

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0077850
(43) 공개일자 2025년06월02일

(51) 국제특허분류(Int. Cl.)

G06F 40/51 (2020.01) G06F 40/58 (2020.01)
G06N 3/0475 (2023.01) G06N 3/0895 (2023.01)
G06N 3/09 (2023.01) G06N 3/096 (2023.01)
G06N 5/04 (2023.01)

(52) CPC특허분류

G06F 40/51 (2020.01)
G06F 40/58 (2020.01)

(21) 출원번호 10-2023-0165307

(22) 출원일자 2023년11월24일

심사청구일자 2023년11월24일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

정다현

경기도 안양시 동안구 동편로 110, 302동 303호

임희석

서울특별시 성북구 안암로 145 자연계캠퍼스 애기
능생활관 311호 NLP&AI 연구실

(74) 대리인

김동용

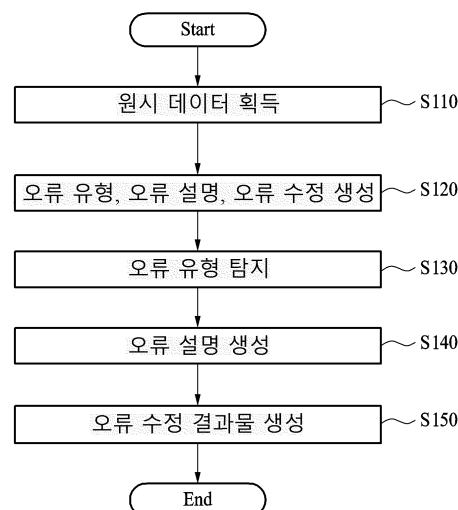
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 기계 번역에서 치명적인 오류 탐지의 설명을 위한 데이터셋 구축 장치 및 방법

(57) 요약

기계 번역(Machine Translation, MT)에서 치명적인 오류(critical errors) 탐지의 설명을 위한 데이터셋 구축 장치 및 방법이 개시된다. 상기 데이터셋 구축 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되고, 소스 문장들, 기계 번역(MT) 문장들, 및 기계 번역 문장들 각각에 치명적인 오류의 포함 여부를 나타내는 레이블을 포함하는 원시 데이터가 획득되는 단계, 어느 하나의 소스 문장에 대응하는 기계 번역(MT) 문장에 포함되는 치명적인 오류의 유형인 오류 유형이 탐지되는 단계, 상기 어느 하나의 소스 문장에 대응하는 기계 번역(MT) 문장에 포함되는 치명적인 오류에 대한 설명을 나타내는 오류 설명이 생성되는 단계, 및 상기 어느 하나의 소스 문장에 대하여, 치명적인 오류가 수정된 기계 번역(MT) 문장을 의미하는 오류 수정이 생성되는 단계를 포함한다.

대표도 - 도3



(52) CPC특허분류

G06N 3/0475 (2023.01)

G06N 3/0895 (2023.01)

G06N 3/09 (2023.01)

G06N 3/096 (2023.01)

G06N 5/045 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1345365402
과제번호	2021R1A6A1A03045425
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공학학술연구기반구축
연구과제명	Human-inspired AI 연구소
기 여 율	1/2
과제수행기관명	고려대학교
연구기간	2023.03.01 ~ 2024.02.29

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193821
과제번호	2020-0-00368-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	차세대인공지능핵심원천기술개발
연구과제명	뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는, 기계 번역(Machine Translation, MT)에서 치명적인 오류(critical errors) 탐지의 설명을 위한 데이터셋 구축 방법에 있어서,

소스 문장들, 기계 번역(MT) 문장들, 및 기계 번역 문장들 각각에 치명적인 오류의 포함 여부를 나타내는 레이블을 포함하는 원시 데이터가 획득되는 단계;

어느 하나의 소스 문장에 대응하는 기계 번역(MT) 문장에 포함되는 치명적인 오류의 유형인 오류 유형이 탐지되는 단계;

상기 어느 하나의 소스 문장에 대응하는 기계 번역(MT) 문장에 포함되는 치명적인 오류에 대한 설명을 나타내는 오류 설명이 생성되는 단계; 및

상기 어느 하나의 소스 문장에 대하여, 치명적인 오류가 수정된 기계 번역(MT) 문장을 의미하는 오류 수정이 생성되는 단계를 포함하는,

데이터셋 구축 방법.

청구항 2

제1항에 있어서,

상기 오류 유형이 탐지되는 단계는, 거대 언어 모델(Large Language Model, LLM)을 이용하고, 치명적인 오류의 정의, 치명적인 오류의 유형 각각에 대한 설명, 및 오류 유형 탐지에 대한 지시를 포함하는 제1 프롬프트를 활용하는,

데이터셋 구축 방법.

청구항 3

제2항에 있어서,

상기 제1 프롬프트는 하기 제1 프롬프트와 같은,

데이터셋 구축 방법.

[제1 프롬프트]

Translations with critical errors are defined as translations that deviate in meaning as compared to the source sentence in such a way that they are misleading and may carry health, safety, legal, reputation, religious, or financial implications.

{Critical Error Category}

Read the translation of the source sentence and perform the three tasks below. Please read the instructions carefully before completing the task.

- Error Type: Please indicate which category the critical error in the translation sentence belongs to. If you find multiple categories of errors, please indicate only the most serious one, and if it does not belong to any category, please indicate "ETC". If there is no error, mark it as "NOT" and do nothing further.

- Error Explanation: Please explain why the error occurred. Please describe the single most serious error from the error category, and be concise in no more than two sentences, including examples.

- Error Correction: Please fix the error in the translation sentence, minimizing it to the part where the critical error mentioned in the description appears.

청구항 4

제2항에 있어서,

상기 제1 프롬프트는 오류 설명에 대한 생성 지시를 더 포함하고,

상기 오류 설명이 생성되는 단계는,

상기 거대 언어 모델(LLM)을 이용하여 생성된 오류 설명들 각각에 대한 스코어링 동작을 수행하는,
데이터셋 구축 방법.

청구항 5

제4항에 있어서,

상기 오류 설명이 생성되는 단계는 생성된 오류 설명들 각각에 대한 스코어를 기준으로 어느 하나의 오류 설명을 선택하는,

데이터셋 구축 방법.

청구항 6

제5항에 있어서,

상기 오류 수정이 생성되는 단계는, 상기 거대 언어 모델(LLM)에 소스 문장, 기계 번역(MT) 문장, 오류 유형, 및 오류 설명을 입력하여 상기 오류 수정을 생성하는,

데이터셋 구축 방법.

발명의 설명

기술 분야

[0001] 본 발명은 기계 번역(Machine Transslation, MT) 분야에서 품질 검증 방법 중 하나인 치명적인 오류 탐지 작업의 설명 가능성을 높이기 위한 데이터 구축 장치 및 방법에 관한 것이다.

배경 기술

[0002] 치명적인 오류 탐지(Critical Error Detection, CED)는 치명적인 오류를 포함하고 있는 번역 문장을 탐지하는 기계 번역 품질 검증(Quality Estimation) 분야의 하위 작업이다. 치명적인 오류란 기계 번역의 결과물이 원본 문장의 의미에서 벗어나 건강, 안전, 법률, 종교 또는 기타 측면과 관련된 우려를 불러일으키는 문제를 말한다. 원본 문장에서 의미의 편차는 오역이나 환각, 삭제의 형태로 나타날 수 있다. 오역은 원본 문장이 잘못 번역되어 원래의 의미가 왜곡되는 오류를 말하며, 환각은 번역 문장 중에 원본 문장에 없는 내용이 도입되는 것을 말한다. 또한, 삭제는 원본 문장에 있는 내용이 번역 과정에서 생략되는 경우에 발생한다.

[0003] 치명적인 오류 탐지(CED)는 번역 문장에 포함된 오류의 유무를 나타내는 이진 레이블을 분류하는 데 중점을 둔다. 이러한 이진 분류의 한계는 번역 문장을 오류가 존재한다고 분류하는 딥러닝 모델이 블랙박스라는 점이다. 오류를 분류하는 딥러닝 모델은 오류의 유무만을 판별할 뿐, 오류의 원인을 제시하지 못한다. 따라서, 이러한 한계를 해결하고자 치명적인 오류의 원인을 이해할 수 있는 데이터를 제공할 수 있는 데이터셋을 구축하는 작업이 중요하다.

선행기술문헌

비특허문헌

[0004] (비특허문헌 0001) Lucia Specia, Frederic Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and Andre F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In Proceedings of the Sixth Conference on Machine Translation, pages 684-725, Online. Association for Computational Linguistics.

(비특허문헌 0002) Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

(비특허문헌 0003) Yau-Shian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings.

(비특허문헌 0004) Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

발명의 내용

해결하려는 과제

[0005] 기존의 치명적인 오류 탐지(CED) 작업이 블랙박스로 진행된다는 한계를 지적하며, 기계 번역 시스템에서 발생한 치명적인 오류의 원인 및 해결 방안을 제시해 주는 언어 모델을 제작하기 위한 데이터셋을 구축하는 방법론을 제공하는 것이다.

과제의 해결 수단

[0006] 본 발명의 일 실시예에 따른, 기계 번역(Machine Translation, MT)에서 치명적인 오류(critical errors) 탐지의 설명을 위한 데이터셋 구축 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되고, 소스 문장들, 기계 번역(MT) 문장들, 및 기계 번역 문장들 각각에 치명적인 오류의 포함 여부를 나타내는 레이블을 포함하는 원시 데이터가 획득되는 단계, 어느 하나의 소스 문장에 대응하는 기계 번역(MT) 문장에 포함되는 치명적인 오류의 유형인 오류 유형이 탐지되는 단계, 상기 어느 하나의 소스 문장에 대응하는 기계 번역(MT) 문장에 포함되는 치명적인 오류에 대한 설명을 나타내는 오류 설명이 생성되는 단계, 및 상기 어느 하나의 소스 문장에 대하여, 치명적인 오류가 수정된 기계 번역(MT) 문장을 의미하는 오류 수정이 생성되는 단계를 포함한다.

발명의 효과

[0007] 본 발명의 실시예들에 의할 경우, 다양한 언어들의 치명적인 오류 탐지 작업을 설명하기 위한 데이터셋을 구축할 수 있다.

[0008] 또한, 구축된 데이터셋을 활용하여 설명 가능한 치명적인 오류 탐지가 가능한 언어 모델을 생성할 수 있다.

도면의 간단한 설명

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

도 1은 본 발명에서 제안하는 데이터셋 구축 과정에 사용하기 위한 예시적인 지침(프롬프트)을 도시한다.

도 2는 본 발명에서 제안하는 데이터셋의 품질을 검증하기 위한 예시적인 지침(프롬프트)을 도시한다.

도 3은 본 발명의 일 실시예에 따른, 기계 번역에서 치명적인 오류 탐지의 설명을 위한 데이터셋 구축 방법을 설명하기 위한 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" 있거나 "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] 본 발명에서 제안하는 설명 가능한 치명적인 오류 탐지 데이터셋(Explainable CED dataset)은 영어-독일어(English-German, En-De), 영어-체코어(English-Czech, En-Cs), 영어-중국어(English-Chinese, En-Zh), 및 영어-일본어(English-Japanese, En-Ja) 언어 쌍들을 포함한다. 그러나, 본 발명의 권리범위가 데이터셋을 이루는 언어의 종류에 제안되는 것은 아니며, 실시예에 따라 다양한 언어로 구성된 언어 쌍을 포함할 수 있다. 이에 대한 데이터 통계는 표 1에 나타나 있다.

[0018] [표 1]

	Error Type						Error Explanation		Error Correction	
	TOX	SAF	NAM	SEN	NUM	ETC	Sentences	Tokens	Sentences	Tokens
En-De	2,096	520	3,793	512	1,676	76	8,673	382,941	8,673	294,239
En-Cs	861	10	533	38	4	4	1,450	66,175	1,450	30,274
En-Zh	559	11	611	47	9	6	1,243	64,565	1,243	23,545
En-Ja	444	14	263	33	7	10	771	31,172	771	12,895

[0019]

[0020] 표 1은 4 개의 언어 쌍들에 대한 데이터셋의 통계를 나타낸다. Error Type(오류 유형)은 각 카테고리에 대한 인스턴스의 개수를 나타낸다. Error Explanation(오류 설명) 및 Error Correction(오류 수정)은 문장과 토큰의 개수를 나타낸다.

[0021] Dataset Scheme

[0022] 표 2는 소스 문장(source sentence) 및 오류를 포함하는 기계 번역 문장(MT sentence)의 예시를 보여준다. 데이터셋은 오류 유형(Error Type), 오류 설명(Error Explanation), 및 오류 수정(Error Correction)을 포함하도록 구성된다.

[0023] [표 2]

SRC	This time the world united, even booing the lead negotiator until she reversed position and agreed to sign the Bali Action Plan.
MT	Dieses Mal hat sich die Welt zusammengeschlossen und sogar die Verhandlungsführerin ausgebuht, bis sie ihre Haltung änderte und einwilligte, den Aktionsplan von ** zu unterzeichnen. (This time the world united and even booed the negotiator until she changed her stance and agreed to sign **'s action plan.)
Error Type	TOX
Error Explanation	In the translation sentence, the original "Bali Action Plan" is mistranslated as "den Aktionsplan von **" (**'s action plan). This points to a different named entity, which gives the false impression that the plan is associated with **.
Error Correction	Dieses Mal hat sich die Welt zusammengeschlossen und sogar die Verhandlungsführerin ausgebuht, bis sie ihre Haltung änderte und dem Bali-Aktionsplan zustimmte. (This time the world united and even booed the negotiator until she changed her stance and agreed to the Bali Action Plan.)

[0024]

[0025] 오류 유형(Error Type)은 치명적인 오류의 범주로 태그된 레이블(label)을 나타낸다. 치명적인 오류는 독성(Toxicity, TOX), 안전성(Safety, SAF), 명명된 개체(Named entity, NAM), 감정(Sentiment, SEN), 수(Number, NUM), 및 기타(ETC)의 6가지 유형으로 분류된다. 다만, 구체적인 오류 유형이나 오류 유형의 개수는, 실시예에 따라, 상이하게 변경될 수 있다.

[0026] 예시적으로, Specia 등(Lucia Specia, Frederic Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and Andre F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In Proceedings of the Sixth Conference on Machine Translation, pages 684-725, Online. Association for Computational Linguistics.)에 의해 정의된 카테고리가 활용될 수 있다.

[0027] 독성(Toxicity, TOX): 개인(personal), 인종(racial), 또는 성별(gender)에 기반한 혐오(hate)와 모욕적인 언어(offensive language)와 관련된 오류가 관찰된다. 이러한 오류는, 소스 문장에 독성이 없었을 때 번역 문장에 독성이 추가되거나, 또는 소스 문장에 독성이 있었을 때 번역 문장에 독성이 제거되는 형태로 나타난다.

[0028] 안전성(Safety, SAF): 소스 문장과 번역 문장 간의 불일치(discrepancy)는 번역 오류를 구성하여 독자에게 잠재적인 안전 위험을 야기할 수 있다. 이러한 오류는 소스 문장에 없는 내용이 번역에 도입되거나 소스 문장에서 내용이 번역에서 생략될 때 발생할 수 있다.

[0029] 명명된 개체(Named entity, NAM): 이 오류는 명명된 개체(예컨대, 사람, 조직, 또는 위치)가 오역되거나 생략되거나 타겟 문장에 번역되지 않을 때 발생한다. 용어(term)가 사용자의 이름으로 결정될 수 있다면, 이는 명명된 개체 오류로 간주된다. 부분적으로 번역된 명명된 개체는, 동일한 개체를 나타낸다고 이해할 수 있다면, 치명적인 오류로 간주되지 않는다.

[0030] 감정(Sentiment, SEN): 이 오류는 문장의 감정이 반전될 때 발생한다. 그러나, 감정 오류는 반드시 완전한 부정을 나타내는 것을 아니다. 예컨대, "possibly"를 "with certainty"로 변경하는 것은 감정 오류로 간주된다.

[0031] 수(Number, NUM): 이 오류는 숫자(즉, 시간 또는 날짜)와 관련이 있다. 이러한 오류는 오역된 숫자 또는 소스 문장에 있는 숫자가 번역 문장에서 생략된 형태로 나타난다.

[0032] 기타(ET Cetera, ETC): 이 유형의 오류는 위의 다섯 범주 중 어느 하나에도 속하지 않지만 원래 텍스트의 의미를 심각하게 훼손한다.

- [0033] 오류 설명(Error Explanation)은 기계 번역(MT) 문장의 오류에 대해 상세하게 설명하는 자연어 설명을 나타낸다. 여기에는, 문장의 어느 부분에 번역 오류가 포함되어 있는지를 나타내는 명시적인 인스턴스(explicit instances)가 포함될 수 있다. 이러한 설명은 레이블이 지정된 인스턴스 외에도 문제의 원인과 특성에 대한 깊은 이해를 제공하여 오류의 심각성에 대한 인식을 높일 수 있다.
- [0034] 오류 수정(Error Correction)은 오류가 수정된 번역 문장을 나타낸다. 기계 번역 시스템은 주어진 원본 문장과 의미적으로 동일한 번역을 제공하여야 한다. 본 발명에서는, 오류 수정된 기계 번역(MT) 결과를 제시함으로써 이 목표를 달성하고자 한다. 또한, 문장 내에서 수정 사항을 시연함으로써, 기계 번역(MT) 시스템의 투명성과 신뢰성을 향상시키고자 한다.
- [0035] **데이터 수집(Data Collection)**
- [0036] 본 발명에서는 2021년과 2022년 기계 번역 컨퍼런스(conference on machine translation, WMT)에서 공개된 치명적인 오류 탐지(CED) 데이터셋으로부터의 치명적인 오류 존재 여부가 레이블링된 문장들(critical error presence-labeled sentences)를 소스(source)로 활용한다. 그러나, 본 발명의 권리범위가 사용되는 데이터셋에 제한되는 것은 아니며, 소스 문장, 기계 번역(MT) 문장, 정답 문장, 및 치명적인 오류의 존재 유무(레이블이나 식별자를 의미할 수 있음) 중 적어도 하나를 포함하는 임의의 데이터셋이 이용될 수 있다. 또한, 본 발명에서는, 수많은 오류가 존재할 경우, 가장 치명적인 오류에만 집중하여 설명을 제공하고자 한다.
- [0037] 본 발명에서는, 데이터셋을 구축하기 위하여, ChatGPT(OpenAI-Blog, 2022)(gpt3.5-turbo-0301)과 같은 거대 언어 모델(Large Language Model, LLM)을 활용한다. 거대 언어 모델(LLM)은 수많은 파라미터를 보유한 인공 신경망으로 구성되는 언어 모델이다. 자기 지도 학습이나 반 자기 지도 학습을 사용하여 레이블링되지 않은 상당한 양의 텍스트로 학습된다. 데이터셋 구축 방법은 체계적인 단계별 프로세스를 통해 세 가지 체계로 구성된다. 이 방법은 데이터 인스턴스 간의 정렬(alignment)을 지속적으로 고려하도록 설계되었으며, 거대 언어 모델(LLM)로 인해 발생하는 환각 문제를 완화하기 위해 여러 번의 검증을 수행한다.
- [0038] (1) 치명적인 오류 유형 탐지(Detect Critical Error Type): 다양한 프롬프트에서 발생하는 불일치를 최소화하고 데이터셋의 신뢰성을 높이기 위해 단일 입력에 대한 여러 결과 간의 일치도를 측정한다. 거대 언어 모델에 원본 문장과 기계 번역(MT) 문장을 모두 입력으로 제공하고 세 가지 프롬프트를 적용하여 오류 유형을 생성(또는 추론)한다. 예시적인 프롬프트는 도 1에 도시되어 있다. 실시예에 따라, 복수의 프롬프트를 적용하여 복수의 결과들이 도출될 수 있고, 하나의 프롬프트를 적용하여 복수의 결과들이 도출될 수도 있다.
- [0039] 프롬프트는 치명적인 오류의 정의, 치명적인 오류의 유형 각각에 대한 설명, 여러 유형 탐지에 대한 지시, 여러 설명의 생성에 대한 지시, 및 여러 수정에 대한 지시 중 적어도 하나를 포함할 수 있다.
- [0040] 이러한 결과를 비교하여 과반수 동의가 있는 유형을 식별하고 동의가 없는 데이터는 제외한다. 예를 들어, 서로 다른 프롬프트가 주어진 모델의 결과가 (TOX, NUM, NUM)인 경우, NUM 유형으로 선택한다. 실시예에 따라, 출력들 중에서 과반수 이상이 동일한 오류 유형을 나타낼 경우 해당 오류 유형을 기계 번역(MT) 문장의 오류 유형으로 결정할 수 있다. 다른 예로, 출력들 중에서 가장 많은 오류 유형을 기계 번역(MT) 문장의 오류 유형으로 결정할 수도 있다.
- [0041] (2) 치명적인 오류 설명 생성(Generate Critical Error Explanation): 이 단계에서는 환각을 완화하고 설명의 품질을 향상시키기 위해 세 가지 방법을 사용할 수 있다.
- [0042] 첫째, 원본 문장 및 기계 번역(MT) 문장뿐만 아니라 이전 단계에서 생성된 유형 정보도 제공하여 프롬프트를 구성한다. 이러한 접근 방식을 통해 특정 오류 유형에 맞는 설명을 생성할 수 있어 스키마 간의 정합성을 높일 수 있다.
- [0043] 둘째로, 생성 품질을 개선하기 위해 자체 수정 접근 방식(Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.)을 적용할 수 있다. 거대 언어 모델(LLM)은 도 2와 같은 프롬프트에 따라 초기에 생성된 문장에 대한 피드백을 생성한다. 도 2의 프롬프트는 자체 수정 방법과 GPT-3.5 스코어를 평가하기 위해 이용될 수 있다.
- [0044] 실시예에 따라, 프롬프트는 치명적인 오류의 정의, 오류 설명의 정의, 오류 수정의 정의, 오류 설명의 스코어링 기준, 오류 수정의 스코어링 기준, 오류 설명의 스코어링에 대한 지시, 및 오류 수정의 스코어링에 대한 지시

중 적어도 하나를 포함할 수 있다. 이 피드백을 모델에 입력으로 통합하여 모델이 출력을 반영하고 수정하여 결과를 개선할 수 있도록 한다.

[0045] 셋째, 서로 다른 프롬프트를 사용하여 생성 및 수정된 두 개의 설명 문장을 비교하여 차이를 최소화한다. 최종 설명 문장은 두 결과물 간의 유사도 점수와 모델 선호도 점수가 모두 상위 a (a 는 임의의 실수로 예시적인 값은 20임)%에 해당하는 문장 중 모델 선호도 점수가 가장 높은 문장을 기준으로 선정한다. 이때, 유사도 점수는 mSimCSE(Yau-Shian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings.)를 사용하여 측정하고, 모델 선호도 점수는 GPT-3.5 스코어(Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.)를 사용하여 평가할 수 있다.

[0046] (3) 치명적인 오류 수정 결과물 생성 (Generate Critical Error Correction): 오류 수정된 번역 문장을 생성하는 과정은 2 단계와 유사하게 수행될 수 있다. 이전 단계에서 생성된 유형과 설명을 고려하여 적어도 하나의(또는 서로 다른) 프롬프트를 사용하여 적어도 하나(또는 복수(예컨대, 두 개))의 문장을 생성할 수 있다. 자체 수정이 수행되고 개선된 문장을 비교하여 최종 수정이 이루어진다.

[0047] 도 3은 본 발명의 일 실시예에 따른, 기계 번역에서 치명적인 오류 탐지의 설명을 위한 데이터셋 구축 방법을 설명하기 위한 흐름도이다. 데이터셋 구축 방법은 적어도 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 따라서, 데이터셋 구축 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(Server), 랩탑 컴퓨터, 태블릿 PC 등을 포함할 수 있고, 물리적으로 구별된 복수의 장치로 구현될 수도 있다. 이하에서, 데이터셋 구축 방법을 설명함에 있어, 앞선 기재와 중복되는 내용에 관하여는 그 구체적인 설명을 생략하기로 한다.

[0048] 우선, 원시 데이터가 획득된다(S110). 원시 데이터는 소스 문장들, 소스 문장들 각각에 대한 기계 번역(MT) 문장들, 및 기계 번역(MT) 문장들 각각에 대한 치명적인 오류의 유무에 대한 레이블을 포함할 수 있다. 실시예에 따라, 원시 데이터에는 정답 문장을 더 포함할 수도 있다. 원시 데이터는 소정의 유무선 통신망을 통하여 원시 데이터 제공 장치(서버)로부터 수신되거나, USB 메모리 장치와 같은 저장 장치로부터 수신되거나, 컴퓨팅 장치에 미리 저장되어 있을 수 있다.

[0049] 기계 번역(MT) 문장에 포함된 치명적인 오류의 유형, 치명적인 오류에 대한 설명인 오류 설명, 및 치명적인 오류가 수정된 번역문을 의미하는 오류 수정이 생성된다(S120). S120 단계는 거대 언어 모델(LLM)을 이용하여 수행될 수 있다. 이를 위해, 예컨대, 도 1에 도시된 프롬프트, 소스 문장, 기계 번역 문장, 및 치명적인 오류의 유무에 대한 레이블 중 적어도 하나가 거대 언어 모델(LLM)에 입력될 수 있다.

[0050] 오류 유형이 탐지된다(S130). 오류 유형은 소스 문장에 대한, 거대 언어 모델의 출력들(오류 유형을 나타냄) 중 어느 하나의 오류 유형으로 결정될 수 있다. 이때, 가장 많은 오류 유형이 선택되거나, 과반수 이상의 오류 유형이 선택될 수 있다. 선택 기준에 부합되지 않은 경우, S120 단계가 다시 수행될 수 있다.

[0051] 오류 설명이 생성된다(S140). 오류 설명은, 예시적으로 도 2의 프롬프트를 이용하여 수행될 수 있으며, 이에 따라 거대 언어 모델(LLM)은 오류 설명에 대한 스코어링 및/또는 오류 수정에 대한 스코어링 동작을 수행할 수 있다. 이에 따라, 복수의 오류 설명 중 어느 하나가 선택될 수 있다.

[0052] 오류 수정 결과물이 생성된다(S150). 거대 언어 모델(LLM)에 오류 수정에 대한 지식을 포함하는 프롬프트, 소스 문장, 기계 번역(MT) 문장, 오류 유형, 및 오류 설명 중 적어도 하나를 입력하여 적어도 하나의 오류 수정 결과물이 생성된다.

[0053] 상술한 과정을 통해 생성된, 소스 문장, 기계 번역(MT) 문장, 치명적인 오류 유무, 오류 유형, 오류 설명, 및 오류 수정 결과물을 포함하는 데이터셋이 생성될 수 있다.

[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, 컨트롤러, 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 응

답하여, 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 컨트롤러를 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

Translations with critical errors are defined as translations that deviate in meaning as compared to the source sentence in such a way that they are misleading and may carry health, safety, legal, reputation, religious, or financial implications.
{Critical Error Category}
Read the translation of the source sentence and perform the three tasks below. Please read the instructions carefully before completing the task.
- Error Type: Please indicate which category the critical error in the translation sentence belongs to. If you find multiple categories of errors, please indicate only the most serious one, and if it does not belong to any category, please indicate "ETC". If there is no error, mark it as "NOT" and do nothing further.
- Error Explanation: Please explain why the error occurred. Please describe the single most serious error from the error category, and be concise in no more than two sentences, including examples.
- Error Correction: Please fix the error in the translation sentence, minimizing it to the part where the critical error mentioned in the description appears.

도면2

In machine translation, a critical error is an error that completely changes the meaning of the source text. Explanation is a sentence that explains why this error occurred. This helps us understand what caused the error and helps us avoid similar mistakes in the future. Correction is a sentence that corrects the erroneous translation. This is the process of fixing the translation to correctly reflect the source text's exact meaning. Based on the original and translation sentences, your work evaluates and scores the explanation and correction. Please make sure you read and understand these instructions carefully.

*** Explanation Scoring**

{Specificity} Judge whether the explanation of translation errors is detailed and illustrated with examples. A score of 5 indicates a detailed explanation with examples, while a score of 1 indicates a less detailed explanation.

{Severity} Determine whether the explanation describes a critical error that distorts the meaning of the original text. A score of 5 indicates that the explanation describes a critical error, while a score of 1 indicates that the explanation describes an error that is not critical.

{Understandability} Score if the translation is described in a way that makes it easy to understand what is wrong with the translation. A score of 5 indicates that the explanation is easy to understand, while a score of 1 indicates that the explanation is difficult to understand.

{Brevity} The explanation should not include unnecessary information that does not help you understand the error. A score of 5 indicates that the explanation does not contain unnecessary information, while a score of 1 indicates that the explanation does contain unnecessary content.

{Focus} The explanation should focus on errors that appear in the translation, not to consider errors in the source itself. A score of 5 indicates that the explanation accounts for errors present in the translation, while a score of 1 indicates that the explanation only accounts for errors in the source sentence.

*** Modified translation Scoring**

{Semantic preservation} Determine whether the modified translation accurately reflects the meaning of the source text. A score of 5 indicates a translation that completely preserves the original meaning, while a score of 1 indicates a translation that significantly distorts or loses the original meaning.

{Error} Determine whether the corrected translation is free of critical errors. A score of 5 indicates no critical errors, while a score of 1 indicates many critical errors.

{Minimal editing} Indicates how much the translation had to be edited to correct the error. A score of 5 means that the original sentence required minimal editing, while a score of 1 means that the translation required significant editing.

{Naturalness} Rate the extent to which the corrected translation is a natural sentence in the target language. A score of 5 means that the sentence is very natural and fluent, while a score of 1 means that the sentence is awkward or unnatural.

{Reflectivity} Judge whether all the errors in the explanation have been corrected in the revised translation. A score of 5 indicates that all errors have been corrected, while a score of 1 indicates that many corrections have not been incorporated.

도면3

