(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0045688
(43) 공개일자 2023년04월05일

(51) 국제특허분류(Int. Cl.)
G06F 40/58 (2020.01) *G06F 40/205* (2020.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 40/58 (2020.01)
G06F 40/205 (2020.01)
(21) 출원번호 10-2021-0127438
(22) 출원일자 2021년09월27일
심사청구일자 2021년09월27일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
박찬준
서울특별시 송파구 중대로 24, 306동 1303호 (문정동, 올림픽훼밀리타운)
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스애기능생활관 311호 NLP&AI연구실
(74) 대리인
김홍석

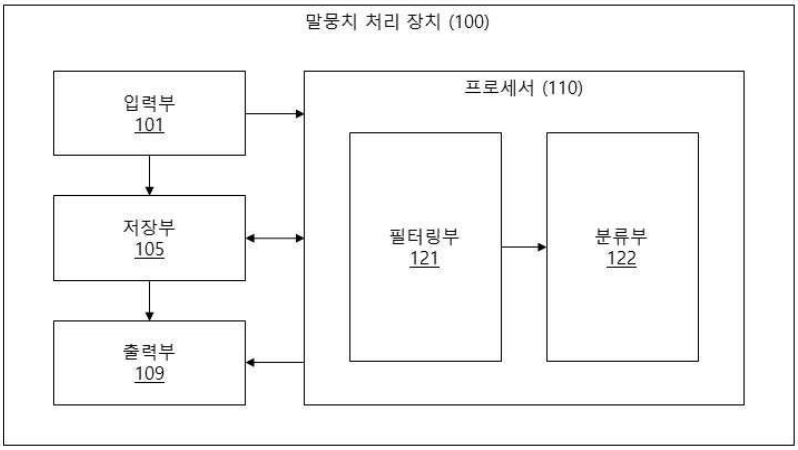
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **말뭉치 처리 장치, 기계 번역 장치 및 방법**

(57) 요 약

말뭉치 처리 장치, 기계 번역 장치 및 방법에 관한 것으로, 말뭉치 처리 장치는 적어도 하나의 말뭉치가 저장되는 저장부 및 상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하고, 상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하는 프로세서를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2023.01)

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 하나의 말뭉치가 저장되는 저장부; 및

상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하고, 상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하는 프로세서;를 포함하는 말뭉치 처리 장치.

청구항 2

제1항에 있어서,

상기 프로세서는, 상기 적어도 하나의 말뭉치 내의 서로 쌍을 이루는 원 문장 및 목표 문장을 비교하고, 상기 원 문장 및 상기 목표 문장이 동일하면 상기 원 문장 및 상기 목표 문장 중 적어도 하나를 상기 말뭉치에서 제거하거나,

상기 원 문장 및 상기 목표 문장 내의 모든 문자 중에서 일정 비율 이상의 문자가 미리 정의된 문자가 아니라면, 상기 원 문장 및 상기 목표 문장을 제거하거나,

어절의 길이가 미리 정의된 길이가 미리 정의된 길이보다 길거나, 문장 내의 무의미한 기호의 비율이 일정 비율보다 크거나, 문장 내의 특수 기호의 개수가 미리 정의된 개수보다 크거나 또는 상기 원 문장 및 상기 목표 문장 간의 대응 관계에 오류가 존재하는 경우, 상기 원 문장 및 상기 목표 문장을 제거함으로써 상기 필터링된 말뭉치를 획득하는 말뭉치 처리 장치.

청구항 3

제1항에 있어서,

상기 프로세서는, 사전훈련-미세조정 접근(PFA: Pretrain-Finetuning Approach) 기법을 이용하여 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하는 말뭉치 처리 장치.

청구항 4

적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하는 필터링부; 및

상기 필터링된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시키거나 또는 훈련된 인공신경망 기계 번역 모델을 이용하여 원 문장에 대응하는 목표 문장을 획득하는 학습 처리부;를 포함하는 기계 번역 장치.

청구항 5

제4항에 있어서,

상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하여 필터링 및 분류된 말뭉치를 획득하는 분류부;를 더 포함하는 기계 번역 장치.

청구항 6

제5항에 있어서,

상기 학습 처리부는, 미리 정의된 설정 또는 사용자의 조작에 따라 적어도 하나의 도메인을 선택하고, 상기 선택된 적어도 하나의 도메인을 기반으로 상기 필터링 및 분류된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시켜 도메인에 적응 모델을 획득하는 기계 번역 장치.

청구항 7

적어도 하나의 말뭉치를 획득하는 단계;

상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하는 단계; 및

상기 필터링된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시키는 단계;를 포함하는 기계 번역 방법.

청구항 8

제7항에 있어서,

상기 필터링된 말뭉치를 이용하여 훈련된 인공신경망 기계 번역 모델을 이용하여 원 문장에 대응하는 목표 문장을 획득하는 단계;를 더 포함하는 기계 번역 방법.

청구항 9

제7항에 있어서,

상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하여 필터링 및 분류된 말뭉치를 획득하는 단계;를 더 포함하는 기계 번역 방법.

청구항 10

제9항에 있어서,

미리 정의된 설정 또는 사용자의 조작에 따라 적어도 하나의 도메인을 선택하는 단계;를 더 포함하는 기계 번역 방법.

청구항 11

제10항에 있어서,

상기 필터링된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시키는 단계는,

상기 선택된 적어도 하나의 도메인을 기반으로 상기 필터링 및 분류된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시켜 도메인에 적응 모델을 획득하는 단계;를 포함하는 기계 번역 방법.

청구항 12

제7항에 있어서,

상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하는 단계는,

상기 적어도 하나의 말뭉치 내의 서로 쌍을 이루는 원 문장 및 목표 문장을 비교하고, 상기 원 문장 및 상기 목표 문장이 동일하면 상기 원 문장 및 상기 목표 문장 중 적어도 하나를 상기 말뭉치에서 제거하는 단계;

상기 원 문장 및 상기 목표 문장 내의 모든 문자 중에서 일정 비율 이상의 문자가 미리 정의된 문자가 아니라면, 상기 원 문장 및 상기 목표 문장을 제거하는 단계; 및

어절의 길이가 미리 정의된 길이가 미리 정의된 길이보다 길거나, 문장 내의 무의미한 기호의 비율이 일정 비율보다 크거나, 문장 내의 특수 기호의 개수가 미리 정의된 개수보다 크거나, 또는 상기 원 문장 및 상기 목표 문장 간의 대응 관계에 오류가 존재하는 경우, 상기 원 문장 및 상기 목표 문장을 제거하는 단계; 중 적어도 하나를 포함하는 기계 번역 방법.

발명의 설명

기 술 분 야

[0001] 말뭉치 처리 장치, 기계 번역 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 기계 번역(Machine Translation)이란, 컴퓨터 등의 정보처리장치가 원 문장(source sentence)를 목표 문장(target sentence)로 자동으로 번역하는 소프트웨어 또는 시스템을 의미한다. 이러한 기계 번역은 규칙, 통계 또는 인공 신경망을 이용하여 구현되어 왔다. 구체적으로 규칙 기반 기계 번역은, 알려진 언어 규칙을 기반으로 형태소 분석, 구문 분석 또는 의미 분석 등의 전통적인 자연언어 처리 방법에 따라 번역을 수행하는 방법이고, 통계 기반 기계 번역은, 대규모의 병렬 말뭉치(Parallel Corpus, 2개 언어 이상의 번역된 문장 등을 조합하여 제작된 말뭉치로, 서로 대응하는 문장 또는 문단을 포함할 수 있음)로부터 단어들 간의 정렬 정보나 빈도수 등을 획득하여 통계 정보를 학습시키고, 학습된 통계 정보를 이용하여 번역을 수행하는 방법이다. 인공 신경망 기계 번역(NMT: Neural Machine Translation)은 딥러닝 등의 인공 신경망을 이용하여 기계 번역을 수행하는 방법으로, 하위 모듈(형태소 분석기나 구문 분석기 등)의 성능에 크게 구애되지 않고 우수한 품질의 번역을 수행할 수 있는 장점이 있다. 최근에는 신경망 기계 번역의 성능 향상에 따라서 다수의 기계 번역 서비스는 종래의 규칙 기반 또는 통계 기반 번역 대신에 신경망 기계 번역을 채용하여 제작되고 있다. 그러나, 최근 소개된 인공 신경망 기반 기계 번역 방법들은, 모델의 사이즈가 매우 크거나, 대량의 처리할 파라미터를 요구하기 때문에, 리소스 부족이나 학습 속도 저하 등과 같은 여러 한계점을 드러내고 있었다.

발명의 내용

해결하려는 과제

[0003] 인공신경망 기계 번역을 위한 고품질의 학습용 데이터를 자동으로 구축, 정제 또는 관리할 수 있는 말뭉치 처리 장치, 기계 번역 장치 및 방법을 제공하는 것을 해결하고자 하는 과제로 한다.

과제의 해결 수단

[0004] 상술한 과제를 해결하기 위하여 말뭉치 처리 장치, 기계 번역 장치 및 방법이 제공된다.

[0005] 말뭉치 처리 장치는, 적어도 하나의 말뭉치가 저장되는 저장부 및 상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하고, 상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하는 프로세서를 포함할 수 있다.

[0006] 상기 프로세서는, 상기 적어도 하나의 말뭉치 내의 서로 쌍을 이루는 원 문장 및 목표 문장을 비교하고, 상기 원 문장 및 상기 목표 문장이 동일하면 상기 원 문장 및 상기 목표 문장 중 적어도 하나를 상기 말뭉치에서 제거하거나, 상기 원 문장 및 상기 목표 문장 내의 모든 문자 중에서 일정 비율 이상의 문자가 미리 정의된 문자가 아니라면, 상기 원 문장 및 상기 목표 문장을 제거하거나, 어절의 길이가 미리 정의된 길이가 미리 정의된 길이보다 길거나, 문장 내의 무의미한 기호의 비율이 일정 비율보다 크거나, 문장 내의 특수 기호의 개수가 미리 정의된 개수보다 크거나, 또는 상기 원 문장 및 상기 목표 문장 간의 대응 관계에 오류가 존재하는 경우, 상기 원 문장 및 상기 목표 문장을 제거함으로써 상기 필터링된 말뭉치를 획득할 수도 있다.

[0007] 상기 프로세서는, 사전훈련-미세조정 접근(PFA: Pretrain-Finetuning Approach) 기법을 이용하여 필터링된 말뭉치를 적어도 하나의 도메인으로 분류할 수도 있다.

[0008] 기계 번역 장치는, 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하는 필터링부 및 상기 필터링된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시키거나 또는 훈련된 인공신경망 기계 번역 모델을 이용하여 원 문장에 대응하는 목표 문장을 획득하는 학습 처리부를 포함할 수 있다.

[0009] 기계 번역 장치는, 상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하여 필터링 및 분류된 말뭉치를 획득하는 분류부를 더 포함하는 것도 가능하다.

[0010] 상기 학습 처리부는, 미리 정의된 설정 또는 사용자의 조작에 따라 적어도 하나의 도메인을 선택하고, 상기 선택된 적어도 하나의 도메인을 기반으로 상기 필터링 및 분류된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시켜 도메인에 적응 모델을 획득할 수도 있다.

[0011] 기계 번역 방법은, 적어도 하나의 말뭉치를 획득하는 단계, 상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링된 말뭉치를 획득하는 단계 및 상기 필터링된 말뭉치를 이용하여 인공신경망 기계 번역

모델을 훈련시키는 단계를 포함할 수도 있다.

[0012] 기계 번역 방법은, 상기 필터링된 말뭉치를 이용하여 훈련된 인공신경망 기계 번역 모델을 이용하여 원 문장에 대응하는 목표 문장을 획득하는 단계를 더 포함하는 것도 가능하다.

[0013] 기계 번역 방법은, 상기 필터링된 말뭉치를 적어도 하나의 도메인으로 분류하여 필터링 및 분류된 말뭉치를 획득하는 단계를 더 포함할 수도 있다.

[0014] 기계 번역 방법은, 미리 정의된 설정 또는 사용자의 조작에 따라 적어도 하나의 도메인을 선택하는 단계를 더 포함할 수도 있다.

[0015] 상기 필터링된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시키는 단계는, 상기 선택된 적어도 하나의 도메인을 기반으로 상기 필터링 및 분류된 말뭉치를 이용하여 인공신경망 기계 번역 모델을 훈련시켜 도메인에 적응 모델을 획득하는 단계를 포함하는 것도 가능하다.

[0016] 상기 적어도 하나의 말뭉치에 대해 필터링을 수행하여 상기 적어도 하나의 말뭉치의 적어도 하나의 문장을 제거하거나 또는 상기 적어도 하나의 말뭉치 중 적어도 하나를 제거하여 필터링 된 말뭉치를 획득하는 단계는, 상기 적어도 하나의 말뭉치 내의 서로 쌍을 이루는 원 문장 및 목표 문장을 비교하고, 상기 원 문장 및 상기 목표 문장이 동일하면 상기 원 문장 및 상기 목표 문장 중 적어도 하나를 상기 말뭉치에서 제거하는 단계, 상기 원 문장 및 상기 목표 문장 내의 모든 문자 중에서 일정 비율 이상의 문자가 미리 정의된 문자가 아니라면, 상기 원 문장 및 상기 목표 문장을 제거하는 단계 및 어절의 길이가 미리 정의된 길이가 미리 정의된 길이보다 길거나, 문장 내의 무의미한 기호의 비율이 일정 비율보다 크거나, 문장 내의 특수 기호의 개수가 미리 정의된 개수보다 크거나, 또는 상기 원 문장 및 상기 목표 문장 간의 대응 관계에 오류가 존재하는 경우, 상기 원 문장 및 상기 목표 문장을 제거하는 단계 중 적어도 하나를 포함할 수도 있다.

발명의 효과

[0017] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 인공신경망 기계 번역을 위한 고품질의 학습용 데이터를 자동으로 구축, 정제 또는 관리할 수 있게 되는 효과를 얻을 수 있다.

[0018] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 보다 효과적이면서 효율적으로 고품질의 병렬 말뭉치를 획득할 수 있게 되어 인공신경망 기계 번역의 성능 향상 및 경제성 증진을 도모할 수 있는 장점도 얻을 수 있다.

[0019] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 다양한 분야의 말뭉치를 각 도메인 별로 관리할 수 있게 되어 기계 번역의 품질을 개선할 수 있는 긍정적인 효과도 얻을 수 있다.

[0020] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 자동으로 병렬 말뭉치의 획득 및 분류, 모델 전처리, 기계 번역 모델의 훈련 또는 이의 배포 등이 수행될 수 있으므로, 인공 신경망 기계 번역에 대한 지식이 없거나 부족한 비전문가들도 보다 용이하고 편리하게 자신이 원하는 바에 따라 인공 신경망 기계 번역기를 제작할 수 있는 장점도 얻을 수 있다.

[0021] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 사용자 특유의 인공 신경망 기계 번역기를 제작할 수 있게 되는 장점도 얻을 수 있다.

[0022] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 주어진 단어나 문장에 대해 시간 대비 고품질 번역 결과를 획득할 수 있어, 다국어 등의 번역 서비스의 품질 향상을 도모할 수 있는 효과도 획득할 수 있다.

[0023] 상술한 말뭉치 처리 장치, 기계 번역 장치 및 방법에 의하면, 번역에 필요한 학습 데이터를 용이하게 확보할 수 있을 뿐만 아니라, 이들의 구축에 소요되는 노동력, 시간 및 비용을 절감할 수 있게 되는 장점도 얻을 수 있다.

도면의 간단한 설명

[0024] 도 1은 말뭉치 처리 장치의 일 실시예에 대한 블록도이다.

도 2는 기계 번역 장치의 일 실시예에 대한 블록도이다.

도 3은 프로세서의 일 실시예에 대한 블록도이다.

도 4는 기계 번역 방법의 일 실시예에 대한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0025] 이하 명세서 전체에서 동일 참조 부호는 특별한 사정이 없는 한 동일 구성요소를 지칭한다. 이하에서 사용되는 '부'가 부가된 용어는, 소프트웨어 및/또는 하드웨어로 구현될 수 있으며, 실시예에 따라 하나의 '부'가 하나의 물리적 또는 논리적 부품으로 구현되거나, 복수의 '부'가 하나의 물리적 또는 논리적 부품으로 구현되거나, 하나의 '부'가 복수의 물리적 또는 논리적 부품들로 구현되는 것도 가능하다. 명세서 전체에서 어떤 부분이 다른 부분과 연결되어 있다고 할 때, 이는 어떤 부분과 다른 부분이 상호 간에 물리적으로 연결되었음을 의미할 수도 있고, 및/또는 전기적으로 연결되었음을 의미할 수도 있다. 또한, 어떤 부분이 다른 부분을 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 부분 이외의 또 다른 부분을 제외하는 것이 아니며, 설계자의 선택에 따라서 또 다른 부분을 더 포함할 수 있음을 의미한다. 또한 단수의 표현은 문맥상 명백하게 예외가 있지 않는 한, 복수의 표현을 포함할 수 있다.

[0026] 이하 도 1 내지 도 3을 참조하여 말뭉치 처리 장치 및 기계 번역 장치 각각의 일 실시예를 설명하도록 한다.

[0027] 도 1은 말뭉치 처리 장치의 일 실시예에 대한 블록도이다.

[0028] 도 1에 도시된 바와 같이, 말뭉치 처리 장치(100)는, 일 실시예에 있어서, 프로세서(110)를 포함할 수 있으며, 입력부(101), 저장부(105) 및 출력부(109) 중 적어도 하나를 더 포함할 수 있다. 입력부(101), 저장부(105), 출력부(109) 및 프로세서(110) 중 적어도 둘은 케이블, 회로 또는 무선 통신 네트워크를 통해 일방으로 또는 상호 간에 데이터나 명령, 지시 등을 전달할 수 있게 마련된다.

[0029] 입력부(101)는, 말뭉치 처리 장치(100)의 동작에 필요한 데이터나 명령 등을 입력 받을 수 있다. 예를 들어, 입력부(101)는 사용자의 조작에 따라 또는 적어도 하나의 다른 외부의 장치(들)(일례로 메모리 장치나, 데스크톱 컴퓨터나, 스마트폰이나, 서버용 하드웨어 장치 등)으로부터 하나 이상의 문장이나 단어 등을 입력 받거나, 하나 이상의 문장 또는 단어 등을 포함하는 말뭉치를 입력 받을 수 있다. 여기서, 말뭉치는 병렬 말뭉치를 포함할 수 있다. 이 경우, 입력부(101)는 다른 웹 사이트의 크롤링 결과에 따라 획득된 학습용 데이터(즉, 문장 또는 이들 문장을 포함하는 말뭉치)를 입력 받을 수도 있다. 또한, 입력부(101)는, 말뭉치 처리 등에 관한 적어도 하나의 명령 등을 사용자나 다른 장치 등으로부터 입력 받을 수 있고, 말뭉치 처리 장치(100)의 동작을 위한 적어도 하나의 프로그램(앱, 애플리케이션 또는 소프트웨어 등으로 지칭 가능함)을 수신할 수도 있다. 입력부(101)가 수신한 데이터, 명령 또는 프로그램 등은 저장부(105)나 프로세서(110) 등으로 전달 가능하다. 입력부(101)는, 말뭉치 처리 장치(100)와 일체형으로 마련된 것일 수도 있고, 또는 물리적으로 분리 가능하게 마련된 것일 수도 있으며, 예를 들어, 키보드, 마우스, 태블릿, 터치 스크린, 터치 패드, 트랙 볼, 트랙패드, 스캐너 장치, 영상 촬영 모듈, 초음파 스캐너, 동작 감지 센서, 진동 센서, 수광 센서, 감압 센서, 근접 센서 및/또는 마이크로 폰 등을 포함할 수도 있고, 외부의 다른 장치로부터 데이터 등의 수신이 가능한 데이터 입출력 단자나, 외부의 다른 장치와 유무선 통신 네트워크를 통해 연결되는 통신 모듈(일례로 랜 카드, 근거리 통신 모듈 또는 이동 통신 모듈 등) 등을 포함할 수도 있다.

[0030] 저장부(105)는 말뭉치 처리 장치(100)의 동작을 위한 데이터나 프로그램 등을 일시적 또는 비일시적으로 저장할 수 있다. 예를 들어, 저장부(105)는 입력부(101) 등을 통해 획득한 말뭉치를 저장할 수도 있고, 프로세서(110)에 의해 필터링된 말뭉치를 저장할 수도 있으며, 필터링된 말뭉치에 부가된 분류 정보(도메인)를 저장할 수도 있다. 저장부(105)는, 주기억장치 및 보조기억장치 중 적어도 하나를 이용하여 구현 가능하다. 여기서, 주기억장치는 롬(ROM) 및/또는 램(RAM)과 같은 반도체 저장 매체 등을 포함할 수 있고, 보조기억장치는, 플래시 메모리 장치(일례로 솔리드 스테이트 드라이브(SSD: Solid State Drive) 등), SD(Secure Digital) 카드, 하드 디스크 드라이브, 자기 드럼, 콤팩트 디스크, 디브이디 또는 레이저 디스크 등을 포함할 수 있다.

[0031] 출력부(109)는, 저장부(105)에 저장된 데이터나, 프로세서(110)의 처리 결과 등을 외부로 출력할 수 있다. 예를 들어, 출력부(109)는 저장된 말뭉치나, 필터링 되거나 및/또는 분류된 말뭉치나, 프로세서(110)의 처리 과정 등을 사용자 등에게 시각적 또는 청각적으로 출력하거나 또는 외부의 다른 장치로 전달할 수 있다. 출력부(109)는, 필요에 따라, 말뭉치 처리 장치(100)와 일체형으로 마련된 것일 수도 있고, 또는 물리적으로 분리 가능하게 마련된 것일 수도 있으며, 예를 들어, 디스플레이, 프린터 장치, 스피커 장치, 영상 출력 단자, 데이터 입출력 단자 및/또는 통신 모듈 등을 포함할 수 있다.

[0032] 프로세서(110)는, 입력부(101)를 통해 입력되거나 저장부(105)에 저장된 말뭉치 처리에 대한 소정의 동작을 순차적으로 또는 동시에 수행하여 처리된 말뭉치를 획득할 수 있다. 예를 들어, 프로세서(110)는 하나 이상의 말뭉치, 일례로 하나 이상의 병렬 말뭉치를 획득하고, 획득한 하나 이상의 말뭉치나 도메인 별로 분류된 하나 이

상의 말뭉치에 대해 필터링을 수행하여 필터링된 말뭉치를 획득하거나, 및/또는 하나 이상의 입력 또는 저장된 말뭉치나 하나 이상의 필터링 처리된 말뭉치를 도메인 별로 분류할 수도 있다.

[0033] 도 1에 도시된 일 실시예에 따르면, 프로세서(110)는, 필터링부(121) 및 분류부(122) 중 적어도 하나를 포함할 수 있다. 만약 프로세서(110)가 필터링부(121) 및 분류부(122) 모두를 포함한다면, 필터링부(121) 및 분류부(122)는 설계에 따라 논리적으로 구분되는 것일 수도 있고 또는 물리적으로 구분되는 것일 수도 있다. 논리적으로 구분되는 경우, 필터링부(121) 및 분류부(122)는 하나의 반도체 칩 등을 이용하여 구현될 수도 있다. 물리적으로 구분되는 경우 필터링부(121) 및 분류부(122)는 각각 서로 상이한 둘 이상의 반도체 칩 등을 이용하여 구현될 수도 있다.

[0034] 필터링부(121)는 주어진 말뭉치 내에서 적절한 문장 또는 단어를 선별하고, 불필요한 문장 또는 단어를 제거하거나, 주어진 다수의 말뭉치 중 일부의 말뭉치를 제거함으로써, 기계 번역 모델을 훈련시키기 위한 최적의 말뭉치를 획득하도록 할 수 있다. 이에 따라 좋은 품질의 인공신경망 기계 번역 모델을 훈련 및 획득하기 위한 양질의 말뭉치가 구축될 수 있게 된다. 예를 들어, 필터링부(121)는 병렬 말뭉치 내의 서로 쌍을 이루는 문장들(즉, 원 문장과 목표 문장)을 비교하고, 비교 결과 원 문장과 목표 문장이 동일하면, 이들 중 어느 하나를 제거하거나 양자를 모두 제거함으로써 필터링을 수행할 수도 있다. 다른 예를 들어, 필터링부(121)는 문장(일례로 원 문장 및 목표 문장 중 적어도 하나를 포함할 수 있음) 내의 모든 문자 중에서 일정 비율(일례로 50%) 이상의 문자가 미리 정의된 문자(예를 들어, 알파벳이나 한글 등)가 아니라면, 원 문장과 목표 문장 중 어느 하나를 제거하거나 또는 양자 모두를 제거하도록 마련된 것일 수도 있다. 또 다른 예를 들어, 필터링부(121)는 어절의 길이가 미리 정의된 길이(예를 들어 500 어절)보다 긴 문장이나, 문장의 일정 비율(일례로 30%) 이상이 무의미한 기호(예를 들어, 화이트 스페이스나 탭 등)로 이루어진 문장이나, 특수 기호가 일정 개수(일례로 9개) 이상 포함되어 있는 문장이나, 및/또는 원 문장과 목표 문장 사이의 대응 관계에 오류가 존재하는 문장에 대해서도 원 문장과 목표 문장 중 적어도 하나를 제거하도록 마련되는 것도 가능하다. 이외에도 필터링부(121)는 설계자가 미리 정한 하나 이상의 조건에 따라서 병렬 말뭉치 내의 서로 대응하는 문장, 즉, 원 문장 및 목표 문장 중 적어도 하나를 제거할 수도 있다. 이와 같이 필터링된 말뭉치를 이용하는 경우, 입력 데이터의 품질이 향상되므로 더 우수한 학습 모델의 훈련 결과를 획득할 수 있다. 예를 들어, 필터링된 말뭉치를 이용하면, 필터링 처리되지 않은 말뭉치를 이용한 경우보다 기계 번역 모델에 대한 다중 언어 평가 언더스터디 점수(BLEU: Bilingual Evaluation Understudy) score)가 더 높은 경향을 보인다.

[0035] 분류부(122)는 말뭉치를 특정한 하나 또는 둘 이상의 도메인으로 분류할 수 있다. 일반적인 기계 번역을 이용하는 경우, 특정 도메인(예를 들어, 논문, 보안, 특허, 의학, 약학 또는 스포츠 등)의 특유의 용어나 언어의 사용 예를 충분히 반영하지 못해, 번역의 품질이 낮은 경우가 많다. 분류부(122)는 이와 같이 각각의 도메인마다의 번역 품질 개선을 위해 입력부(101) 등을 통해 획득된 말뭉치나 필터링부(121)에 의해 필터링된 말뭉치를 도메인 별로 분류하고, 이를 기록함으로써 특정 도메인에 특화된 병렬 말뭉치를 획득할 수 있다. 일 실시예에 있어서, 분류부(122)는, 사전훈련-미세조정 접근(PFA: Pretrain-Finetuning Approach) 기법 등을 이용하여 말뭉치에 대한 도메인 분류를 수행할 수도 있다.

[0036] 필터링부(121)에 의해 필터링 되거나 및/또는 분류부(122)에 의해 특정 도메인으로 분류된 말뭉치(즉, 필터링 및/또는 분류된 말뭉치)는, 일시적으로 또는 비일시적으로 저장부(105)에 저장될 수 있고, 필요에 따라 출력부(109)를 통해 외부로 시각적 또는 청각적으로 출력될 수 있으며, 기계 번역 모델의 학습을 위해 마련된 다른 정보처리장치(예를 들어, 외장 메모리 장치나, 기계 번역 장치(도 2의 200))로 유무선통신네트워크 등을 통해 전달될 수도 있다.

[0037] 일 실시예에 의하면, 상술한 프로세서(110)는, 중앙 처리 장치(CPU: Central Processing Unit), 마이크로 컨트롤러 유닛(MCU: Micro Controller Unit), 애플리케이션 프로세서(AP: Application Processor), 전자 제어 유닛(ECU: Electronic Controlling Unit), 기본보드 관리 컨트롤러(BMC: Baseboard Management Controller), 마이컴(Micom: Micro Processor) 및/또는 이외 각종 연산 및 제어 처리를 수행할 수 있는 적어도 하나의 전자 장치 등을 이용하여 구현될 수 있다. 이들 처리 또는 제어 장치는, 예를 들어, 하나 또는 둘 이상의 반도체 칩, 회로 또는 관련 부품 등을 단독으로 이용하거나 조합 이용하여 제작 가능하다.

[0038] 도 2는 기계 번역 장치의 일 실시예에 대한 블록도이고, 도 3은 프로세서의 일 실시예에 대한 블록도이다.

[0039] 도 2에 도시된 바와 같이 기계 번역 장치(200)는 프로세서(210)를 포함할 수 있으며, 필요에 따라 입력부(201), 저장부(205) 및 출력부(209) 중 적어도 하나를 포함할 수 있다.

[0040] 입력부(201)는, 기계 번역 장치(200)의 동작에 필요한 명령, 데이터 또는 프로그램 등을 수신할 수 있으며, 예를 들어, 병렬 말뭉치 등과 같은 말뭉치를 사용자 또는 다른 장치로부터 입력 받을 수도 있고, 학습 또는 번역의 개시 또는 종료 등을 위한 명령을 입력 받을 수도 있다.

[0041] 저장부(205)는 기계 번역 장치(200)의 동작을 위한 데이터나 프로그램 등을 일시적 또는 비일시적으로 저장할 수 있으며, 예를 들어, 입력부(201)를 통해 입력된 말뭉치나 프로세서(210)에 의해 필터링 되거나 특정 도메인으로 분류된 말뭉치를 저장하거나, 번역할 문장(즉, 원 문장) 및 번역된 문장(즉, 목표 문장) 중 적어도 하나를 저장하거나, 및/또는 훈련되거나 또는 훈련되지 않은 인공신경망 기계 번역 모델(예를 들어, 트랜스포머(transformer) 등)을 저장할 수도 있다.

[0042] 출력부(209)는, 저장부(205)에 저장되거나 프로세서(210)의 처리에 따라 획득된 데이터 등을 외부로 출력할 수 있다. 예를 들어, 출력부(209)는, 프로세서(210)가 저장되거나 처리된 말뭉치나, 훈련된 인공신경망 기계 번역 모델이나, 원 문장에 대응하는 목표 문장 등을 사용자 등에게 시각적 또는 청각적으로 출력하거나 또는 외부의 다른 장치로 전달할 수 있다.

[0043] 입력부(201), 저장부(205) 및 출력부(209)는 말뭉치 처리 장치(100)의 입력부(101), 저장부(105) 및 출력부(109)와 동일하거나 일부 변형된 형태로 구현 가능하므로, 이하 자세한 설명은 생략하도록 한다.

[0044] 프로세서(210)는, 기계 번역을 위해 필요한 동작을 수행할 수 있으며, 구체적으로 예를 들어, 입력부(201)를 통해 입력되거나 저장부(205)에 저장된 말뭉치에 대한 적어도 하나의 처리를 수행하거나, 처리된 말뭉치를 이용하여 인공신경망 기계 번역 모델의 훈련을 수행하거나, 및/또는 훈련된 인공신경망 기계 번역 모델을 이용하여 원 문장에 대응하는 목표 문장을 생성하는 동작 등을 동시에 또는 순차적으로 수행할 수도 있다. 실시예에 따라서, 프로세서(210)는 이들 동작 중 일부만을 수행할 수도 있고, 이들 동작 모두를 수행할 수도 있다. 프로세서(210)는, 말뭉치 처리 동작, 인공신경망 기계 번역 모델의 훈련 동작 및/또는 목표 문장의 획득 동작 등을 수행하기 위해 저장부(205)에 저장된 프로그램을 실행시킬 수도 있다. 프로세서(210)는, 상술한 말뭉치 처리 장치(100)의 프로세서(110)와 동일하게, 중앙 처리 장치 등을 이용하여 구현 가능하다.

[0045] 일 실시예에 있어서, 프로세서(210)는, 도 2 및 도 3에 도시된 바와 같이, 말뭉치 처리부(220) 및 학습 처리부(230)를 포함할 수 있다. 말뭉치 처리부(220) 및 학습 처리부(230)는, 논리적으로 구분되는 것일 수도 있고, 물리적으로 구분되는 것일 수도 있다. 실시예에 따라서, 프로세서(210)는 학습 처리부(230)만을 포함하는 것도 가능하다. 이 경우, 프로세서(210)는 사전에 다른 정보처리장치(일례로 말뭉치 처리 장치(100)) 등에 의해 필터링 또는 분류 처리된 말뭉치를 입력부(201) 또는 저장부(205)로부터 획득하고, 이를 기반으로 학습 처리를 수행할 수도 있다.

[0046] 말뭉치 처리부(220)는 하나 이상의 말뭉치(일례로 병렬 말뭉치)를 획득하고, 획득한 말뭉치에 대해 필터링 및/또는 도메인 별 분류 등과 같은 소정의 처리를 수행할 수 있다. 말뭉치 처리부(220)는 필요에 따라 생략되는 것도 가능하다. 일 실시예에 따르면, 말뭉치 처리부(220)는, 주어진 적어도 하나의 말뭉치(입력부(201)를 통해 입력되거나 분류부(222)에 의해 분류된 말뭉치 등을 포함 가능함) 또는 적어도 하나의 말뭉치에 속하는 하나 이상의 문장이 학습을 위해 적합한지 여부를 판단하고, 판단 결과를 기반으로 모든 말뭉치의 전부 또는 일부를 제거하거나 또는 주어진 말뭉치 내의 적어도 하나의 데이터(문장이나 단어 등)을 제거하여 처리된 말뭉치를 획득하는 필터링부(221)와, 하나 이상의 말뭉치 각각을 소정의 기준에 따라 적어도 하나의 도메인으로 분류하여 분류 결과를 획득하는 분류부(222)를 포함할 수 있다. 필터링부(221)에 의해 필터링 처리되거나, 분류부(222)에 의해 분류되거나 또는 필터링부(221) 및 분류부(222) 양자에 의해 처리된 말뭉치는 바로 학습 처리부(230)로 전달될 수도 있고, 일시적 또는 비일시적으로 저장부(205)에 저장된 후, 사용자의 조작이나 미리 정의된 설정에 따라 학습 처리부(230)로 전달될 수도 있다. 필터링부(221) 및 분류부(222)는, 말뭉치 처리 장치(100)에서 설명한 필터링부(121) 및 분류부(122)와 실질적으로 동일하거나 일부 변형된 방식으로 구현 가능하므로, 이하 자세한 설명은 생략하도록 한다.

[0047] 학습 처리부(230)는 말뭉치 처리부(220)에 의해 처리된 말뭉치를 이용하여 기계 번역 모델의 훈련 및 훈련된 기계 번역 모델을 이용한 결과(일례로 적어도 하나의 원 문장에 대응하는 적어도 하나의 목표 문장 등)의 획득 중 적어도 하나를 수행할 수 있다.

[0048] 학습 처리부(230)는 하나 이상의 학습 모델을 단독으로 또는 조합 이용하여 구현될 수 있다. 여기서, 하나 이상의 학습 모델은, 예를 들어, 트랜스포머, 시퀀스투시퀀스(Seq2Seq) 모델, 심층 신경망(DNN: Deep Neural Network), 콘볼루션 신경망(CNN: Convolutional Neural Network), 순환 신경망(RNN: Recurrent Neural

Network), 콘볼루션 순환 신경망(CRNN: Convolutional Recurrent Neural Network), 심층 신뢰 신경망(DBN: Deep Belief Network), 심층 Q-네트워크(Deep Q-Networks), 장단기 메모리(LSTM: Long short term memory), 다층 퍼셉트론(Multi-layer Perceptron), 서포트 벡터 머신(SVM: support vector machine), 생성적 적대 신경망(GAN: Generative Adversarial Network) 및/또는 조건적 생성적 적대 신경망(cGAN: Conditional GAN) 등을 포함할 수 있으나, 이에 한정되는 것은 아니다. 학습 모델은 훈련을 통해 학습을 수행하고 학습 결과를 기반으로 데이터 처리를 수행할 수 있도록 설계자가 고려 가능한 적어도 하나의 학습 알고리즘이나, 학습 알고리즘을 이용하여 작성된 적어도 하나의 프로그램 코드나, 또는 이(들)의 전부 또는 일부를 기반으로 또는 이들을 전체적으로 또는 부분적으로 조합하여 구현된 적어도 하나의 프로그램 패키지 등을 포함할 수 있다.

[0049] 일 실시예에 의하면, 학습 처리부(230)는, 도 3에 도시된 바와 같이, 도메인 선택부(231), 전처리부(232), 훈련부(233), 도메인 적응부(234), 평가부(235) 및 배포처리부(236) 등을 포함할 수 있다.

[0050] 도메인 선택부(231)는, 미리 정의된 설정이나 입력부(201)를 통한 사용자의 선택 등에 따라 제작하고자 하는 인공신경망 기계 번역 모델의 도메인을 결정할 수 있다. 도메인 선택부(231)의 결정 결과(즉, 선택된 도메인)는 도메인 적응부(234)로 전달될 수 있다.

[0051] 전처리부(232)는 인공신경망 기계번역 모델의 생성을 위한 적어도 하나의 데이터(말뭉치)에 대한 전처리를 수행할 수 있다. 예를 들어, 전처리부(232)는 서브워드 분리(subword tokenization) 과정을 수행할 수도 있다. 서브워드 분리는 입력 문장 내의 하나 이상의 단어를 더 작은 단위의 서브워드로 분리하는 과정으로, 어휘 이외의 단어(OOV: Out of vocabulary)를 완화시킬 수 있는 장점이 있다. 서브워드 분리로는, 예를 들어, 바이트 페어 인코딩(BPE: Byte Pair Encoding)이나 센텐스 피이스(Sentencepiece) 등이 있다. 또한, 전처리부(223)는 주어진 말뭉치를 이용하여 모조 병렬 말뭉치를 더 생성할 수도 있다. 이 경우, 전처리부(223)는, 반대 번역(back translation) 방법이나 복제된 번역(copied translation) 방법 등을 이용할 수도 있다.

[0052] 훈련부(233)는 전처리된 말뭉치 내의 문장들을 이용하여 기계 번역 모델에 대한 훈련을 수행할 수 있다. 훈련부(233)는, 예를 들어, 트랜스포머, 순환신경망, 어텐션 메커니즘 또는 버트(BERT: Bidirectional Encoder Representation from Transformer) 등을 이용하여 훈련을 수행할 수 있다.

[0053] 도메인 적응부(234)는 훈련부(233)에 의해 훈련된 기계 번역 모델에 대해 도메인 적응(domain adaptation)을 수행하여, 도메인에 특화된 훈련된 기계 번역 모델(이하 도메인 적응 모델)을 획득할 수 있다. 예를 들어, 도메인 적응부(234)는 훈련부(233)로부터 훈련된 기계 번역 모델을 획득하고, 도메인 적응을 위해 기 학습된 적어도 하나의 학습 알고리즘을 이용하여 훈련된 기계 번역 모델에 대응하는 도메인 적응 모델을 획득할 수 있다. 이 경우, 이용되는 학습 알고리즘은 도메인 선택부(231)에 의해 선택된 도메인에 대응하는 도메인 적응 학습 알고리즘을 포함할 수 있다.

[0054] 평가부(235)는 기계 번역 모델을 평가하고, 평가 결과에 따라서 이를 배포 처리부(236)로 전달하거나, 또는 훈련부(233)나 도메인 적응부(234)로 전달할 수 있다. 여기서, 평가 대상이 되는 모델은, 훈련부(233)에 의해 훈련된 기계 번역 모델 및 도메인 적응부(234)에 의해 획득된 도메인 적응 모델 중 적어도 하나를 포함할 수 있다. 평가부(235)는 손실 함수나 대한 다중 언어 평가 언더스터디 점수 등과 같이 사전에 평가를 위해 마련된 별도의 함수 등을 기반으로 도메인 적응 모델의 번역 품질 등을 평가할 수 있다.

[0055] 배포처리부(236)는 획득된 기계 번역 모델(예를 들어, 훈련된 기계 번역 모델이나 도메인 적응 모델)의 배포를 제어 및 처리할 수 있다. 예를 들어, 배포처리부(236)는, 평가부(235)의 평가 결과에 따라서 훈련부(233)에 의해 훈련된 기계 번역 모델 또는 도메인 적응부(234)에 의해 획득된 도메인 적응 모델의 배포 여부를 결정하고, 만약 이들의 배포를 결정하면, 미리 정의된 바에 따라 또는 사용자의 요청에 따라 출력부(209)를 제어하여 훈련된 기계 번역 모델 또는 도메인 적응 모델이 적어도 하나의 다른 정보처리장치로 전달되도록 할 수도 있다.

[0056] 상술한 말뭉치 처리 장치(100) 및 기계 번역 장치(200) 중 적어도 하나는, 실시예에 따라, 상술한 말뭉치의 처리 및/또는 기계 번역 모델의 학습 처리 등을 수행할 수 있도록 특별히 고안된 장치를 이용하여 구현될 수도 있고, 또는 적어도 하나의 정보처리장치를 단독으로 또는 조합하여 이용함으로써 구현될 수도 있다. 여기서, 적어도 하나의 정보처리장치는, 예를 들어, 데스크톱 컴퓨터, 랩톱 컴퓨터, 서버용 하드웨어 장치, 스캐너 장치, 프린터 장치, 삼차원 프린터 장치, 스마트 폰, 태블릿 피씨, 스마트 시계, 스마트 태그, 스마트 밴드, 두부 장착형 디스플레이(HMD: Head Mounted Display) 장치, 휴대용 게임기, 내비게이션 장치, 스마트 키, 리모트 컨트롤러(리모컨), 디지털 텔레비전, 셋 톱 박스, 디지털 미디어 플레이어 장치, 미디어 스트리밍 장치, 디브이디 재생 장치, 컴팩트 디스크 재생 장치, 음향 재생 장치(인공 지능 스피커 등), 가전 기기(일례로 냉장고, 선풍기,

공조기, 오븐 또는 세탁기 등), 유인 또는 무인 이동체(일례로 차량(승용차, 버스나 이륜차 등), 이동성 로봇, 무선 모형 차량, 로봇 청소기 등), 유인 또는 무인 비행체(일례로 항공기, 헬리콥터, 드론, 모형 비행기 또는 모형 헬리콥터 등), 교통 제어기, 가정용, 산업용 또는 군사용 로봇, 산업용 또는 군사용 기계 또는 전자 광고판 등을 포함할 수 있으나, 이에 한정되는 것은 아니다. 설계자나 사용자 등은 상황이나 조건에 따라서 상술한 정보처리장치 이외에도 정보의 연산 처리 및 제어가 다양한 장치 중 적어도 하나를 상술한 말뭉치 처리 장치(100) 또는 기계 번역 장치(200)로 채용할 수 있다.

[0057] 이하 도 4를 참조하여 기계 번역 방법의 일 실시예에 대해서 설명하도록 한다.

[0058] 도 4는 기계 번역 방법의 일 실시예에 대한 흐름도이다.

[0059] 도 4에 도시된 바를 참조하면, 인공신경망 기계 번역 방법의 수행을 위해서 먼저 인공신경망 기계 번역 모델의 학습에 이용될 적어도 하나의 말뭉치가 획득될 수 있다(300). 적어도 하나의 말뭉치는 각각 적어도 하나의 단어나 문장 등으로 구축된 것일 수 있다. 적어도 하나의 말뭉치는 실시예에 따라서, 일 언어의 문장(또는 단어) 및 일 언어의 문장(또는 단어)에 대응하는 번역된 다른 언어의 문장(또는 단어)로 이루어진 병렬 말뭉치를 포함할 수도 있다. 적어도 하나의 말뭉치의 획득은, 사용자에 의해 직접 입력 받아 수행될 수도 있고, 다른 정보처리장치나 데이터 저장 장치로부터 전달받아 수행될 수도 있다. 예를 들어, 말뭉치의 획득은 웹 사이트 등의 크롤링을 통해 수행되는 것도 가능하다.

[0060] 적어도 하나의 말뭉치가 획득되면, 적어도 하나의 말뭉치에 대한 필터링이 수행될 수 있다(302). 이에 따라 주어진 말뭉치 내에서 상대적으로 우수한 품질의 문장이나 단어가 선별되어 양질의 말뭉치가 획득될 수 있게 된다. 필터링 과정은, 예를 들어, 병렬 말뭉치 내의 서로 쌍을 이루는 원 문장과 목표 문장을 비교하여 동일 여부를 판별하고, 원 문장과 목표 문장이 동일하면, 이들 중 어느 하나를 제거하거나 양자를 모두 제거함으로써 수행될 수도 있다. 다른 예를 들어, 필터링 과정은, 원 문장 및/또는 목표 문장 내의 각각의 문자가 미리 정의된 문자에 해당하는지 여부를 판별하고, 판별 결과 모든 문자 중에서 일정 비율 이상의 문자가 미리 정의된 문자에 해당하지 않는다면, 해당 문장(원 문장 및/또는 목표 문장)을 제거함으로써 수행될 수도 있다. 또한, 또 다른 예를 들어, 필터링 과정은, 어절의 길이, 문장 내에 불필요한 기호의 비중, 특수 기호의 개수 및/또는 원 문장 및 목표 문장 간의 대응 관계에 대한 오류 등을 기반으로 문장(원 문장 및/또는 목표 문장)을 제거함으로써 수행될 수도 있다.

[0061] 또한, 일 실시예에 의하면, 말뭉치에 대한 분류가 수행될 수도 있다(304). 이에 따라 말뭉치는 하나 이상의 특정한 도메인으로 분류될 수 있게 된다. 여기서, 도메인은 사용자나 설계자 등에 의해 미리 정의된 것일 수 있으며, 예를 들어, 논문, 보안, 특허 또는 의학 등과 같이 일상적인 언어와는 상이한 전문적인 용어를 사용하는 하나 이상의 분야를 포함할 수 있다. 말뭉치에 대한 분류는 필터링 과정을 통해 필터링된 말뭉치에 대해 수행되는 것도 가능하고, 필터링 되기 전의 말뭉치에 대해 수행되는 것도 가능하다. 도 3에는 필터링이 말뭉치의 분류에 선행하여 수행되는 것으로 도시되어 있으나, 설계자의 선택에 따라서 말뭉치의 분류가 필터링에 선행하여 수행되는 것도 가능하다. 말뭉치의 분류는, 예를 들어, 사전훈련-미세조정 접근 기법 등을 이용하여 수행될 수도 있으나, 이에 한정되는 것은 아니다.

[0062] 일 실시예에 의하면, 학습 처리 이전에 또는 학습 처리 도중에 인공신경망 기계 번역 모델이 적응될 도메인이 사용자나 또는 미리 정의된 바에 따라서 선택 및 결정될 수 있다(306).

[0063] 말뭉치에 대한 필터링(302) 및/또는 말뭉치의 분류(304)가 수행되면, 사전 설정에 따라 일정한 시간이 경과된 후 또는 사용자 등의 요청에 따라서 인공신경망 기계 번역 모델에 대한 학습이 수행될 수 있다(308). 여기서, 학습 처리는, 트랜스포머, 시퀀스투시퀀스 모델, 심층신경망 또는 순환신경망 등을 단독으로 또는 조합하여 마련된 학습 모델을 이용하여 수행될 수 있다.

[0064] 보다 구체적으로 예를 들어, 학습 처리는, 먼저 말뭉치에 대한 전처리(일례로 서브워드 분리나 모조 말뭉치의 생성 등)가 수행되고, 전처리된 말뭉치 내의 문장(또는 이의 임베딩 값)을 입력 값으로 하여 상술한 학습 모델에 대한 훈련이 수행되고, 사용자나 미리 정의된 바에 따라 선택된 도메인에 따라서 훈련된 학습 모델에 대한 도메인 적응을 수행하여 도메인 적응 모델을 획득하고, 획득한 도메인 적응 모델에 대해 평가를 수행하는 과정을 통해 수행될 수도 있다. 도메인 적응 모델의 획득이나 평가 과정은 필요에 따라 생략 가능하다.

[0065] 학습 처리가 종료되면, 인공신경망 기계 번역 모델이 획득될 수 있다(310). 인공신경망 기계 번역 모델은 사용자의 설정이나 수동 조작에 따라서 저장부에 저장될 수도 있고, 입출력단자나 유무선 통신 네트워크 등을 통해 다른 장치로 전달되어 배포될 수도 있다.

[0066] 상술한 과정(300 내지 310)은 모두 동일한 하나의 장치에 의해 수행될 수도 있고, 또는 물리적으로 분리된 다른 장치에 의해 수행될 수도 있다. 예를 들어, 모든 과정(300 내지 310)이 상술한 기계 번역 장치에 의해 수행되는 것도 가능하고, 모든 과정(300 내지 310) 중 말뭉치 획득 내지 말뭉치 분류(300 내지 304)는, 상술한 말뭉치 처리 장치에 의해 수행되고, 도메인 선택 내지 기계 번역 모델의 획득(306 내지 310)는 상술한 기계 번역 장치에 의해 수행되는 것도 가능하다.

[0067] 상술한 실시예에 따른 기계 번역 방법은, 컴퓨터 장치에 의해 구동될 수 있는 프로그램의 형태로 구현될 수 있다. 프로그램은, 명령어, 라이브러리, 데이터 파일 및/또는 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며, 기계어 코드나 고급 언어 코드를 이용하여 설계 및 제작된 것일 수 있다. 프로그램은 상술한 방법을 구현하기 위하여 특별히 설계된 것일 수도 있고, 컴퓨터 소프트웨어 분야에서 통상의 기술자에게 기 공지되어 사용 가능한 각종 함수나 정의를 이용하여 구현된 것일 수도 있다. 또한, 여기서, 컴퓨터 장치는, 프로그램의 기능을 실현 가능하게 하는 프로세서나 메모리 등을 포함하여 구현된 것일 수 있으며, 필요에 따라 통신 장치를 더 포함할 수도 있다. 상술한 기계 번역 방법을 구현하기 위한 프로그램은, 컴퓨터에 의해 판독 가능한 기록 매체에 기록될 수 있다. 컴퓨터에 의해 판독 가능한 기록 매체는, 예를 들어, 롬, 램 또는 플래시 메모리 등을 기반으로 하는 반도체 저장 매체나, 하드 디스크 또는 플로피 디스크 등과 같은 자기 디스크 저장 매체나, 콤팩트 디스크나 디브이디 등과 같은 광 기록 매체나, 및/또는 플롭티컬 디스크 등과 같은 자기-광 기록 매체 등 컴퓨터 등의 호출에 따라 실행되는 하나 이상의 프로그램을 일시적 또는 비일시적으로 저장 가능한 적어도 한 종류의 물리적 저장 매체를 포함할 수 있다.

[0068] 이상 말뭉치 처리 장치, 기계 번역 장치 및 기계 번역 방법의 여러 실시예에 대해 설명하였으나, 말뭉치 처리 장치, 기계 번역 장치 또는 기계 번역 방법은 오직 상술한 실시예에 한정되는 것은 아니다. 해당 기술 분야에서 통상의 지식을 가진 자가 상술한 실시예를 기초로 수정 및 변형하여 구현할 수 있는 다른 다양한 장치나 방법 역시 상술한 말뭉치 처리 장치, 기계 번역 장치 또는 기계 번역 방법의 일 실시예가 될 수 있다. 예를 들어, 설명된 방법(들)이 설명된 바와 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소(들)가 설명된 바와 다른 형태로 결합, 연결 또는 조합되거나 다른 구성 요소 또는 균등물 등에 의하여 대치 또는 치환되더라도, 상술한 말뭉치 처리 장치, 기계 번역 장치 및/또는 기계 번역 방법의 일 실시예가 될 수 있다.

부호의 설명

[0069] 100: 말뭉치 처리 장치 101: 입력부

105: 저장부 109: 출력부

110: 프로세서 121: 필터링부

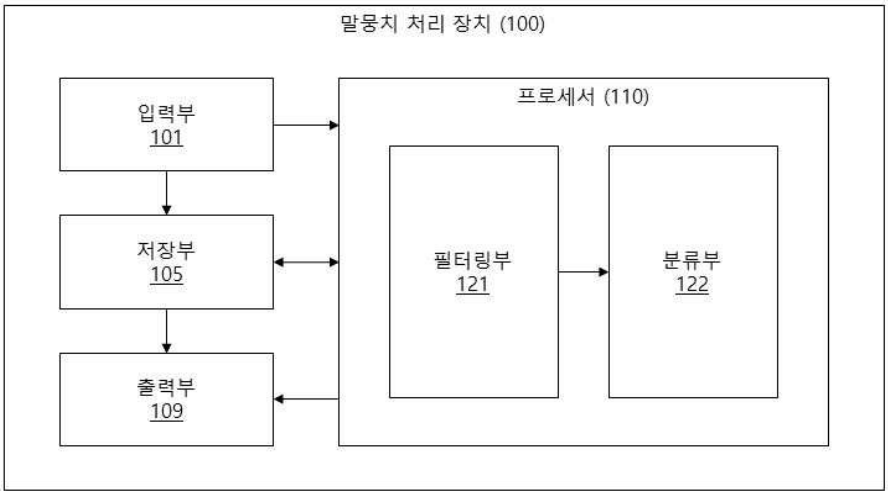
122: 분류부 200: 기계 번역 장치

210: 프로세서 220: 말뭉치 처리부

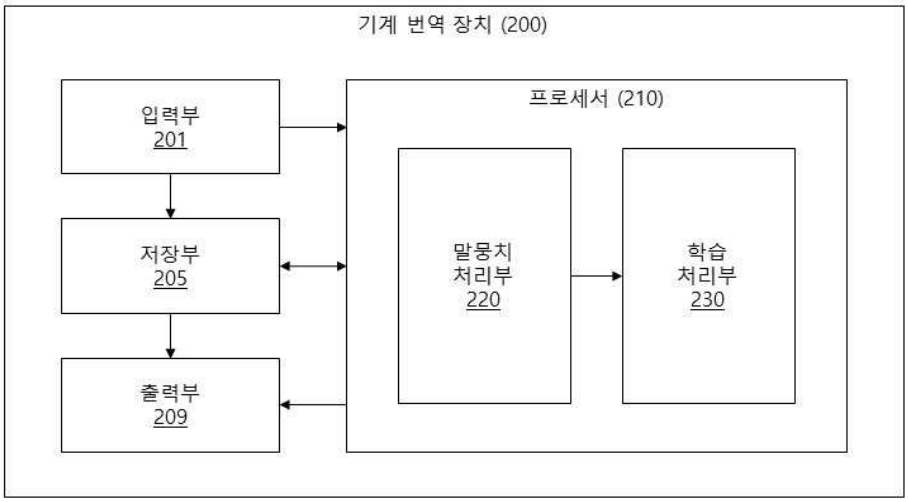
230: 학습처리부

도면

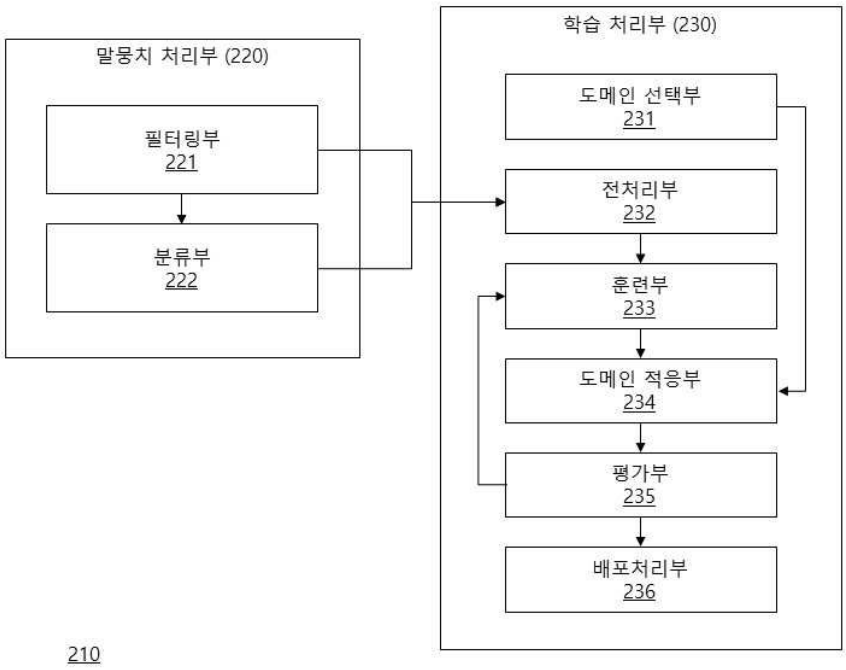
도면1


말뭉치 처리 장치 (100)
입력부
101
저장부
105
출력부
109
프로세서 (110)
필터링부
121
분류부
122

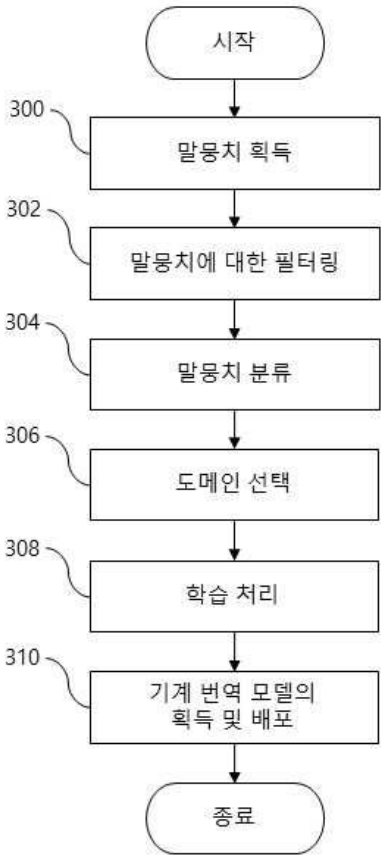
도면2


기계 번역 장치 (200)
입력부
201
저장부
205
출력부
209
프로세서 (210)
말뭉치
처리부
220
학습
처리부
230

도면3


말뭉치 처리부 (220)
필터링부
221
분류부
222
학습 처리부 (230)
도메인 선택부
231
전처리부
232
훈련부
233
도메인 적응부
234
평가부
235
배포처리부
236
210

도면4


시작
300
말뭉치 획득
302
말뭉치에 대한 필터링
304
말뭉치 분류
306
도메인 선택
308
학습 처리
310
기계 번역 모델의
획득 및 배포
종료