(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0078945
(43) 공개일자 2024년06월04일

(51) 국제특허분류(Int. Cl.)
G06F 40/51 (2020.01) *G06F 40/186* (2020.01)
G06F 40/284 (2020.01) *G06F 40/58* (2020.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 40/51 (2020.01)
G06F 40/186 (2020.01)
(21) 출원번호 10-2022-0161686
(22) 출원일자 2022년11월28일
심사청구일자 2022년11월28일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
어수경
경상남도 합천군 삼가면 금리4길 17
(74) 대리인
김등용

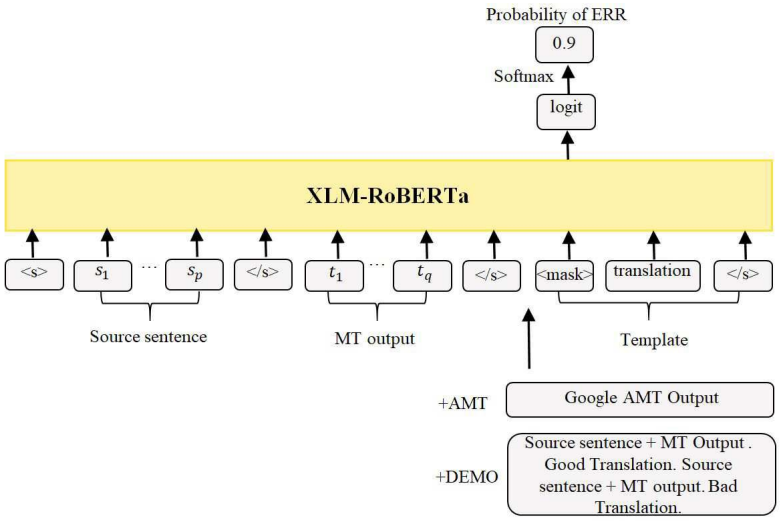
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **프롬프트를 활용한 기계번역 결과 치명적 오류 감지 방법 및 장치**

(57) 요 약

프롬프트를 활용한 기계번역 결과 치명적 오류 감지 모델 생성 장치 및 방법과 이를 이용한 치명적 오류 감지 장치 및 방법이 개시된다. 상기 치명적 오류 감지 모델 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되고, 소스 문장, 번역 문장, 및 치명적 오류의 유무를 나타내는 레이블(label)을 포함하는 학습 데이터를 소스 문장, 번역 문장, 및 디스크립션(description)을 포함하는 프롬프트 템플릿(prompt template)으로 변환하는 단계, 상기 레이블을 상기 레이블의 의미가 상기 디스크립션 내에서 표현되도록 하는 구체적인 단어로 변환하는 단계, 및 상기 프롬프트 템플릿을 이용하여 사전학습 언어 모델(Pre-trained Language Model, PLM)을 학습하여 상기 치명적 오류 감지 모델을 생성하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 40/284 (2020.01)
G06F 40/58 (2020.01)
G06N 20/00 (2021.08)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711159969
과제번호 2018-0-01405-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345349433
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교 산학협력단
연구기간 2022.03.01 ~ 2023.02.28

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 치명적 오류 감지 모델 생성 방법에 있어서,

소스 문장, 번역 문장, 및 치명적 오류의 유무를 나타내는 레이블(label)을 포함하는 학습 데이터를 소스 문장, 번역 문장, 및 디스크립션(description)을 포함하는 프롬프트 템플릿(prompt template)으로 변환하는 단계;

상기 레이블을 상기 레이블의 의미가 상기 디스크립션 내에서 표현되도록 하는 구체적인 단어로 변환하는 단계; 및

상기 프롬프트 템플릿을 이용하여 사전학습 언어 모델(Pre-trained Language Model, PLM)을 학습하여 상기 치명적 오류 감지 모델을 생성하는 단계를 포함하는 치명적 오류 감지 모델 생성 방법.

청구항 2

제1항에 있어서,

상기 레이블에 대응하는 <mask> 토큰을 포함하는 상기 디스크립션은 기계번역 결과에 치명적 오류가 있는지 여부를 알려주는 문장을 나타내는,

치명적 오류 감지 모델 생성 방법.

청구항 3

제2항에 있어서,

상기 치명적 오류 감지 모델을 생성하는 단계 이전에,

상기 학습 데이터로부터 치명적 오류가 있는 적어도 하나의 예제와 치명적 오류가 없는 적어도 하나의 예제를 추출하여 상기 프롬프트 템플릿에 추가하는 단계를 더 포함하는,

치명적 오류 감지 모델 생성 방법.

청구항 4

제2항에 있어서,

상기 치명적 오류 감지 모델을 생성하는 단계 이전에,

임의의 기계번역 모델을 이용하여 생성된 상기 소스 문장에 대한 번역문을 상기 프롬프트 템플릿에 추가하는 단계를 더 포함하는,

치명적 오류 감지 모델 생성 방법.

청구항 5

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 치명적 오류 감지 방법에 있어서,

소스 문장과 번역 문장을 수신하는 단계;

상기 소스 문장과 상기 번역 문장을 이용하여 프롬프트 템플릿을 생성하는 단계; 및

치명적 오류 감지 모델에 상기 프롬프트 템플릿을 입력하여 치명적 오류의 유무를 결정하는 단계를 포함하는 치명적 오류 감지 방법.

청구항 6

제5항에 있어서,

상기 치명적 오류 감지 모델은 제1항에 기재된 방법에 의해 생성된,

치명적 오류 감지 방법.

청구항 7

제5항에 있어서,

상기 프롬프트 템플릿은 상기 소스 문장, 상기 번역 문장, 및 디스크립션을 포함하고,

<mask> 토큰을 포함하는 상기 디스크립션은 기계번역 결과에 치명적 오류가 있는지 여부를 알려주는 문장을 나타내는,

치명적 오류 감지 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공지능 기술 중 자연어 처리 분야에 관한 것으로, 보다 상세하게는 기계번역의 품질 예측 기술에 관한 것이다.

배 경 기 술

[0002] CED(Critical Error Detection, 치명적 오류 감지)는 기계번역 품질 예측 분야(Quality Estimation, QE)의 하위 태스크로, 정답 문장 없이 소스 문장과 번역 문장만을 참고하여 이진 분류를 통해 치명적인 기계번역 오류가 발생한 경우를 탐지(또는 감지)하는 태스크이다. 치명적 번역 오류는 기계번역 시 그 결과가 경제, 법률, 안전 및 윤리, 사회적으로 문제가 될 수 있는 발언으로 변질되는 경우를 의미한다. 예시적으로, 번역 오류로 인한 금융 분야에서의 손실, 위법의 가능성, 잘못된 정보로 인한 건강 측면에서의 손해, 종교나 성, 인종 차별적 발언을 통한 사회적 이슈 조장 등이 있다.

[0003] 치명적 번역 오류는 단순 번역 결과 오류임에도 불구하고 의도치 않은 심각한 사회적 파장 또는 손실을 초래할 수 있고, 그 피해로 인한 책임 또는 보상에 대한 구체적인 대책 및 방안이 마련되어 있지 않기 때문에 반드시 해결해야 할 문제이다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 대한민국 공개특허 제2010-0031800호 (2010.03.25. 공개)

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

발명의 내용

해결하려는 과제

[0005] 종래 기술은 대용량 언어 모델을 활용해 문장 분류를 수행하는 방식으로 기계번역 결과 중 치명적 오류를 탐지(또는 감지)하였다. 그러나 언어 모델의 사전학습 수행 방식은 문장의 입력에 대해 임의로 문장을 <mask> 토큰으로 치환하고 좌우 문맥을 통해 마스킹된 토큰을 예측하는 방식으로 학습되며, 이는 입력 문장들을 통해 이진 분류를 수행하는 종래 기술과 성격이 완전히 다르다. 이는 기존의 학습된 파라미터 값이 완전히 다른 태스크를 학습함으로써 깨지게 되고 언어 모델의 능력을 완전히 활용하지 못한 채 오히려 우수하지 못한 결과를 낼 가능성이 있다.

과제의 해결 수단

[0006] 본 발명의 일 실시예에 따른 치명적 오류 감지 모델 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 치명적 오류 감지 모델 생성 방법으로서, 소스 문장, 번역 문장, 및 치명적 오류의 유무를 나타내

는 레이블(label)을 포함하는 학습 데이터를 소스 문장, 번역 문장, 및 디스크립션(description)을 포함하는 프롬프트 템플릿(prompt template)으로 변환하는 단계, 상기 레이블을 상기 레이블의 의미가 상기 디스크립션 내에서 표현되도록 하는 구체적인 단어로 변환하는 단계, 상기 프롬프트 템플릿을 이용하여 사전학습 언어 모델(Pre-trained Language Model, PLM)을 학습하여 상기 치명적 오류 감지 모델을 생성하는 단계를 포함한다.

[0007] 본 발명의 일 실시예에 따른 치명적 오류 감지 방법 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 치명적 오류 감지 방법으로서, 소스 문장과 번역 문장을 수신하는 단계, 상기 소스 문장과 상기 번역 문장을 이용하여 프롬프트 템플릿을 생성하는 단계, 및 치명적 오류 감지 모델에 상기 프롬프트 템플릿을 입력하여 치명적 오류의 유무를 결정하는 단계를 포함한다.

발명의 효과

[0008] 본 발명은 사전학습 방식과 동일하게 마스킹된 토큰을 예측함으로써 이진 분류를 수행할 수 있도록 프롬프트(prompt) 방식을 활용한다. 이는 언어 모델의 사전학습 방식을 보존하고, 사전학습 시 모델이 학습한 내용을 최대한으로 활용할 수 있는 방식이다. 특히나 데이터가 16, 32, 64개 등 극한으로 적은 상황을 고려한 few-shot setting에서는 종래 기술에서 활용한 미세조정(fine-tuning) 방식보다 월등하게 높은 성능을 보인다.

도면의 간단한 설명

[0009] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 의한 전체 학습 구조도를 도시한다.

도 2는 본 발명에서 사용되는 예시적인 프롬프트 템플릿을 도시한다.

도 3은 본 발명의 일 실시예에 따른 기계번역의 치명적 오류 감지 모델 생성 방법을 설명하기 위한 흐름도이다.

도 4는 도 3에 도시된 치명적 오류 감지 모델 생성 방법에 의해 생성된 치명적 오류 감지 모델을 이용한 치명적 오류 감지 방법을 설명하기 위함 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0010] 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.

[0011] 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.

[0012] 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.

[0013] 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.

[0014] 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.

[0015] 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.

[0016] 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.

[0017] **1. 프롬프트 기반 미세조정(Prompt-based Fine-tuning)**

[0018] 본 발명에서는 미세조정과 사전학습의 학습 목적(training objectives) 사이의 차이를 감소시키기 위하여 프롬프트 기반 미세조정을 이용한다. 프롬프트 기반 미세조정을 적용하여, 사전학습 단계에서 획득한 사전학습 언어 모델의 언어적 능력(linguistic capability)를 보존할 수 있는 CED 모델을 제안한다.

[0019] 본 명세서에서, CED 모델의 학습 데이터셋 D를 (src,mt,y)∈D로 표기한다. 여기서, src와 mt는 각각 소스 문장(source sentence)과 이의 기계번역 문장(translated sentence)를 나타낸다. 그리고, y는 이에 대응하는 레이블(label, 예컨대, OK, BAD)로써, 치명적 오류의 유무를 나타낸다. 즉, y가 OK일 때 기계번역에 치명적 오류가 없고, y가 BAD일 때 기계번역에 치명적 오류가 있는 것으로 볼 수 있다. 레이블을 이루는 OK나 BAD는 실시예에 따라, 다른 단어나 숫자 등으로 표현될 수도 있다. 또한, 본 발명에서는 CED 태스크에서 프롬프트 기반 미세조정을 구현하기 위하여 D 내의 데이터를 변환하는 두 가지의 맵핑 함수 T(제1 맵핑 함수), L(제2 맵핑 함수)을 정의한다.

[0020] 템플릿 함수(template function) T는 src와 mt를 마스크된 토큰(masked token)을 갖는 디스크립션(description)을 포함하는 단일의 입력 시퀀스(single input sequence)로 변환한다. 입력 시퀀스를 생성하는 과정에서, T는 또한 스페셜 <mask> 토큰이 채워지는 위치를 정의한다. 학습 과정에서, 모델이, 전체적인 문맥과 일관성을 갖는, 대응되는 <mask> 토큰 위치에 맞는 적절한 워드를 유추하도록 유도할 수 있다. 여기서, 디스크립션은 제1 디스크립션이라 명명될 수 있고, 기계번역 결과에 치명적 오류가 있는지 없는지를 알려주는 문장을 의미할 수 있다.

[0021] 다음으로, 버벌라이저(verbalizer)로 명명되는 레이블 워드 함수(label word function) L은 주어진 레이블 y를 T를 통해 변환된 입력 시퀀스의 마스크된 위치에 위치할 적절한 레이블 워드로 변환한다. 레이블 워드는 레이블의 의미가 제1 디스크립트 내에서 표현되도록 하는 구체적인 단어를 의미할 수 있다.

[0022] 예컨대, src가 "indigenous peoples constitute just 0.7% of the global population"이고, mt가 "Indigene Völker machen nur 5% der Welt bevölkerung aus"이고, 대응하는 레이블 y가 레이블 워드들 "OK:great, BAD:terrible"를 갖는 BAD일 때, T는 이러한 문장들을 "<s> indigenous peoples constitute just 0.7% of the global population </s> Indigene Völker machen nur 5% der Welt bevölkerung aus </s>. It was <mask> translation."으로 변환하고, L은 이의 레이블을 "terrible"로 변환한다. 그러면, 레이블이 "OK"인지 "BAD"인지를 결정하는 CED의 오리지널 미세조정 목적은 <mask> 토큰 위치에 대한 올바른 워드를 예측하는 것으로 변환된다. 즉, 변환된 레이블 워드가 모델의 출력이 된다. 보다 상세하게, 모델은 수학식 1의 확률을 예측하도록 학습된다.

[0023] [수학식 1]

[0024] $$P(y|src, mt) = P(\langle mask \rangle = L(y)|T(src, mt))$$

[0025] WMT22 CED 태스크의 스코어링 방법을 고려하여, 모델의 추론 결과를 OK 또는 BAD 태그로 이진화하지 않았다. 대신에, 수학식 2와 같이 전체 스코어를 정규화기 위해 소프트맥스 함수(softmax function)를 이용하고, <mask> 위치 내의 디코딩된 토큰이 BAD에 맵핑될 확률을 추출한다. 이러한 확률은 mt의 평가된 품질 스코어(estimated quality score)로 간주한다. 그러나, 본 발명이 이에 제한되는 것은 아니다. 실시예에 따라, 모델의 추론 결과가 OK 또는 BAD 태그로 이진화하는 것도 가능하다.

[0026] [수학식 2]

[0027] $$\text{score}(src, mt) = \frac{\exp(P(\text{BAD}|src, mt))}{\sum_{y' \in \{\text{OK},\text{BAD}\}} \exp(P(y'|src, mt))}$$

[0028] 2. Prompt and Answer Engineering

[0029] 도 1에는 본 발명의 일 실시예에 의한 전체 학습 구조도가 도시되어 있고, 도 2에는 본 발명에서 사용되는 예시적인 프롬프트 템플릿이 도시되어 있다. CED 태스크를 위한 효율적인 프롬프트는 알려지지 않았기 때문에, 본 발명에서는 다양한 프롬프트 후보들을 디자인하였다. 모델 입력을 "src mt. It was <mask> translation, A <mask> translation of src is mt"와 같이 자연스런 문맥(natural context)으로 조직화하려 시도하였다. 레이블 워드들(label words)에 관하여, "grat/terrible"와 "good/bad"와 같은 두 개의 구별되는 워드들을 선택하였다. 명확하게 대조적인 레이블 워드들을 설정함으로써, 나이브한 오류들(naive errors)이 훌륭한 번역 결과로 고려되지는 않지만, 모델 학습 과정에서 애매성(ambiguity)를 제거하고자 하였다. (프롬프트) 템플릿들과 레이블 워드들의 모든 타입은 표 1에 표시되어 있다.

[0030] [표 1]

Template
<s> **src** </s> **mt**. <mask> translation.</s>
<s> **src** </s> **mt**. It was <mask> translation.</s>
<s> A <mask> translation of **src** is **mt**.</s>
<s> **src** </s> **mt** <mask></s>
<s> **src** </s> **mt**? <mask></s>
<s> **src** </s> **mt**? <mask>,</s>
<s> **src** </s> **mt**? "<mask>"</s>
Label Words
OK: "great", BAD: "terrible"
OK: "good", BAD: "bad"
OK: "!", BAD: "?"
OK: "nice", BAD: "poor"
OK: "yes", BAD: "no"

[0031]

[0032] 3. 보조 디스크립션(Auxiliary Description)

[0033] 본 발명에서는 모델 입력에 추가 정보(supplementary information)를 제공하기 위하여 보조 디스크립션들(auxiliary descriptions, 제2 디스크립션이라 명명될 수 있음)을 추가할 수 있다. 두 가지 타입의 보조 디스크립션들, 즉 데모(demonstration)와 기계번역 결과들(예컨대, 구글 번역 결과들) 중 적어도 하나가 선택될 수 있다.

[0034] 데모는, GPT3와 LM-BFF에서 제안된 in-context learning approach와 유사하게 학습 데이터로부터 각 클래스를 위한 예제를 추출하고 추출된 예제를 입력 시퀀스에 결합한다(concatenates). LM-BFF와는 대조적으로, 본 발명에서는, 의미적 유사성(semantic similarity)에 기초하여 데모들을 추출할 때 발생할 수 있는 의도치 않은 바이어스(unintended bias)를 피하기 위해서, 샘플링에 대한 어떠한 제약 조건 없이, 랜덤하게 학습 예제들(training examples)을 선택한다.

[0035] 예시적으로, 추출된 예제는 레이블이 OK인(즉, 치명적 오류가 존재하지 않는) 적어도 하나의 예제와 레이블이 BAD인(즉, 치명적 오류가 존재하는) 적어도 하나의 예제를 포함할 수 있다. 추가적으로, 레이블이 OK인 적어도 하나의 예제에 치명적 오류가 존재하지 않음을 의미하는 디스크립션과 레이블이 BAD인 적어도 하나의 예제에 치명적 오류가 존재함을 의미하는 디스크립션이 더 포함될 수 있다.

[0036] 구글 번역 결과들은 상용화된 MT(Machine Translation) 시스템으로부터의 번역 결과들을 덧붙인다. 이전의 연구들에서도 증명되었듯이, 구글 번역 결과들(임의의 번역 모델에 의한 번역 결과들)을 부가하는 것은 성능 개선에 기여할 수 있다. 이를 고려하여, 본 발명에서는 mt′ 을 생성하기 위해, 전체 데이터 내의 각 src를 번역함으로

써, 임의의 기계번역 모델(예컨대, 구글 번역(Google Translate))을 이용한다. 이를 입력 시퀀스에 추가함으로써, 외부 MT 시스템의 지식을 제안 모델로 증류(distill)할 수 있다.

[0037] 새로운 입력 시퀀스를 구성하기 위하여, 보조 디스크립션들은 D 내의 각 예제와 결합된다. 이를 통하여, 모델이 치명적 오류를 결정하도록 유도할 수 있다.

[0038] **4. Prompt Ensemble**

[0039] 프롬프트 기반 미세조정은 디자인된 프로프트들에 따라 모델 성능에서 편차를 보인다. 본 발명에서는 멀티플 프롬프트들로부터의 결과들을 집계함으로써 성능을 높이고자 한다. 앙상블을 위해, 높은 MCC(Matthew's correlation coefficient) 결과를 갖는 상위 K 값들을 더할 수 있다.

[0040] 도 3은 본 발명의 일 실시예에 따른 기계번역의 치명적 오류 감지 모델 생성 방법을 설명하기 위한 흐름도이다.

[0041] 치명적 오류 감지 방법으로 명명될 수도 있는 치명적 오류 감지 모델 생성 방법은 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 따라서, 치명적 오류 감지 모델 생성 방법의 단계들 중 적어도 일부는 컴퓨팅 장치의 프로세서에 의한 동작으로 이해될 수 있고, 컴퓨팅 장치는 치명적 오류 감지 모델 생성 장치 또는 치명적 오류 감지 장치 등으로 명명될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 태블릿 PC, 노트북, 스마트폰(smart phone) 등을 포함한다.

[0042] 우선, 학습 데이터에 대한 제1 변환이 수행된다(S110). 제1 변환은 템플릿 함수에 의한 (프롬프트) 템플릿 생성 과정으로 이해될 수 있다. 구체적으로, 템플릿 함수는 학습 데이터를 이루는 src(소스 문장), 번역 문장(mt), 및 레이블(y)로부터 src, mt, 및 제1 디스크립션을 포함하는 단일의 입력 시퀀스를 생성한다.

[0043] 다음으로, 학습 데이터에 대한 제2 변환이 수행된다(S120). 제2 변환은 레이블 워드 함수에 의한 동작으로 이해될 수 있다. 구체적으로, 레이블 원드 함수는 주어진 레이블 y를 T를 통해 변환된 입력 시퀀스의 마스크된 위치에 위치할 적절한(미리 정해진) 레이블 워드로 변환한다.

[0044] 상술한 과정을 통해, 레이블이 "OK"인지 "BAD"인지를 결정하는 CED의 오리지널 미세조정의 목적은 <mask> 토큰 위치에 대한 올바른 워드를 예측하는 것으로 변환된다.

[0045] 실시예에 따라, 적어도 하나의 보조 디스크립션이 (프롬프트) 템플릿에 부가되어 최종적인 모델 입력이 생성될 수 있다(S130). 다만, 본 단계는 생략이 가능하며, 추가되는 보조 디스크립션은 실시예에 따라 상이할 수 있다.

[0046] 생성된 모델 입력을 이용하여 사전학습 언어 모델(Pre-trained Language Model)을 학습(미세조정)시킴으로써, 치명적 오류 감지 모델이 생성될 수 있다(S140).

[0047] 추가적으로, 생성된 치명적 오류 감지 모델을 이용하여 주어진 소스 문장과 기계번역 문장에 대한 치명적 오류 유무를 감지(또는 탐지)하는 동작이 수행될 수도 있다.

[0048] 도 4는 도 3에 도시된 치명적 오류 감지 모델 생성 방법에 의해 생성된 치명적 오류 감지 모델을 이용한 치명적 오류 감지 방법을 설명하기 위함 흐름도이다.

[0049] 치명적 오류 감지 방법은 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 따라서, 치명적 오류 감지 방법의 단계들 중 적어도 일부는 컴퓨팅 장치의 프로세서에 의한 동작으로 이해될 수 있고, 컴퓨팅 장치는 치명적 오류 감지 장치 등으로 명명될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 태블릿 PC, 노트북, 스마트폰(smart phone) 등을 포함한다.

[0050] 소스 문장과 번역 문장이 수신될 수 있다(S210). 컴퓨팅 장치는 소정의 유무선 통신 프로토콜을 통하여 외부 장치(단말이나 서버)로부터 판단 대상인 소스 문장과 번역 문장을 수신할 수 있다. 실시예에 따라, 컴퓨팅 장치는 소정의 입출력 인터페이스를 이용하여 사용자로부터 소스 문장과 번역 문장을 직접 수신할 수도 있다.

[0051] 다른 실시예로, 소정의 기계번역 모델이 저장되어 있는 컴퓨팅 장치는 사용자로부터 또는 외부의 장치(단말이나 서버)로부터 소스 문장만을 수신하고, 기계번역 모델을 이용하여 소스 문장에 대한 번역 문장을 생성하는 것도 가능하다. 즉, 이 경우는, 컴퓨팅 장치가 번역 문장의 생성과 번역 문장에 대한 치명적 오류 감지 동작을 모두 수행하는 경우로 이해될 수 있다.

[0052] 다음으로, 소스 문장과 번역 문장을 이용하여 모델의 입력을 생성할 수 있다(S220). 모델의 입력은 (프롬프트) 템플릿을 의미할 수 있고, 도 2에 도시된 다양한 템플릿들 중 어느 하나가 이용될 수 있다. 모델의 입력을 생성하는 구체적인 과정은 상술하였기에, 이에 관한 구체적인 설명은 생략하기로 한다.

[0053] 마지막으로, 생성된 모델 입력을 치명적 오류 감지 모델에 입력하여, 소스 문장과 번역 문장에 대한 치명적 오류 유무를 감지할 수 있다(S230).

[0054] 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.

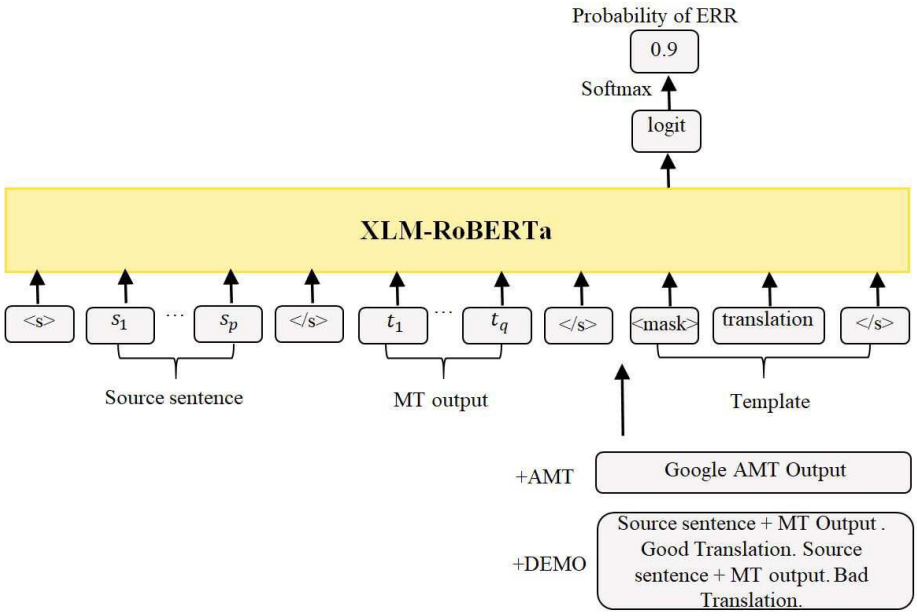
[0055] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 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.

[0056] 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.

[0057] 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.

도면

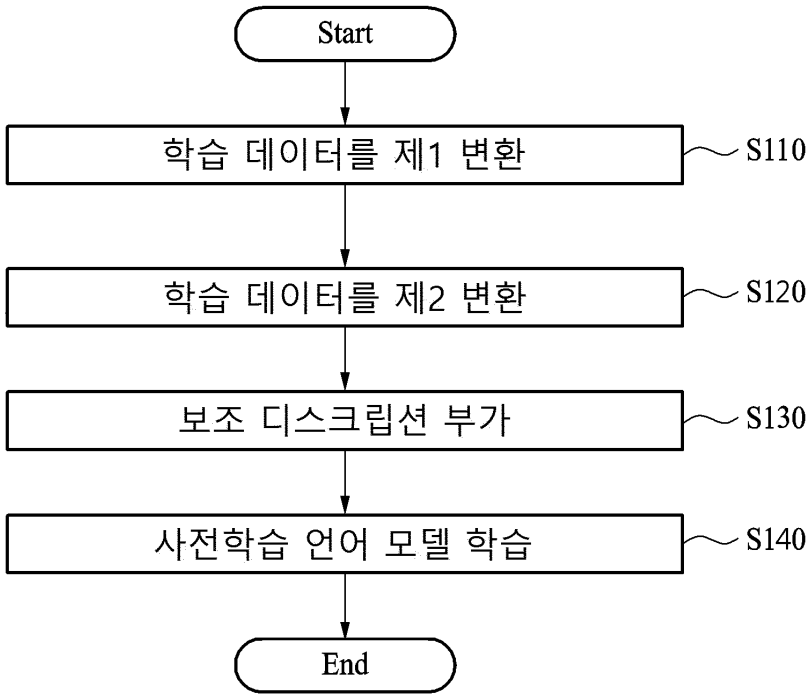
도면1


Probability of ERR
0.9
Softmax
logit
XLM-RoBERTa
<s>
s_1
s_p
</s>
t_1
t_q
</s>
<mask>
translation
</s>
Source sentence
MT output
Template
+AMT
Google AMT Output
+DEMO
Source sentence + MT Output . Good Translation. Source sentence + MT output. Bad Translation.

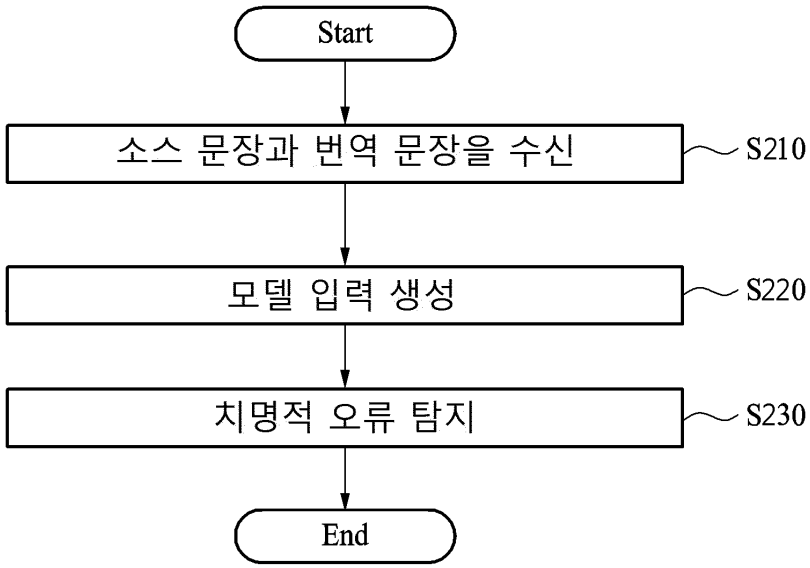
도면2

Type	Index	Template	Label Words	MCC (Binary)	MCC	P&R
Plain Template	1	**src mt**. <mask>translation.	great / terrible	0.9156	0.9153	0.9200
	2	**src mt**. <mask>translation.	good / bad	0.9158	0.9087	0.9137
	3	**src mt** <mask>	! / ?	0.9153	0.9131	0.9179
	4	A <mask>translation of **src** is **mt**.	good / bad	0.9189	0.9142	0.9189
	5	A <mask>translation of **src** is **mt**.	great / terrible	0.9161	0.9131	0.9179
	6	**src mt**. It was <mask>translation.	great / terrible	0.9124	0.9042	0.9095
	7	**src mt**. It was <mask>translation.	nice / poor	0.9161	0.9131	0.9179
	8	**src mt**? <mask>	yes / no	0.9184	***0.9120***	0.9168
Template with Demo	1	**demo_ok demo_bad srcmt**. <mask>translation.	great / terrible	0.9149	0.9064	0.9116
	2	**demo_ok demo_bad src mt**. <mask>translation.	good / bad	0.9089	0.9098	0.9147
	3	**demo_ok demo_bad src mt**. It was <mask>translation.	great / terrible	0.9125	0.9064	0.9116
	4	**demo_ok demo_bad src mt**. It was <mask>translation.	nice / poor	0.9183	**0.9187**	0.9232
	5	**demo_ok demo_bad src mt**. It was <mask>translation.	! / ?	0.9084	0.9042	0.9095
	6	**demo_ok demo_bad src mt**? <mask>	yes / no	0.9109	0.9075	0.9126
	7	**src mt demo_ok demo_bad** <mask>translation.	great / terrible	0.9095	0.8964	0.9021
	8	**src mt demo_ok demo_bad** <mask>translation.	good / bad	0.9060	0.9042	0.9095
	9	**src mt demo_ok demo_bad** . It was <mask>translation.	great / terrible	0.9123	0.9064	0.9116
	10	**src mt demo_ok demo_bad** . It was <mask>translation.	nice / poor	0.9138	0.9098	0.9147
	11	**src mt demo_ok demo_bad** ? <mask>	yes / no	0.9175	0.9098	0.9147
Template with Google Translate	1	**src mt**? <mask>**gmt**	great / terrible	0.9161	0.9098	0.9147
	2	**src mt**? <mask>**gmt**	good / bad	0.9198	**0.9165**	0.9211
	3	**src mt**? <mask>**gmt**	! / ?	0.9218	**0.9165**	0.9211
	4	**src mt**? <mask>**gmt**	yes / no	0.9121	0.9087	0.9137
	5	**src mt**? It was <mask>. **gmt**	great / terrible	0.9173	0.9053	0.9105
	6	**src mt**? It was <mask>. **gmt**	good / bad	0.9166	0.9087	0.9137
	7	**src mt**? It was <mask>. **gmt**	! / ?	0.9172	0.9153	0.9200
	8	**src mt**? It was <mask>. **gmt**	yes / no	0.9158	***0.9120***	0.9168
	9	**src mt**? "<mask>", **gmt**	! / ?	0.9176	***0.9120***	0.9168
	10	**src mt**? "<mask>", **gmt**	good / bad	0.9175	0.9064	0.9116
	11	**src mt**? <mask>, **gmt**	! / ?	0.9103	0.9087	0.9137
	12	**src mt**? <mask>, **gmt**	good / bad	0.9111	0.9087	0.9137
	13	**src mt gmt**. <mask>translation.	great / terrible	0.9133	0.9009	0.9063
	14	**src mt gmt**. <mask>translation.	good / bad	0.8872	~~0.8875~~	0.8937
	15	**src mt gmt**. <mask>	! / ?	0.9151	0.9064	0.9116
	16	A <mask>translation of **src** is **mt gmt**.	good / bad	0.9209	**0.9165**	0.9211
	17	A <mask>translation of **src** is **mt gmt**.	great / terrible	0.9134	***0.9109***	0.9158
	18	**src mt gmt**. It was <mask>translation.	great / terrible	0.9134	0.9075	0.9126
	19	**src mt gmt**. It was <mask>translation.	nice / poor	0.9160	**0.9165**	0.9211
	20	**src mt gmt**? <mask>	yes / no	0.9147	0.9098	0.9147

도면3


Start
학습 데이터를 제1 변환
S110
학습 데이터를 제2 변환
S120
보조 디스크립션 부가
S130
사전학습 언어 모델 학습
S140
End

도면4


Start
소스 문장과 번역 문장을 수신
S210
모델 입력 생성
S220
치명적 오류 탐지
S230
End