(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년07월28일
(11) 등록번호 10-2561817
(24) 등록일자 2023년07월26일

(51) 국제특허분류(Int. Cl.)
G06F 40/58 (2020.01) *G06F 40/263* (2020.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 40/58 (2020.01)
G06F 40/263 (2020.01)
(21) 출원번호 10-2021-0097108
(22) 출원일자 2021년07월23일
심사청구일자 2021년07월23일
(65) 공개번호 10-2023-0015675
(43) 공개일자 2023년01월31일
(56) 선행기술조사문헌
Tan, Xu, et al, Multilingual neural machine translation with knowledge distillation., arXiv preprint arXiv:1902.10461, 2019*
Tan, Xu, et al, Multilingual neural machine translation with language clustering., arXiv preprint arXiv:1908.09324, 2019*
Kim, Y., & Rush, A. M., Sequence-level knowledge distillation, arXiv preprint arXiv:1606.07947, 2016*
Johnson, Melvin, et al., Google's multilingual neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5, 2017*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
도희진
경상북도 포항시 남구 청암로 77, 여자기숙사 2동 308호
이근배
서울특별시 서초구 서운로 221, 103동 1203호
(74) 대리인
특허법인이상

전체 청구항 수 : 총 20 항 심사관 : 김영신

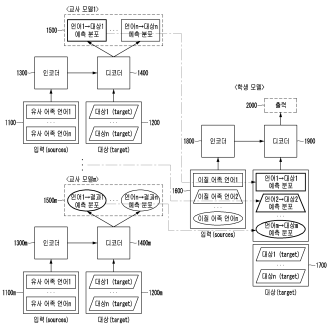
(54) 발명의 명칭 **어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템, 장치 및 방법**

(57) 요 약

본 발명의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템은, 언어의 어족 정보를 활용한 지식 증류 기법으로 다국어 기계 번역을 수행하는 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템에 있어서, m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로

(뒷면에 계속)

대 표 도 - 도2



번역하는 적어도 하나 이상의 다국어 교사 모델 및 n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델을 포함한다.

(52) CPC특허분류
G06N 3/08 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711125943
과제번호 2019-0-01906-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 인공지능핵심고급인재양성
연구과제명 인공지능대학원지원(포항공과대학교)
기 여 율 1/2
과제수행기관명 포항공과대학교 산학협력단
연구기간 2019.09.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711126317
과제번호 2020-0-01789-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 대학ICT연구센터육성지원사업
연구과제명 High Performance Knowledge System 개발 및 인력양성
기 여 율 1/2
과제수행기관명 동국대학교 산학협력단
연구기간 2020.07.01 ~ 2027.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

언어의 어족 정보를 활용한 지식 증류 기법으로 다국어 기계 번역을 수행하는 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템에 있어서,

m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 기능이 사전학습된 적어도 하나 이상의 다국어 교사 모델; 및

다국어 교사 모델의 사전학습 결과에 기반하여 n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델; 을 포함하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 2

청구항 1에 있어서, 상기 시스템은,

다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 3

청구항 1에 있어서, 상기 시스템은,

다국어 학생 모델에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 입력 소스(source) 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의 출력 분포를 목표 분포로 지정하여 학습하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 4

청구항 1에 있어서, 상기 시스템은,

다국어 학생 모델의 예측을 다국어 교사 모델의 출력 분포에 일치시키는 표준 접근법을 사용하여 지식 증류 적용을 수행하고,

시퀀스(sequence) 단위 증류 적용 시 및 단어 단위 증류 적용 시의 다국어 학생 모델의 성능 간 비교 결과에 기반하여 시퀀스 단위 보간(interpolation) 및 단어 단위 보간 중 적어도 하나 이상을 선택적으로 사용하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 5

청구항 1에 있어서, 상기 시스템은,

목표 분포(target distribution)가 원본 대상 데이터의 원-핫 레이블(one-hot label)인 음의 로그 우도(negative log-likelihood) 함수를 이용하는 제1 손실 함수, 및 목표 분포(target distribution)가 다국어 교사 모델의 출력 분포로서, 상기 출력 분포의 값에 일치하도록 학습되는 제2 손실 함수 간의 가중치 합에 의하여 생성되는 전체 손실 함수를 다국어 교사 모델의 출력을 반영하여 조정함으로써 학습이 진행되는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 6

청구항 1에 있어서, 다국어 교사 모델은,

입력이 유사 혹은 동일 어족에 속하는 2개 이상의 언어들에 대한 데이터를 가지는 입력 소스(source) 모듈;

대상이 각 입력 언어들과 쌍을 이루며 대상 언어들로 번역된 데이터를 가지는 대상 타겟(Target) 모듈;

인코더를 사용하여 인코딩하는 인코더 모듈;

디코더를 사용하여 디코딩하는 디코더 모듈; 및

입력 소스 데이터로부터 대상 타겟 언어로 번역되어 생성된 결과를 출력하는 결과 예측 분포 출력 모듈; 을 포함하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 7

청구항 1에 있어서, 다국어 학생 모델은,

입력이 이질 어족에 속하는 2개 이상의 언어들에 대한 데이터를 가지는 입력 소스(source) 모듈;

대상이 입력 데이터와 매칭되는 원본 대상 데이터와 입력 데이터에 대해 사전 학습된 교사 모델에서 저장된 출력인 top-k 예측 분포가 동시에 고려되는 대상 타겟(Target) 모듈;

인코더를 사용하여 인코딩하는 인코더 모듈;

디코더를 사용하여 디코딩하는 디코더 모듈;

입력 데이터로부터 대상 언어로 번역되어 생성된 결과를 출력하는 결과 예측 분포 출력 모듈; 을 포함하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 8

청구항 6에 있어서, 다국어 교사 모델의 입력 소스(source) 모듈 및 대상 타겟(Target) 모듈은,

모든 입력 언어들에 바이트 패어 인코딩(Byte Pair Encoding)을 적용하여 공용 어휘 집합을 공유하고,

단어 임베딩을 통해 각 단어를 밀집 벡터로 매칭시켜 표현하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 9

청구항 6에 있어서, 다국어 교사 모델의 인코더 모듈의 인코더 및 디코더 모듈의 디코더는,

트랜스포머(transformer) 모델인,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 10

청구항 6에 있어서, 다국어 교사 모델의 결과 예측 분포 출력 모듈은,

출력은 토큰에 대한 확률 분포 형식이며,

빔 검색(Beam search)을 통해 출력 분포 중 예측 확률이 가장 높은 k 개(top-k)의 토큰 인덱스와 그 확률 분포를 저장하고,

결과 분포는 각 입력 언어에 대해 개별적으로 저장하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 11

청구항 7에 있어서, 다국어 학생 모델의 입력 소스(source) 모듈 및 대상 타겟(Target) 모듈은,

모든 입력 언어들에 바이트 패어 인코딩(Byte Pair Encoding)을 적용하여 공용 어휘 집합을 공유하고,

단어 임베딩을 통해 각 단어를 밀집 벡터로 매칭시켜 표현하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 12

청구항 7에 있어서, 다국어 학생 모델의 인코더 모듈의 인코더 및 디코더 모듈의 디코더는,

트랜스포머(transformer) 모델인,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 13

청구항 7에 있어서, 다국어 학생 모델의 결과 예측 분포 출력 모듈은,

출력은 토큰에 대한 확률 분포 형식이며,

빔 검색(Beam search)을 통해 출력 분포 중 예측 확률이 가장 높은 k 개(top-k)의 토큰 인덱스와 그 확률 분포를 저장하고,

결과 분포는 각 입력 언어에 대해 개별적으로 저장하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템.

청구항 14

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치에 있어서,

프로세서(processor);

프로세서를 통해 실행되는 적어도 하나의 명령이 저장된 메모리(memory); 를 포함하되,

적어도 하나의 명령은 상기 프로세서가:

m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 기능이 사전학습된 적어도 하나 이상의 다국어 교사 모델 설정 단계;

n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델 설정 단계;

다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득하는 단계; 및

다국어 교사 모델의 사전학습 결과 및 출력 분포의 k-best 목록에 기반한 다국어 학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의 출력 분포를 목표 분포로 지정하여 학습하는 단계; 를 수행하도록 구성되는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치.

청구항 15

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법에 있어서,

m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 기능이 사전학습된 적어도 하나 이상의 다국어 교사 모델 설정 단계;

n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델 설정 단계;

다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득하는 단계; 및

다국어 교사 모델의 사전학습 결과 및 출력 분포의 k-best 목록에 기반한 학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의

출력 분포를 목표 분포로 지정하여 학습하는 단계; 를 포함하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법.

청구항 16

청구항 15에 있어서, 상기 방법은,

다국어 교사 모델의 예측 결과 분포 출력을 언어마다 개별적으로 저장하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법.

청구항 17

청구항 15에 있어서, 상기 방법은,

다국어 학생 모델의 훈련에서 다국어 학생 모델의 예측을 다국어 교사 모델의 출력 분포에 일치시키는 표준 접근법을 사용하여 학습하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법.

청구항 18

청구항 15에 있어서, 상기 방법은,

유사 어족 다국어 모델로부터 이질 어족 다국어 모델로 지식을 증류하도록 훈련하는,

어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법.

청구항 19

청구항 15 내지 청구항 18 중 어느 한 항의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법을 구현하기 위한 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 20

청구항 15 내지 청구항 18 중 어느 한 항의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법을 구현하기 위한 컴퓨터 프로그램을 기록한 컴퓨터 판독 가능한 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 신경망 네트워크를 이용한 기계 번역에 관한 것으로, 어족 기반 유사성 접근 방식을 활용한 지식 증류 기법을 다국어 신경망 기계 번역에 적용하는 기술이다.

배 경 기 술

[0002] 다국어 신경망 기계 번역 기술은 단일 모델로 여러 언어 간 번역을 수행한다. 이때 다루는 언어가 많아지고, 다양화될수록 오직 하나의 언어 쌍만 번역하는 개별 기계 번역에 비해 경쟁력 있는 성능을 유지하기가 어렵다는 한계가 있다. 이 문제를 해결하기 위해 기존 연구들은 다국어 번역 모델에서 비슷한 어족의 언어들만 다루는 방식으로 성능을 유지해왔다. 하지만, 어족 별로 모델을 따로 훈련하는 것은 다국어 기계 번역이 궁극적으로 추구하고자 하는 바와 거리가 멀다.

[0003] 본 특허에서는 이질적인 어족의 언어들을 단일 모델에서 훈련하는 것을 피하는 대신, 오히려 그 성능 개선에 초점을 둔다. 성격이 비슷한 어족의 언어들이 동일한 번역 모델에서 학습될 때 번역 성능이 더 높다는 점에 착안하여, 유사 어족 모델로부터 이질 어족 모델로 지식을 증류하는 기술을 제안한다. 이를 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역이라 명명한다.

[0004] 단일 모델에서 여러 언어 간 번역을 수행하는 다국어 신경망 기계 번역(Multilingual NMT)은 훈련 비용 감소와 저 자원(low-resource) 언어의 번역 성능향상 측면에서 이점을 가진다. 빅데이터와 고성능 하드웨어가 뒷받침되며 급속도로 발전해왔지만, 여전히 한 모델에서 다루는 언어 수와 다양성이 증가할수록 개별 기계 번역

(Individual NMT)에 비해 경쟁력 있는 성능을 유지하기 어렵다는 한계가 있다. 이에 기존 연구들은 다국어 번역 모델의 입력 소스(source) 또는 대상 타겟(target) 측에 유사 어족의 언어들만 포함하는 방식으로 성능 향상을 위한 시도를 해왔다. 그러나, 어족별로 별개의 모델을 훈련하는 것은 하나의 시스템으로 가능한 한 많은 언어를 번역하고자 하는 다국어 기계 번역의 궁극적 목표와 충돌한다.

발명의 내용

해결하려는 과제

[0005] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은 신경망 네트워크를 이용한 기계 번역에 관한 것으로, 어족 기반 유사성 접근 방식을 활용한 지식 증류 기법을 다국어 신경망 기계 번역에 적용하는 기술로써, 이질적 어족을 다루는 다국어 번역 모델의 성능을 향상시키는 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0006] 상기 목적을 달성하기 위한 본 발명의 일실시예의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템은, 언어의 어족 정보를 활용한 지식 증류 기법으로 다국어 기계 번역을 수행하는 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템에 있어서, m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 교사 모델; 및 n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델; 을 포함할 수 있다.

[0007] 상기 시스템은, 다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 할 수 있다.

[0008] 상기 시스템은, 다국어 학생 모델에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 입력 소스(source) 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의 출력 분포를 목표 분포로 지정하여 학습할 수 있다.

[0009] 상기 시스템은, 다국어 학생 모델의 예측을 다국어 교사 모델의 출력 분포에 일치시키는 표준 접근법을 사용하여 지식 증류 적용을 수행하고, 시퀀스(sequence) 단위 증류보다 단어 단위 증류가 더 잘 수행되는 점을 고려하여 단어 단위 보간(interpolation)법을 사용할 수 있다.

[0010] 상기 시스템은, 다국어 교사 모델의 출력을 반영하여 수학식 1, 수학식 2, 수학식 3을 만족하는 손실 함수를 조정하며, 전체 손실함수(L_{all})는 L_{nll}과 L_{kd}로 구성되고, L_{nll}은 목표 분포(target distribution)가 원본 대상 데이터의 원-핫 레이블(one-hot label)인 음의 로그 우도(negative log-likelihood) 함수이고, L_{kd}는 목표 분포(target distribution)가 다국어 교사 모델의 출력 분포(q)로 그 값에 일치하는 방향으로 학습이 진행될 수 있다.

[0011] 다국어 교사 모델은, 입력이 유사 혹은 동일 어족에 속하는 2개 이상의 언어들에 대한 데이터를 가지는 입력 소스(source) 모듈; 대상이 각 입력 언어들과 쌍을 이루며 대상 언어들로 번역된 데이터를 가지는 대상 타겟(Target) 모듈; 인코더를 사용하여 인코딩하는 인코더 모듈; 디코더를 사용하여 디코딩하는 디코더 모듈; 및 입력 소스 데이터로부터 대상 타겟 언어로 번역되어 생성된 결과를 출력하는 결과 예측 분포 출력 모듈; 을 포함할 수 있다.

[0012] 다국어 학생 모델은, 입력이 이질 어족에 속하는 2개 이상의 언어들에 대한 데이터를 가지는 입력 소스(source) 모듈; 대상이 입력 데이터와 매칭되는 원본 대상 데이터와 입력 데이터에 대해 사전 학습된 교사 모델에서 저장된 출력인 top-k 예측 분포가 동시에 고려되는 대상 타겟(Target) 모듈; 인코더를 사용하여 인코딩하는 인코더 모듈; 디코더를 사용하여 디코딩하는 디코더 모듈; 입력 데이터로부터 대상 언어로 번역되어 생성된 결과를 출력하는 결과 예측 분포 출력 모듈; 을 포함할 수 있다.

[0013] 다국어 교사 모델의 입력 소스(source) 모듈 및 대상 타겟(Target) 모듈은, 모든 입력 언어들에 바이트 페어 인코딩(Byte Pair Encoding)을 적용하여 공용 어휘 집합을 공유하고, 단어 임베딩을 통해 각 단어를 밀집 벡터로 매칭시켜 표현할 수 있다.

[0014] 다국어 교사 모델의 인코더 모듈의 인코더 및 디코더 모듈의 디코더는, 트랜스포머(transformer) 모델일 수 있다.

[0015] 다국어 교사 모델의 결과 예측 분포 출력 모듈은, 출력은 토큰에 대한 확률 분포 형식이며, 빔 검색(Beam search)을 통해 출력 분포 중 예측 확률이 가장 높은 k 개(top-k)의 토큰 인덱스와 그 확률 분포를 저장하고, 결과 분포는 각 입력 언어에 대해 개별적으로 저장할 수 있다.

[0016] 다국어 학생 모델의 입력 소스(source) 모듈 및 대상 타겟(Target) 모듈은, 모든 입력 언어들에 바이트 패어 인코딩(Byte Pair Encoding)을 적용하여 공용 어휘 집합을 공유하고, 단어 임베딩을 통해 각 단어를 밀집 벡터로 매칭시켜 표현할 수 있다.

[0017] 다국어 학생 모델의 인코더 모듈의 인코더 및 디코더 모듈의 디코더는, 트랜스포머(transformer) 모델일 수 있다.

[0018] 다국어 학생 모델의 결과 예측 분포 출력 모듈은, 출력은 토큰에 대한 확률 분포 형식이며, 빔 검색(Beam search)을 통해 출력 분포 중 예측 확률이 가장 높은 k 개(top-k)의 토큰 인덱스와 그 확률 분포를 저장하고, 결과 분포는 각 입력 언어에 대해 개별적으로 저장할 수 있다.

[0019] 본 발명의 다른 목적을 달성하기 위한 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치는, 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치에 있어서, 프로세서(processor); 프로세서를 통해 실행되는 적어도 하나의 명령이 저장된 메모리(memory); 를 포함하되, 적어도 하나의 명령은 상기 프로세서가: m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 교사 모델 설정 단계; n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델 설정 단계; 다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득하는 단계; 및 다국어 학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의 출력 분포를 목표 분포로 지정하여 학습하는 단계; 를 수행하도록 구성될 수 있다.

[0020] 본 발명의 또 다른 목적을 달성하기 위한 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법은, 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법에 있어서, m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 교사 모델 설정 단계; n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델 설정 단계; 다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득하는 단계; 및 학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의 출력 분포를 목표 분포로 지정하여 학습하는 단계; 를 포함할 수 있다.

[0021] 상기 방법은, 다국어 교사 모델의 예측 결과 분포 출력을 언어마다 개별적으로 저장할 수 있다.

[0022] 상기 방법은, 다국어 학생 모델의 훈련에서 각 입력 언어가 속한 어족을 다루는 다국어 교사 모델의 예측 분포를 목표로 지정하여 학습할 수 있다.

[0023] 상기 방법은, 유사 어족 다국어 모델로부터 이질 어족 다국어 모델로 지식을 증류하도록 훈련할 수 있다.

[0024] 본 발명의 또 다른 목적을 달성하기 위한 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법을 구현하기 위한 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램일 수 있다.

[0025] 본 발명의 또 다른 목적을 달성하기 위한 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법의 프로그램을 구현하기 위한 컴퓨터 판독 가능한 기록매체일 수 있다.

발명의 효과

[0026] 본 발명의 일 실시예에 따르면, 상대적으로 성능이 좋은 유사 어족 교사 모델로부터 지식을 증류 받아 이질 어족 학생 모델의 성능을 향상시킨다.

[0027] 이로써, 다국어 기계 번역 기술에서 언어 다양성이 초래하는 성능 저하 문제를 개선한다.

[0028] 또한, 지식 증류 시에 개별 모델이 아닌 다국어 모델을 교사로 지정함으로써 자원 효율성을 증가시킨다.

도면의 간단한 설명

[0029] 도 1은 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 모델 개념도이다.

도 2는 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 시스템의 구성도이다.

도 3은 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 장치의 구성도이다.

도 4는 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0030] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0031] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는 데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는"이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0032] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0033] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0034] 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 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0036] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0038] 다국어 신경망 기계 번역은 단일 언어 쌍을 번역하는 개별 기계 번역(Individual NMT)에서 나아가 여러 언어 쌍을 하나의 모델로 번역한다. 모델의 어느 부분을 공유하는지에 따라 다양한 방법이 있는데, 다중 인코더(encoder)와 디코더(decoder)를 사용하는 방법, 단일 인코더만 공유하는 방법, 인코더와 디코더 모두를 공유하는 보편적(universal) 방법이 있다.

[0039] 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 일 실시예에서는, 보편적 모델에 기반을 두며 모든 언어 쌍 학습에 공용 어휘 집합(vocabulary)과 공용 매개변수(parameter)를 사용한다.

[0040] 다국어 기계 번역에서는 언어학적으로 가까운 언어들이 동일 모델로 학습될 때 어휘적, 통사적 유사성으로 더 나은 성능을 보인다. 다국어 기계 번역에서는 어떤 언어들을 다룰지 고려하는 것이 중요하며 언어들을 어족 단위로 클러스터링하여 그룹별 별개의 다국어 모델을 만들었다. 그러나 어족별로 별개 모델을 훈련하는 것은 최대한 다양한 언어를 번역할 수 있는 단일 모델을 개발하려는 다국어 기계 번역의 지향점과 어긋난다. 서로 다른

언어 계열을 포함하면서도 성능 면에서 경쟁력을 갖춘 모델이 필요하다. 이를 위해 전체 매개변수 공유 대신 부분 매개변수 공유를 제안한다.

[0041] 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 일 실시예에서는, 다국어 기계 번역의 궁극적 목표에 중점을 두며 전체 매개변수 공유(full parameter-sharing) 방식을 사용한다.

[0042] 인공 신경망의 크기와 복잡도가 커짐에 따라 모델을 작고 단순하게 만드는 압축기술들이 많이 개발되었다. 지식 증류는 심층학습 분야에서 널리 사용되는 압축 기술 중 하나로, 사전학습(pre-training)된 큰 교사 네트워크로부터 지식을 증류하여 작은 학생 네트워크의 성능을 향상하는 방법이다. 기계 번역 분야에서도 모델 복잡성이 증가함에 따라 모델 압축 기술을 적용하려는 시도가 있었다. 기계 번역에 단어와 시퀀스(sequence) 수준(level)의 지식 증류를 적용하여 해당 모델이 지식 증류를 적용하지 않은 모델보다 성능이 우수하다는 사실이 알려져 있다. 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 일 실시예는 다국어 기계 번역에 지식 증류를 적용한 방법으로, 하나의 다국어(multilingual) 학생 모델 훈련을 위해 다수의 개별(individual) 모델을 교사 모델로 사용한다.

[0043] 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 일 실시예는 마찬가지로 다국어 기계 번역에 지식 증류를 적용하지만, 다음의 두 가지 측면에서 특징이 있다.

[0044] 첫째로, 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은 개별 모델 대신 다국어 모델을 교사 모델로 사용한다.

[0045] 둘째로, 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은 훈련 과정에서 어족을 고려하여 성능을 높인다.

[0046] 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법으로는 N개 언어 쌍 훈련을 위해 N개의 교사 모델을 사전 학습하는 대신, 어족별로 그룹화한 N/k개의 교사 모델만 사전학습하여 전체 훈련에 드는 비용을 절감할 수 있다.

[0047] 전술한 바와 같이, 동일 다국어 모델에서 함께 훈련될 언어들의 계열 고려는 성능에 영향을 미친다. 특히, 유사 어족만 다루는 모델은 이질 어족 그룹으로 훈련하는 모델보다 대체로 성능이 더 우수하다. 이를 바탕으로 본 발명이 제안하는 방법은 사전학습된 유사 어족 다국어 모델의 지식을 증류하여 이질 어족 다국어 모델의 번역 성능을 높이는 것을 목표로 한다. 본 발명이 제안하는 방법은 저성능 네트워크의 훈련 과정에서 지식 증류를 통해 고성능 교사 모델로부터 배우게 된다. 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법에서는 다대일(many-to-one) 번역을 다루기에 언어 유사성 고려는 입력 소스(source) 측에 대한 것이다.

[0049] 도 1은 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 모델 개념도이다.

[0050] 도 1을 참조하면, 본 발명의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은 크게 유사 어족을 다루는 교사 모델과 이질 어족을 다루는 학생 모델 두 부분으로 구성된다.

[0051] 다수의 교사 모델은 유사 어족 언어들을 입력 소스(source)로, 영어를 대상 타겟(target)으로 한다. 예를 들어, 첫 번째 교사 모델은 인도유럽어족 언어들에서 영어로, 두 번째 교사 모델은 오스트로아시아어족 언어들에서 영어로, 세 번째 교사 모델은 알타이어족 언어들에서 영어로 번역한다.

[0052] 학생 모델은 세 개의 서로 다른 어족에 속한 언어들에서 영어로 번역하는 모델이다. 우선, 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득한다. 이후 학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는 교사 모델의 출력 분포를 목표 분포로 지정하여 학습하게 된다.

[0053] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은 지식 증류 적용을 위해 학생의 예측을 교사의 출력 분포에 일치시키는 표준 접근법을 택한다.

[0054] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은 시퀀스(sequence) 단위 증류보다 단어 단위 증류가 더 잘 수행되었다는 점을 고려하여 단어 단위 보간(interpolation)법을 사용한다. 이는 교사의 출력을 반영하여 손실 함수를 조정함으로써 이루어지며, 지식 증류에서 사용되는 구체적 수학식은 수학식 1, 수학식 2, 수학식 3 로 간략하게 표현할 수 있다.

[0056] (수학식 1)

[0057] $$L_{all} = (1-\alpha) \cdot L_{nll} + \alpha \cdot L_{kd}$$

[0059] (수학식 2)

[0060] $$L_{nll} = -\sum_{(x,y)} \log p(y|x)$$

[0062] (수학식 3)

[0063] $$L_{kd} = -\sum_{(x,y)} q(y|x) \cdot \log p(y|x)$$

[0065] 전체 손실함수(L_{all})는 L_{nll}과 L_{kd}로 구성된다. L_{nll}은 목표 분포(target distribution)가 원본 대상 데이터의 원-핫 레이블(one-hot label)인 음의 로그 우도(negative log-likelihood) 함수이며, L_{kd}는 목표 분포가 교사의 출력 분포(q)로 그 값에 일치하는 방향으로 학습이 진행된다.

[0067] <실험예>

[0068] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 실험예를 위해 공개 데이터인 Multitarget TED Talks Task Dataset(MTTT)을 사용하였다. 원본의 검증(validation) 및 테스트(test) 데이터가 적은 점을 고려하여 훈련데이터 무작위 추출로 약 15:1:1 비율을 유지하도록 재구성하였다. 각 언어 쌍별 데이터 크기는 표 1과 같다.

[0069] 교사 모델과 학생 모델이 동일 어휘 집합을 공유하도록 모든 언어에 대해 바이트 페어 인코딩(BPE)과 서브워드(subword) 병합 연산을 적용하였다. 모든 네트워크에서 4개 헤드(head) 2- 레이어(layer) 트랜스포머를 사용하고 총 매개변수 수는 14.8M 개이다. 4 개의 RTX 2080Ti GPU를 이용하여 학습하고 빔 검색 시 k 값은 4로 지정하였다. 표 1은 6개 언어 쌍의 데이터셋 크기를 보여준다.

[0071] (표 1)

[0072]

언어 쌍	Train	Valid	Test
De-En (독일어-영어)	138K	9K	9K
Fr-En (프랑스어-영어)	143K	9K	9K
Ja-En (일본어-영어)	149K	9K	9K
Ko-En (한국어-영어)	139K	9K	9K
Vi-En (베트남어-영어)	99K	7K	7K
Zh-En (중국어-영어)	154K	9K	10K

[0074] 실험예에서는 여러 입력 소스(source) 언어와 하나의 대상 타겟(target) 언어로 구성되는 다대일 사례를 다루었다. 제안 방법 검증을 위해 우선 두 개의 어족 그룹으로 첫 번째 실험 후, 세 그룹으로 범위를 확장해 두 번째 실험을 진행하였다. 유사 어족 집단을 입력 소스(source)로 다루는 각 다국어 교사 모델에서 선택된 언어는 병렬 데이터의 충분성과 언어학적 근접성을 기반으로 결정하였다. 실험 1에서는 {Ko, Ja→En}, {Vi, Zh→En}의 두 교사 모델을 사전학습하고 {Ko, Vi→En}, {Ja, Zh→En}의 두 학생 모델로 지식을 증류하였다. 실험 2에서는 {De,Fr→En} 교사 모델 추가 후 {De, Ko, Vi→En}, {Fr, Ja, Zh→En}의 두 학생 모델로 지식 증류를 적용하였다. 기준(baseline)모델은 학생과 모든 조건이 같지만, 지식 증류를 적용하지 않은 모델이다. 실험 결과는 표 2와 같다.

[0076] (표 2)

[0077]

역할	실험1				실험2					
	Ko-En	Ja-En	Vi-En	Zh-En	Ko-En	Ja-En	Vi-En	Zh-En	De-En	Fr-En
개별 모델	17.32	12.09	26.96	18.38	17.22	12.22	27.02	18.4	33.51	38.53
교사1	17.78	13.37	–	–	17.92	13.44	–	–	–	–
교사2	–	–	27.69	18.94	–	–	27.32	18.96	–	–
교사3	–	–	–	–	–	–	–	–	33.76	39.57
학생1	18.33	–	28.35	–	17.96		28.03	–	33.99	–
학생2	–	13.31	–	19.67	–	12.82	–	19.34	–	39.14
기준1 (baseline)	17.61	–	27.73	–	17.12	–	27.37	–	33.11	–
기준2 (baseline)	–	12.72	–	18.96	–	12.35	–	18.77	–	38.64

[0079] 교사 모델을 각 언어 쌍에 대한 개별모델과 비교하면 유사 어족 언어가 동일 모델에서 훈련되는 것의 효과를 알 수 있다. 모든 교사 모델이 개별모델보다 더 높은 BLEU 점수(Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu, “Bleu: a method for automatic evaluation of machine translation”, In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311-318, 2002)를 보인다. 이는 유사 어족 언어들이 동일 모델에서 학습되는 것이 번역에 도움이 된다는 가정을 공고히 한다. 실험 1과 2 모두에서 제안 방법으로 학습된 학생 1과 2는 모든 언어 쌍에서 기준 1과 2보다 높은 BLEU 점수를 보였다. 또한, 일본어와 프랑스어를 제외한 모든 언어에서 학생 1과 2가 교사보다 뛰어난 가장 높은 성능을 달성했다. 일반적으로 지식 증류의 목표는 교사 모델 능가가 아닌 학생 모델 자체의 성능 향상이다. 그렇기에 교사보다 우수한 학생 모델은 드물다는 점에서 실험 결과는 제안된 모델의 효과를 검증한다.

[0080] 본 발명의 일 실시예에서는 이질 어족을 다루는 다국어 기계 번역의 성능 저하 문제를 개선하기 위하여 어족 기반 지식 증류 방법을 제안한다. Ted Talk 데이터셋을 사용한 실험으로 제안 방법을 검증하였으며 어족 그룹을 추가하여 한 번의 실험을 더 진행함으로써 제안 방법의 확장성을 보였다.

[0081] 본 발명의 일 실시예는 기존의 다국어 기계 번역에서 언어 다양성이 초래하는 성능 저하를 개선하고, 지식 증류 시 개별 모델이 아닌 다국어 모델을 교사로 택함으로써 선행 연구보다 자원 효율성을 증가시켰다. 제안 방법을 사용한다면 다국어 기계 번역에서 다루는 언어 종류가 다양하더라도 경쟁력 있는 성능을 유지할 수 있을 것이다.

[0083] 도 2는 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 시스템의 구성도이다.

[0084] 도 2를 참조하면, 교사 모델의 입력 소스(source) 모듈(1100m)에서, 각 교사 모델의 입력은 유사 혹은 동일 어족에 속하는 2개 이상의 언어들에 대한 데이터로 구성된다. 예를 들어, 교사 모델1의 입력 데이터는 알타이어족의 언어로만 구성되고, 교사 모델2의 입력 데이터는 인도유럽어족의 언어로만 구성되며 교사 모델3의 입력 데이터는 오스트로아시아어족 언어로만 구성된다. 여러 유사 어족 언어들이 동일한 모델을 통해 학습됨으로써 어휘적, 통사적 근접성으로 인해 하나의 언어 쌍만 학습하는 개별 모델보다 더 나은 성능을 보인다. 모든 입력 언어들에 바이트 페어 인코딩(Byte Pair Encoding)을 적용하여 공용 어휘 집합을 공유하도록 한다. 이후, 단어 임베딩을 통해 각 단어를 밀집 벡터로 매칭시켜 표현한다.

[0085] 교사 모델의 대상 타겟(Target) 모듈(1200m)에서, 교사 모델의 대상은 각 입력 언어들과 동일한 의미를 가지며 쌍을 이루는 대상 언어들의 데이터이다. 즉, 각 입력 언어들이 대상 언어로 번역된 데이터이다. 본 모델에서는 여러 개 언어에서 하나의 언어로 번역하는 다대일 경우를 다루었기 때문에 대상 데이터들의 언어는 동일하다. 입력 데이터와 마찬가지로 바이트 페어 인코딩(Byte Pair Encoding)과 단어 임베딩을 적용한다.

[0086] 교사 모델의 인코더 모듈(1300m)에서, 교사 모델의 인코더는 트랜스포머(transformer) 모델의 인코더를 사용한다. 교사 모델의 인코더 관련하여 https://github.com/pytorch/fairseq 을 참고한다.

[0087] 교사 모델의 디코더 모듈(1400m)에서, 교사 모델의 디코더는 트랜스포머(transformer) 모델의 디코더를 사용한다. 교사 모델의 디코더 관련하여 https://github.com/pytorch/fairseq 을 참고한다.

[0088] 교사 모델의 결과 예측 분포 출력 모듈(1500m)에서, 교사 모델 디코더의 출력은 입력 데이터로부터 대상 언어로 번역된 예측 결과이다. 토큰에 대한 확률 분포 형식이며, 빔 검색(Beam search)을 통해 출력 분포 중 예측 확률이 가장 높은 k 개(top-k)의 토큰 인덱스와 그 확률 분포를 저장한다. 학생 모델 학습 과정에서 이 결과 분포를 이용하여 지식 증류 기법을 적용하기 위함이다. 이 때, 결과 분포는 각 입력 언어에 대해 개별적으로 저장한다.

[0089] 학생 모델의 입력 소스(source) 모듈(1600)에서, 학생 모델의 입력은 이질 어족에 속하는 2개 이상의 언어들에 대한 데이터로 구성된다. 예를 들어, 입력 데이터 중 한 언어는 알타이어족에 속하고, 다른 언어는 인도 유럽어족에 속하고, 또 다른 언어는 오스트로아시아어족에 속한다. 교사 모델의 입력 데이터와 마찬가지로 바이트 페

어 인코딩(Byte Pair Encoding)과 단어 임베딩을 적용한다.

[0090] 학생 모델의 대상 타겟(Target) 모듈(1700)에서, 학생 모델의 대상은 입력 데이터와 매칭되는 원본 대상 데이터(교사 모델의 대상과 동일한 형태)와 입력 데이터에 대해 사전 학습된 교사 모델에서 저장된 출력인 top-k 예측 분포가 동시에 고려된다. 각 입력 언어에 대해 해당 언어가 속한 어족을 다루는 사전 학습된 교사 모델의 출력 분포를 가져오게 된다. 학생 모델의 훈련 과정에서 앞의 두 가지 대상을 동시에 반영하여 손실 함수를 계산하며, 이로써 교사 모델로부터의 지식 증류가 일어난다. 입력 데이터와 마찬가지로 원본 대상 데이터에 대해서는 바이트 패어 인코딩(Byte Pair Encoding)과 단어 임베딩을 적용한다.

[0091] 학생 모델의 인코더 모듈(1800)에서, 학생 모델의 인코더는 교사 모델과 마찬가지로 트랜스포머(transformer) 모델의 인코더를 사용한다.

[0092] 학생 모델의 디코더 모듈(1900)에서, 학생 모델의 디코더는 교사 모델과 마찬가지로 트랜스포머(transformer) 모델의 디코더를 사용한다.

[0093] 학생 모델의 결과 예측 분포 출력 모듈(2000)에서, 학생 모델 디코더의 출력은 입력 데이터로부터 대상 언어로 번역되어 생성된 결과이다.

[0095] 도 3은 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치(100)의 구성도이다.

[0096] 도 3을 참조하면, 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치(100)는, 프로세서(110), 메모리(120), 송수신 장치(transceiver, 130), 입력 인터페이스 장치(140), 출력 인터페이스 장치(150), 저장 장치(160) 및 버스(bus)(170)를 포함하여 구성될 수 있다.

[0097] 프로세서(110)는 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 발명의 실시예들에 따른 방법들이 수행되는 전용의 프로세서를 의미할 수 있다.

[0098] 메모리(120) 및 저장 장치(160) 각각은 휘발성 저장 매체 및 비휘발성 저장 매체 중에서 적어도 하나로 구성될 수 있다. 예를 들어, 메모리(120)는 읽기 전용 메모리(read only memory, ROM) 및 랜덤 액세스 메모리(random access memory, RAM) 중에서 적어도 하나로 구성될 수 있다.

[0099] 또한, 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치(100)는, 무선 네트워크를 통해 통신을 수행하는 송수신 장치(transceiver, 130)를 포함할 수 있다.

[0100] 또한, 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치(100)는 입력 인터페이스 장치(140), 출력 인터페이스 장치(150), 저장 장치(160) 등을 더 포함할 수 있다.

[0101] 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치(100)에 포함된 각각의 구성 요소들은 버스(bus)(170)에 의해 연결되어 서로 통신을 수행할 수 있다.

[0102] 본 발명의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 장치(100)의 예를 들면, 통신 가능한 데스크탑 컴퓨터(desktop computer), 랩탑 컴퓨터(laptop computer), 노트북(notebook), 스마트폰(smart phone), 태블릿 PC(tablet PC), 모바일폰(mobile phone), 스마트 워치(smart watch), 스마트 글래스(smart glass), e-book 리더기, PMP(portable multimedia player), 휴대용 게임기, 네비게이션(navigation) 장치, 디지털 카메라(digital camera), DMB(digital multimedia broadcasting) 재생기, 디지털 음성 녹음기(digital audio recorder), 디지털 음성 재생기(digital audio player), 디지털 동영상 녹화기(digital video recorder), 디지털 동영상 재생기(digital video player), PDA(Personal Digital Assistant) 등일 수 있다.

[0104] 도 4는 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법의 순서도이다.

[0105] 도 4를 참조하면, 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은 S1000 단계 내지 S4000 단계를 포함한다.

[0106] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은, 다국어 교사 모델 설정 단계(S1000), 다국어 학생 모델 설정 단계(S2000), 출력 분포의 k-best 목록을 획득하는 단계(S3000) 및 학생 모델 학습 단계(S4000)를 포함한다.

[0107] S1000 단계에서, m 개의 유사 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 교사 모델 설정한다.

[0108] S2000 단계에서, n 개의 이질 어족에 속한 언어들을 입력 소스(source)로, 제1 언어를 대상 타겟(target)으로 번역하는 적어도 하나 이상의 다국어 학생 모델 설정한다.

[0109] S3000 단계에서, 다국어 교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여 각 언어에 대한 출력을 별도 저장함으로써 출력 분포의 k-best 목록을 획득한다.

[0110] S4000 단계에서, 학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상 타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는 다국어 교사 모델의 출력 분포를 목표 분포로 지정하여 학습한다.

[0111] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은, 다국어 교사 모델의 예측 결과 분포 출력을 언어마다 개별적으로 저장한다.

[0112] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은, 다국어 학생 모델의 훈련에서 각 입력 언어가 속한 어족을 다루는 다국어 교사 모델의 예측 분포를 목표로 지정하여 학습한다.

[0113] 본 발명의 일 실시예의 어족 기반 지식 증류 기법을 활용한 다국어 신경망 기계 번역 방법은, 유사 어족 다국어 모델로부터 이질 어족 다국어 모델로 지식을 증류하도록 훈련한다.

[0114] 본 발명의 일 실시예에 따른 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법을 구현하기 위한 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램일 수 있다.

[0115] 본 발명의 일 실시예에 따른 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 방법의 프로그램을 구현하기 위한 컴퓨터 판독 가능한 기록매체일 수 있다.

[0117] 이상, 본 발명의 실시예에 따른 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템, 장치 및 방법에 대해 설명하였다.

[0118] 본 발명의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템, 장치 및 방법의 일 실시예에서, 다국어 기계 번역에서 성격이 다른 어족의 언어들을 다루는 것을 피하는 대신, 오히려 그 성능 개선에 초점을 둔다.

[0119] 본 발명의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템, 장치 및 방법의 일 실시예에서, 이질 어족 언어들을 다루는 다국어 모델의 성능 향상을 위해 유사성 접근 방식을 활용한 지식 증류 방법을 제안한다. 구체적으로, 유사 어족 언어를 다루어 상대적으로 성능이 우수한 다국어 모델을 교사 모델로 지정하고 이질 어족 언어를 다루는 다국어 모델을 학생 모델로 지정하여 지식 증류 기법을 적용한다.

[0120] 본 발명의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템, 장치 및 방법의 일 실시예에서, 학생 모델의 훈련과정에서 각 언어의 어족에 기반을 두어 최적의 교사 모델을 대상화(targeting)함으로써 언어 다양성이 초래하는 성능 저하 문제를 해결한다.

[0121] 본 발명의 어족 기반 지식 증류 기법을 적용한 다국어 신경망 기계 번역 시스템, 장치 및 방법의 일 실시예에서, 공개된 Ted Talk 데이터로 진행한 실험을 통해 제안 방법으로 이질 어족 언어를 다루는 다국어 기계 번역 성능을 높이고, 교사 모델보다도 우수한 성능을 보일 수 있음을 검증한다.

[0123] 본 발명의 실시예에 따른 방법의 동작은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 정보가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

[0124] 또한, 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0125] 본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의

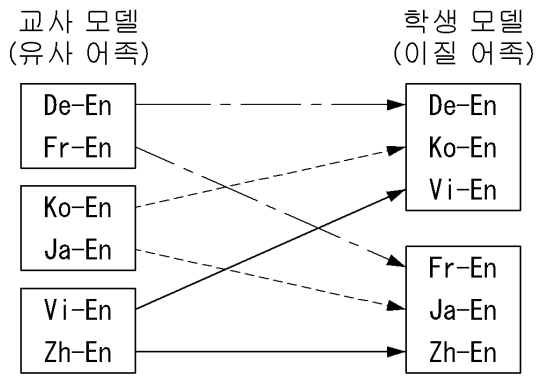
해(또는 이용하여) 수행될 수 있다. 몇몇의 실시예에서, 가장 중요한 방법 단계들의 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0126] 실시예들에서, 프로그램 가능한 로직 장치(예를 들어, 필드 프로그래머블 게이트 어레이)가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시예들에서, 필드 프로그머블 게이트 어레이는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

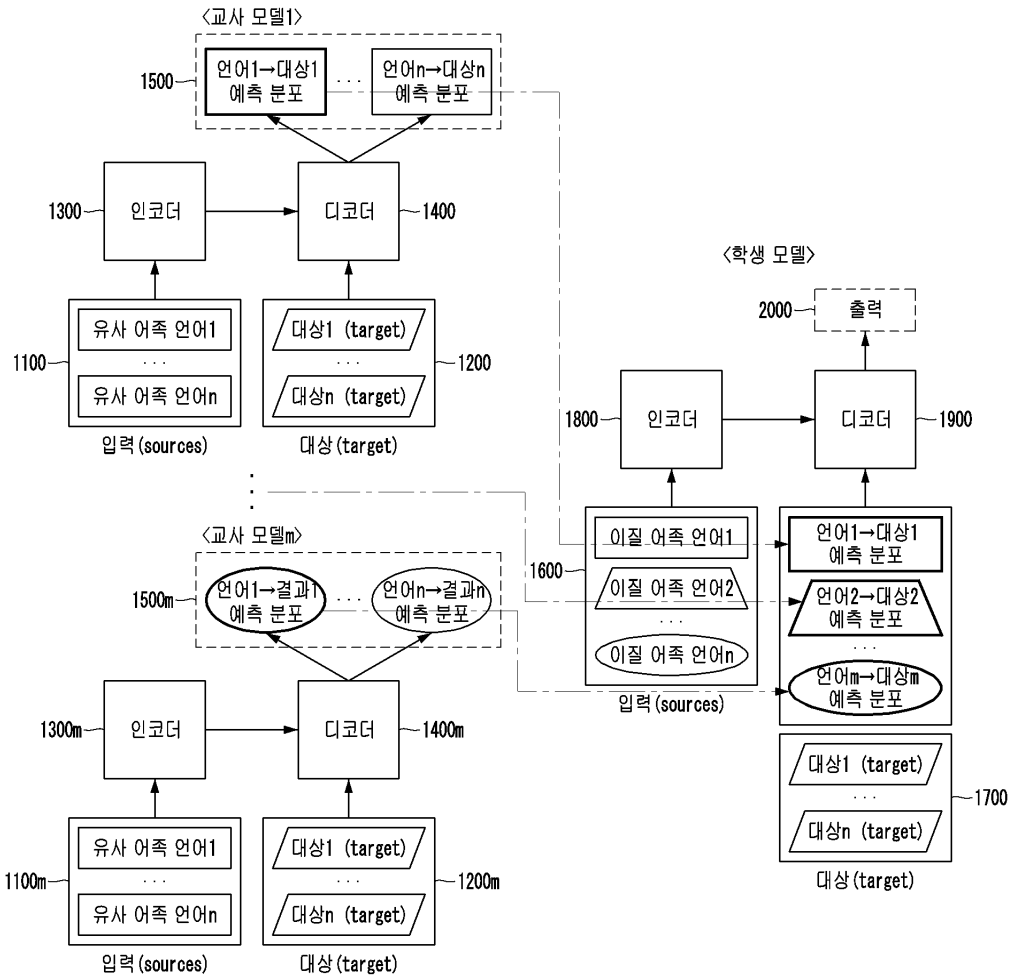
[0127] 이상 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

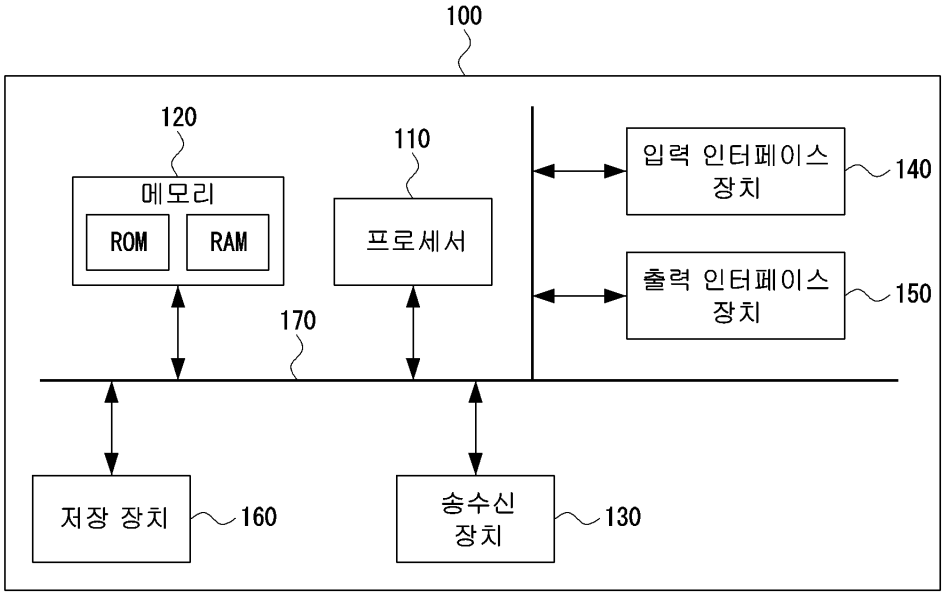
도면1



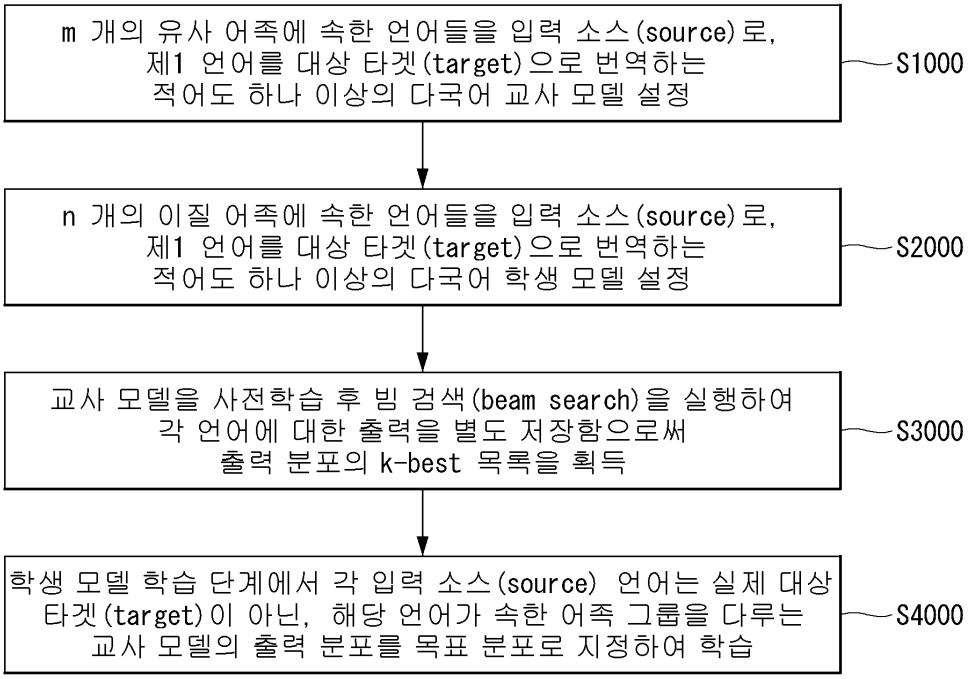
도면2


〈교사 모델1〉
1500
언어1→대상1 예측 분포
언어n→대상n 예측 분포
1300
인코더
디코더
1400
1100
유사 어족 언어1
유사 어족 언어n
입력(sources)
대상1 (target)
대상n (target)
1200
대상(target)
〈학생 모델〉
2000
출력
1800
인코더
디코더
1900
〈교사 모델m〉
1500m
언어1→결과1 예측 분포
언어n→결과n 예측 분포
1600
이질 어족 언어1
이질 어족 언어2
이질 어족 언어n
입력(sources)
언어1→대상1 예측 분포
언어2→대상2 예측 분포
언어m→대상m 예측 분포
1300m
인코더
디코더
1400m
1100m
유사 어족 언어1
유사 어족 언어n
입력(sources)
대상1 (target)
대상n (target)
1200m
대상(target)
1700
대상(target)

도면3


100
120
메모리
ROM
RAM
110
프로세서
입력 인터페이스 장치
140
출력 인터페이스 장치
150
170
저장 장치
160
송수신 장치
130

도면4


m 개의 유사 어족에 속한 언어들을 입력 소스(source)로,
제1 언어를 대상 타겟(target)으로 번역하는
적어도 하나 이상의 다국어 교사 모델 설정
S1000
n 개의 이질 어족에 속한 언어들을 입력 소스(source)로,
제1 언어를 대상 타겟(target)으로 번역하는
적어도 하나 이상의 다국어 학생 모델 설정
S2000
교사 모델을 사전학습 후 빔 검색(beam search)을 실행하여
각 언어에 대한 출력을 별도 저장함으로써
출력 분포의 k-best 목록을 획득
S3000
학생 모델 학습 단계에서 각 입력 소스(source) 언어는 실제 대상
타겟(target)이 아닌, 해당 언어가 속한 어족 그룹을 다루는
교사 모델의 출력 분포를 목표 분포로 지정하여 학습
S4000