에서는 학습 데이터 내의 단어들의 시멘틱(semantic) 정보나 신택틱(syntactic) 정보에 대한 고려 없이 단어를 추출하는 방식으로 노이즈를 생성하기 때문에, 기존 문장에서 동사였던 단어가 형용사로 대체되는 경우가 발생할 수 있다.

[0034] 다음으로, 노이즈 부가부(120)는 원본 문장에 포함된 단어들 중 적어도 하나의 단어를 동일한 품사의 단어로 교체함으로써 노이즈가 부가된 문장을 생성할 수 있다. 원본 문장에 포함된 단어를 품사가 동일한 단어로 교체하는 동작이기 때문에, 추가된 노이즈를 형태소 레벨 노이즈(Morpheme Level Noise)라 명명할 수 있다. 또한, 노이즈 부가부(120)는 워드 리스트에 포함된 복수의 단어들 중 품사가 동일한 단어를 노이즈로 추가함으로써 노이즈가 부가된 문장을 생성할 수 있다.

[0035] 실시예에 따라, 형태소 레벨 노이즈에서는 POS(Part Of Speech) 태깅을 기반으로 한 노이즈가 적용될 수 있다. POS 태깅을 통해 병렬 말뭉치 내의 문장들을 대상으로 각 단어의 품사를 분석하고, 각 POS 태그들에 해당하는 워드 리스트를 구축할 수 있다. 이를 바탕으로 기존 문장 내의 단어를 다른 단어로 교체할 때, 해당 워드 리스트를 참고하여 동일한 태그로 분류되는 단어들 중 하나로 교체할 수 있다. 이는 문장에 노이즈를 부가할 때, 단순하게 랜덤 노이즈 방식으로 부여하는 것이 아닌, 전통적인 NLP 파이프라인(traditional Natural Language Processing pipeline)에 따라 하나의 기준을 세우고 노이즈를 부여했다는 점에서 랜덤 노이즈를 활용했을 때보다 더 높은 성능을 기대할 수 있다.

[0036] 도 3의 예시에서, 노이즈 부가부(120)는 원본 문장(Can I help you?)에 포함된 'help'를 워드 리스트에 포함된 단어들 중 품사(동사, VB)가 동일한 'find'로 교체함으로써 노이즈가 부가된 문장(Can I find you?)을 생성할 수 있다.

[0037] 마지막으로, 노이즈 부가부(120)는 원본 원본 문장에 포함된 단어들 중 적어도 하나의 단어를 동일하거나 유사한 의미를 가진 단어(동의어 또는 유의어)로 교체함으로써 노이즈가 부가된 문장을 생성할 수 있다. 원본 문장에 포함된 단어를 의미가 동일 또는 유사한 단어로 교체하는 동작이기 때문에, 추가된 노이즈를 시멘틱 레벨 노이즈(Semantic Level Noise)라 명명할 수 있다.

[0038] 또한, 노이즈 부가부(120)는 소정의 DB에서 대체할 단어의 동일어 또는 유사어를 검색하고, 검색된 단어들 중 어느 하나를 노이즈로 부가함으로써 노이즈가 부가된 문장을 생성할 수 있다. 예시적으로, 노이즈 부가부(120)는 워드넷(WordNet) DB에서 대체할 단어의 동일어 또는 유사어를 검색할 수 있다.

[0039] 시멘틱 레벨 노이즈에서는 소정의 DB(WordNet, 워드넷)를 활용한 노이즈가 적용될 수 있다. 이는 타겟 문장 내 토큰에 대체 노이즈(substitution noise)를 적용하는 과정에서, 해당 토큰에 대한 워드넷 정보를 검색(retrieve)하여 해당 단어의 유의어(synonym) 중 하나로 대체하는 방법론이다. 본 방법론은 기계번역 결과 내에서, 휴먼 레벨(human level)로 수정해야 하는 오류를 최대한 반영되도록 노이즈를 부가하는 방법이라 말할 수 있다. 휴먼 레벨에서 교정되어야 하는 대표적인 오류로는, 단어 자체의 의미는 같으나, 문맥이나 도메인 내에서 미묘한 차이를 갖는 단어에 대한 수정이 있다. 예를 들어, 'help'와 'assist'는, 단어 자체로는 'give help or assistance; be of service'라는 동일한 의미를 가지고 있지만, 문맥이나 상황을 고려하여 다르게 적용해야 하는 경우가 존재한다. 해당 정보를 워드넷을 통해 노이즈 단계에서 부여함으로써 휴먼 레벨에서 감지할 수 있는 오류에 대한 교정을 APE 모델이 학습할 수 있다.

[0040] 도 3의 예시에서, 노이즈 부가부(120)는 원본 문장(Can I help you?)에 포함된 'help'를 유사한 의미를 갖는

'assist'로 교체함으로써 노이즈가 부가된 문장(Can I assist you?)을 생성할 수 있다.

[0041] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0042] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0043] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0044] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0045] 10 : 학습 데이터 생성 장치

110 : 워드 리스트 생성부

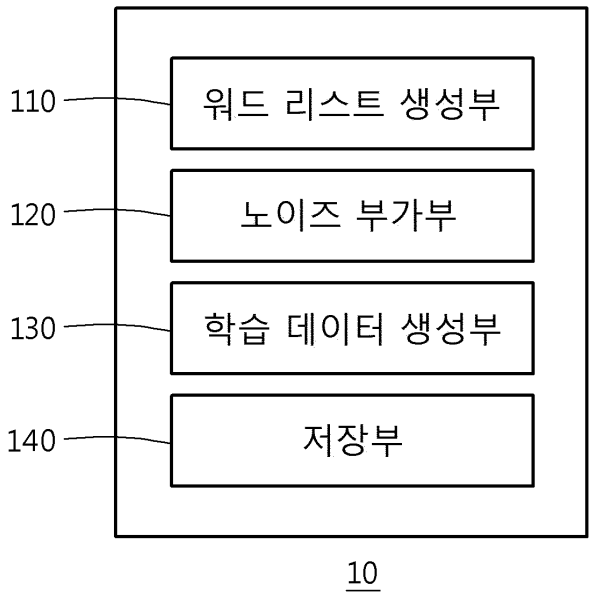
120 : 노이즈 부가부

130 : 학습 데이터 생성부

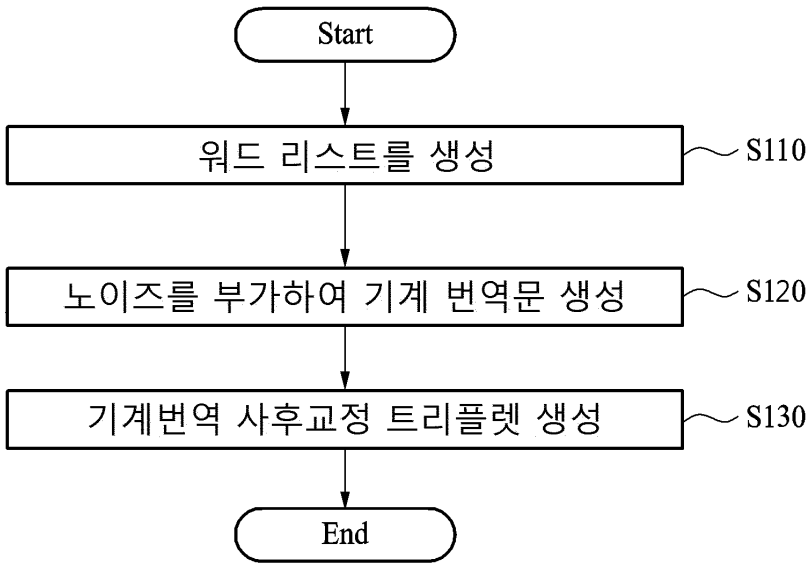
140 : 저장부

도면

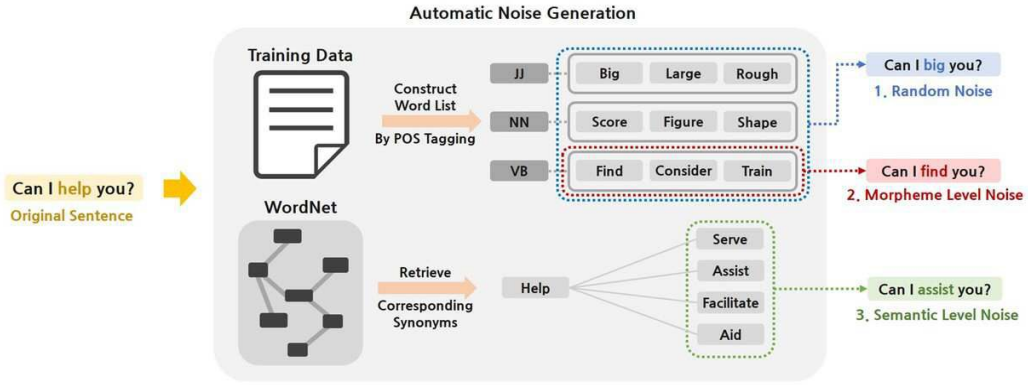
도면1


110
워드 리스트 생성부
120
노이즈 부가부
130
학습 데이터 생성부
140
저장부
10

도면2


Start
워드 리스트를 생성
S110
노이즈를 부가하여 기계 번역문 생성
S120
기계번역 사후교정 트리플렛 생성
S130
End

도면3


Automatic Noise Generation
Training Data
Can I help you?
Original Sentence
Construct
Word List
By POS Tagging
JJ
NN
VB
Big
Large
Rough
Score
Figure
Shape
Find
Consider
Train
Can I big you?
1. Random Noise
Can I find you?
2. Morpheme Level Noise
WordNet
Retrieve
Corresponding
Synonyms
Help
Serve
Assist
Facilitate
Aid
Can I assist you?
3. Semantic Level Noise