한국어 기계 번역에서의 품질 검증을 위한 치명적인 오류 범위 탐지 모델

정다현[1○], 이승윤[1], 어수경[1], 박찬준[2], 이재욱[1], 박기남[3*], 임희석[1,3*]
고려대학교 컴퓨터학과[1], Upstage[2], Human-inspired AI 연구소[3]
{dhaabb55, dltmddbs100, djtnrud, jaewook133, spknn, limhseok}@korea.ac.kr
bcj1210@naver.com

Critical Error Span Detection Model of Korean Machine Translation

Dahyun Jung[1○], Seungyoon Lee[1], Sugyeong Eo[1], Chanjun Park[2], Jaewook Lee[1], Kinam Park[3*], Heuiseok Lim[1,3*]
Department of Computer Science and Engineering, Korea University[1], Upstage[2], Human-inspired AI Research[3]

요 약

기계 번역에서 품질 검증은 정답 문장 없이 기계 번역 시스템에서 생성된 번역의 품질을 자동으로 추정하는 것을 목표로 한다. 일반적으로 이 작업은 상용화된 기계 번역 시스템에서 후처리 모듈 역할을 하여 사용자에게 잠재적인 번역 오류를 경고한다. 품질 검증의 하위 작업인 치명적인 오류 탐지는 번역의 오류 중에서도 정치, 경제, 사회적으로 문제를 일으킬 수 있을 만큼 심각한 오류를 찾는 것을 목표로 한다. 본 논문은 치명적인 오류의 유무를 분류하는 것을 넘어 문장에서 치명적인 오류가 존재하는 부분을 제시하기 위한 새로운 데이터셋과 모델을 제안한다. 이 데이터셋은 거대 언어 모델을 활용하는 구축 방식을 채택하여 오류의 구체적인 범위를 표시한다. 또한, 우리는 우리의 데이터를 효과적으로 활용할 수 있는 다중 작업 학습 모델을 제시하여 오류 범위 탐지에서 뛰어난 성능을 입증한다. 추가적으로 언어 모델을 활용하여 번역 오류를 삽입하는 데이터 증강 방법을 통해 보다 향상된 성능을 제시한다. 우리의 연구는 기계 번역의 품질을 향상시키고 치명적인 오류를 줄이는 실질적인 해결책을 제공할 것이다.

주제어: 기계 번역, 품질 검증, 치명적인 오류 탐지

1. 서론

기계 번역(Machine Translation, MT)은 언어 간의 정보 전달을 도와주는 중요한 기술로, 다양한 산업과 일상 생활에서 광범위하게 활용되고 있다. 그러나 기계 번역 시스템이 만들어 낸 문장이 얼마나 정확하고 신뢰할 수 있는지를 평가하는 것은 여전히 복잡한 문제로 남아 있다 [1]. 이러한 문제에 대응하기 위한 하나의 방법이 바로 품질 검증(Quality Estimation, QE)이다. 품질 검증은 정답 문장 없이 기계 번역된 문장의 정확성 및 일관성, 신뢰성 등을 평가하는 과정으로, 번역의 사용 케이스나 의도에 따라 매우 중요할 수 있다 [2, 3].

특히 품질 검증 내에서도 치명적인 오류 탐지(Critical Error Detection, CED)는 중요한 하위 작업으로 주목받고 있다. 치명적인 오류란 기계 번역 문장에 포함된 오류로 인해 의미의 왜곡이 일어나 개인적, 사회적으로 심각한 피해를 입힐 수 있는 오류를 의미한다 [4, 5]. 이전 연구에서는 영어-한국어 번역쌍을 대상으로 유해성(toxicity), 안전성(safety), 개체명(named entity), 수(number), 감성(sentiment), 공손성(politeness)의 여섯 가지 분류로 치명적인 오류를 정의하고 있다.

기존에는 치명적인 오류의 유무를 단순히 분류하는 연구가 주로 수행되었다 [5, 6]. 이 경우 문장의 어떤 부분에서 오류가 발생했는지를 알 수 없기 때문에 정보량이 부족하다. 따라서 본 연구에서는 문장에서 치명적인 오류가 어느 부분에서 발생했는지 설명하기 위한 목적으로 치명적인 오류의 범위를 찾고자 한다.

우리는 이 작업을 수행하기 위해 치명적인 오류 범위 탐지 데이터셋을 구축한다. 데이터셋 구축을 위한 방법론으로 거대 언어 모델(Large Language Model, LLM)을 맥락 학습(in-context learning) 방식으로 활용한다 [7, 8, 9]. 입력으로 원본 문장과 오류가 포함된 번역 문장, 오류의 범주(category) 정보가 주어지면 거대 언어 모델은 번역 문장에서 오류가 존재하는 범위를 찾는다. 우리는 언어 모델을 통해 얻은 결과를 작업자가 수정하도록 하여 구축 과정에서의 효율성 뿐만 아니라 데이터셋의 품질을 향상시키고자 한다.

우리는 해당 데이터셋을 활용하여 치명적인 오류를 문장에서 추출하는 작업과 오류의 유무 판별하는 작업을 동시에 수행하는 다중 작업 학습(multi-task learning) 모델을 제시한다. 모델은 사전 학습된 언어 모델을 인코더로 사용하여 얻은 임베딩 값을 바탕으로 학습을 진행한다. 학습 과정에서는 모델이 두 작업으로 인해 얻은 손실(loss)를 가중합한다. 더불어, 우리는 언어 모델을 사용하여 번역 데이터에 의도적으로 오류를 삽입하는 방식을 통해 데이터를 증강하여 해당 데이터를 사전 학습하는 프로세스를 추가적으로 구성한다.

우리의 모델은 오류 범위 탐지 작업 뿐만 아니라 오류의 유무 판별 작업에서도 뛰어난 성능을 실험적으로 입증한다. 추가적으로 제시한 데이터 증강 방법도 데이터의 실질적인 양을 증가시키고 작업에 대한 적응성을 높여 모델의 성능에 긍정적으로 작용한다. 우리는 우리의 데이터셋 및 코드를 공개하여 치명적

*교신저자(Corresponding author)

인 번역 오류에 대한 구체적인 해결책을 제시하고자 한다.

2. 관련 연구

치명적인 오류 탐지는 기계 번역 분야의 저명한 학회인 WMT 2021에서 품질 검증을 위해 처음 소개되었다 [6]. 이 연구에서는 원본 문장에서 의미가 왜곡되는 다양한 형태의 오류, 즉 오역(mistranslation), 환각(hallucination), 그리고 삭제(deletion)를 중점적으로 조명한다. 오역은 원본 문장의 의미가 왜곡되어 번역되는 경우, 환각은 원본 문장에 없는 내용이 추가되는 경우, 삭제는 원본 문장의 중요한 부분이 누락되는 경우를 의미한다.

다양한 방법론과 접근법을 통해 이러한 문제에 대한 연구가 활발히 진행되고 있다. [10]은 다양한 언어에 대한 사전 훈련된 모델을 기반으로 추가적으로 불균형 데이터와 특징 엔지니어링을 활용해 성능을 향상시킨다. [11]은 언어 간 사전 훈련된 표현을 기반으로 하여 불균형 데이터와 중요한 오류를 탐지하기 위한 특성 엔지니어링을 적용한 시퀀스 분류 모델을 사용한다.[12]는 프롬프트 기반 미세 조정을 활용하여 언어 모델의 사전 훈련과 미세 조정 간의 간극을 최소화하는 방식으로 작업을 수행한다.

문장 수준의 품질 검증 연구들과 더불어 단어 수준로 세분화하여 오류를 찾기 위한 연구도 계속 되고 있다. COMETKIWI는 COMET 프레임워크를 기반으로 OPENKIWI의 예측 추정기 구조와 연결하여 단어 수준 작업을 수행한다 [13]. 단어 수준의 평가를 위한 데이터셋을 제시하고 Translation Error Rate (TER) 기반 인공 코퍼스를 사람의 선택과 더 가깝게 만들기 위해 태그 구체화 전략과 트리 기반 주석 전략을 사용한 자가 지도 사전 학습을 제안하는 연구도 존재한다 [14].

그러나 현재까지의 연구에서는 치명적인 오류에 대한 구체적인 탐구가 미흡한 실정이다. 이러한 공백을 채우기 위해, 본 연구에서는 번역 문장에서 어떤 부분이 치명적인 오류를 포함하는지에 집중한다.

3. 치명적인 오류 범위 탐지

본 장에서는 먼저 데이터셋 구축 방법을 보이며, 다중 작업 학습 모델 구성 및 증강 방법을 설명한다.

3.1 데이터셋 구축

우리는 작업을 수행하기 위해 기구축된 영어-한국어 치명적인 오류 탐지 데이터를 바탕으로 해당 데이터에 오류 범위 주석을 추가하여 데이터셋을 구축한다. 데이터셋의 통계 정보는 표 1에 나타나 있으며, 각 카테고리에 따른 설명은 다음과 같다: (1) **유해성 (Toxicity, TOX)**: 종교, 성, 인종 등에 관련된 비속어나 증오어가 번역 문장에 포함되는 오류를 말한다, (2) **안전성 (Safety, SAF)**: 번역 오류가 안전이나 건강에 직접적인 영향을 미치는 경우이다, (3) **개체명 (Named Entity, NAM)**: 이름, 기관, 장소 등의 중요한 엔티티가 정확히 번역되지 않는 경우를 의미한다, (4) **수 (Number, NUM)**: 수치나 시간, 날짜 등이 잘못 번역된 경우로, 경제나 법적 문제를 일으킬 수 있으므로 치명적 오류로 분류된다, (5) **감성 (Sentiment, SEN)**: 문장의 긍정이나 부정의 극성이 번역 과정에서 변경되는 경우를 의미한다, (6) **공손성 (Politeness, POL)**: 번역 과정에서 상대방에 대한 존중이 표현되지 않는 경우로, 특히 영어에서 한국어로의 번역에서 주의가 필요하다.

표 1. 데이터셋의 통계 정보

	학습 데이터	검증 데이터	평가 데이터
문장 개수	7,265	500	1,000
오류 없음	6,606	444	924
오류 있음	659	56	76
- 유해성	133	6	7
- 안전성	122	15	10
- 개체명	95	12	20
- 수	116	6	12
- 감성	110	12	14
- 공손성	83	5	13

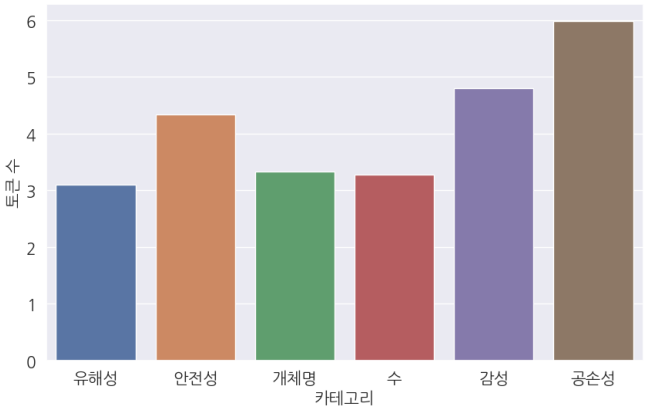

그림 1. 치명적인 오류의 카테고리 별 오류 토큰 수의 평균 비교 그래프

표 2는 오류 카테고리별 데이터의 예시를 보여주고, 그림 1에는 카테고리별 한 문장당 오류로 표시된 토큰의 개수가 명시되어있다.

데이터셋 구축을 위해 거대 언어 모델인 GPT-4 [9]를 활용하여 번역 오류를 자동으로 태깅하는 방법을 채택한다. 이 과정에서 모델에는 원본 문장, 오류가 포함된 번역 문장, 그리고 오류의 유형이 입력으로 제공된다. 일부 문장 내에는 다양한 오

표 2. 번역 문장에 나타난 치명적인 오류에 대한 범위 주석 데이터 예시

원본 문장	번역 문장	오류 유형
I am glad to see that you are considering our furnishing services.	귀하가 당사의 가구 서비스를 고려하고 있다는 것을 알게 되어 기쁩니다.	오류 없음
Wow, can **my friend** and I even eat all of that?	와, 저랑 <v>**머저리가**</v>저걸 다 먹을 수 있을까요?	유해성
At the same time, **it isn't well absorbed into the skin** and can be used year-round without any burden.	동시에 <v>**피부에 잘 흡수되어**</v>부담 없이 1년 내내 사용할 수 있습니다.	안전성
This avoids opening different windows and tabs to view data.	이제는 <v>**환자**</v>데이터를 보기 위해 다른 창과 탭을 열지 않아도 됩니다.	개체명
This electronic digital security safe has **five years** warranty.	이 전자 디지털 보안 금고의 보증은 <v>**5개월**</v>간 가능합니다.	수
I'm very awful of that, and I **appreciate your concern.**	나는 그것에 대해 매우 <v>**조심스럽고**</v>걱정해 주셔서 감사합니다.	감성
Recently, I received a fax from **your** department, but I could not make out the details.	최근에 <v>**니**</v>부서에서 팩스를 받았지만 자세한 내용을 확인할 수 없습니다.	공손성

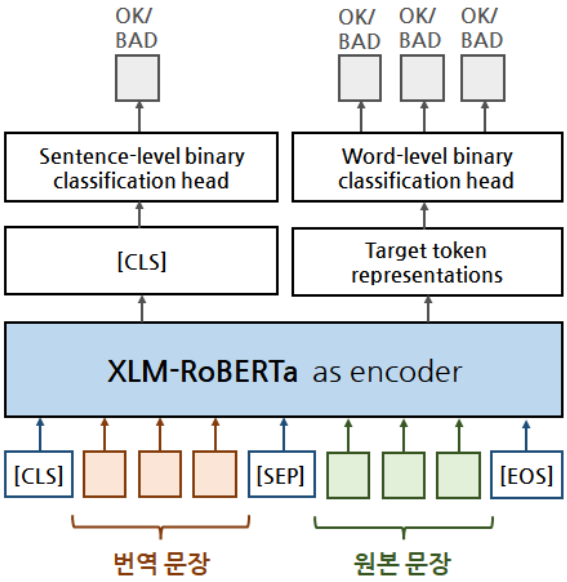

그림 2. 사전 학습된 인코더 모델을 사용한 다중 작업 학습 모델

류가 존재할 수 있으나 우리는 지정된 가장 심각한 오류 유형에 해당하는 하나의 범위에만 태깅한다. 작업의 명확성을 높이기 위해 번역된 문장에 대해서만 태깅을 진행하며, 이해도를 증진시키기 위해 3개의 예시 문장을 제공한다. 또한, 일관된 태깅을 위해 범위를 띄어쓰기 단위로 정의한다. 마지막으로, 모델의 출력을 사람이 다시 한 번 검수하여 데이터셋의 품질을 더욱 높인다.

3.2 제안 모델

우리 모델의 전체적인 구조는 그림 2에 나와 있다. 번역 문장과 그 원본 문장은 연결되어 인코더로 사용되는 XLM-RoBERTa [15]에 입력으로 제공되고 d-차원의 은닉 상태 벡터(hidden state vector)를 생성한다.

문장 수준 모델에서는 [CLS] 토큰의 은닉 상태를 이진 분류 헤드 모듈에 통과시켜 문장 수준의 오류 유무 예측 $\hat{y} \in \{\texttt{ok}, \texttt{bad}\}$를 얻는다. 단어 수준 모델에서는 번역 문장의 각 토큰에 대응되는 은닉 상태 벡터를 추출하고, 이를 단어 수준의 오류 유무 예측을 통해 단어 수준의 예측 $\hat{y}_i \in \{\texttt{ok}, \texttt{bad}\}, 1 \leq i \leq n$를 얻는다. n은 번역 문장의 토큰 개수이다.

우리는 이러한 모델을 사용하여 문장 수준의 주석(label)인 y와 단어 수준의 $\hat{y}$ 예측을 동시에 수행한다. 이와 같은 다중 작업 설정에서는 우리는 다음과 같은 결합 손실 함수 $\mathcal{L}$를 사용한다:

$$\mathcal{L}_{\text{sent}}(\theta) = -w_y \log p_\theta(y) \tag{1}$$

$$\mathcal{L}_{\text{word}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_{y_i} \log p_\theta(y_i) \tag{2}$$

$$\mathcal{L}(\theta) = \lambda \mathcal{L}_{\text{sent}}(\theta) + (1-\lambda)\mathcal{L}_{\text{word}}(\theta), \tag{3}$$

여기서 w는 OK와 BAD 태그에 주어진 클래스 가중치를 나타내며, λ는 문장 수준과 단어 수준의 손실을 조합하기 위해 사용되는 가중치이다. $\lambda = 1$일 경우 문장 수준 모델이 된다.

3.3 증강 방법

치명적인 오류는 쉽게 나타나지는 않지만 한번 나타나면 큰 피해를 입힐 수 있는 오류로, 직접 수집하는 것은 비용이 많이 들고 복잡하다. 따라서 우리는 치명적인 오류 데이터를 찾는 대신 번역 문장에 랜덤한 오류를 삽입하는 형식으로 데이터를 생성한다. 이때 사용된 데이터는 AI hub [1]의 일상생활 및 구어체 한국어-영어 병렬 말뭉치 데이터로 기구축된 치명적인 오류

[1] https://www.aihub.or.kr

표 3. 치명적인 오류의 범위 추출 실험 결과

Method	Model	F1 score	Recall	Precision
Word-level training	XLM-RoBERTa	-	-	-
	InfoXLM	-	-	-
Word + Sentence-level training	XLM-RoBERTa-base	0.0392	0.0400	0.0385
	XLM-RoBERTa-large	0.1644	0.1600	0.1690
	InfoXLM-base	0.1493	0.1333	0.1695
	InfoXLM-large	0.2222	0.1867	0.2745
Word + Sentence-level training + DA	XLM-RoBERTa-base	0.2059	0.1867	0.2295
	XLM-RoBERTa-large	**0.4394**	**0.3867**	**0.5088**
	InfoXLM-base	0.1963	0.2133	0.1818
	InfoXLM-large	0.3310	0.3200	0.3429

탐지 데이터가 이를 바탕으로 만들어졌다. 따라서 우리는 해당 데이터셋과 겹치지 않는 데이터를 선별하여 사용한다.

구체적으로 번역 문장 S에 대하여 무작위로 위치와 개수를 택하여 마스크 토큰 M을 삽입하는 방식으로 새로운 문장 S'을 생성한다.

$$S' = S_1, S_2, \ldots, M_i, M_{i+1}, \ldots, M_j, \ldots, S_n$$

여기서 S'은 XLM-RoBERTa-base 모델에 의해 예측되어 채워지며, i와 j는 각각 마스크되는 토큰의 시작과 끝 번호이다. 이렇게 생성된 문장은 원래 데이터셋을 미세 조정하는 과정 이전에 사전 학습으로 이루어진다. 이를 통해 우리는 번역에서의 오류를 다루는 것에 익숙하지 않은 모델을 학습시켜 치명적인 오류에도 강인한 모델을 만든다.

데이터 증강을 통해 총 1,194,141개의 문장을 생성하였으며, 이 중 에러를 포함하지 않는 문장은 835,899개, 에러를 포함하는 문장은 358,242개이다. 필터링 과정을 거친 후에는 총 799,525개의 문장을 얻었으며, 이 중에서 584,407개는 에러가 없고, 215,118개는 에러를 포함한다.

4. 실험 및 결과

본 장에서는 실험 환경 및 평가 지표, 실험 결과에 대하여 보고한다.

4.1 실험 환경

우리의 구현은 COMETKIWI [13]를 기반으로 하며, 인코더 모델로는 허깅페이스 라이브러리[2]에서 제공하는 XLM-RoBERTa을 채택한다. 백본(backbone) 모델로는 XLM-RoBERTa와 InfoXLM [16]의 base, large 버전을 실험한다. 다중 작업을 위한 가중치 λ는 경험적 판단과 사례 연구를 통해 0.5로 설정한다. 학습은 NVIDIA RTX A 6000 GPU에서 진행되었으며, 70 에포크 동안 배치 사이즈 64로 실행된다. 사전 훈련 단계에서는 3 에포크, 배치 크기는 64로 유지한다. 조기 중지는 미세조정 단계에서만 적용되며, 기준은 50 에포크이다. 학습 도중 2 에포크마다 모델의 성능을 평가하며, 사전 훈련은 5시간 이상, 미세조정은 1시간 이상이 소요된다. 모델 최적화를 위해 AdamW 옵티마이저를 사용하였으며, 사전 훈련과 미세조정 단계에서 모두 학습률을 1.5e-05로 설정한다. 모든 하이퍼파라미터는 수동으로 조정한다.

[2] https://huggingface.co/

4.2 평가 지표

오류 범위 탐지 평가에서는 F1 점수, 재현율(Recall), 정밀도(Precision)을 사용한다. F1 점수는 정밀도와 재현율의 조화 평균을 통해 높은 수준의 정확성과 완전성을 동시에 추구한다. 재현율은 실제 오류 중 얼마나 많은 오류를 성공적으로 탐지했는지, 정밀도는 탐지된 오류 중 실제로 오류인 경우의 비율을 나타낸다. 오류 유무 탐지에는 Matthews Correlation Coefficient(MCC) [17]라는 지표도 함께 제시한다. 이는 이전의 WMT 2021 [6], WMT 2022 [5]에서도 사용되었던 지표로, 실제와 예측 간의 상관 관계를 정량적으로 평가해주는 지표이다.

4.3 오류 범위 탐지 실험

표 3은 원본 문장과 번역 문장을 입력으로 받아 번역 문장에 존재하는 오류의 범위를 예측하는 실험에 대한 성능을 보여준다. Word-level training, 즉 번역 문장에 각 토큰에 대해 오류가 존재하는 지를 학습하는 방식으로만 실험한 결과는 XLM-RoBERTa와 InfoXLM 모델은 base와 large 모두 0에 수렴하는 성능을 보인다. 이처럼 해당 작업은 단독으로 수행하면 모델이 수행하기 어려운 작업이다. 반면 Sentence-level, 즉 우리가 제안하는 문장 안 오류 존재 유무 탐지 작업을 동시에 사용하는

표 4. 치명적인 오류의 유무 분류 작업 실험 결과

Method	Model	MCC	F1 score	Recall	Precision
Sentence-level training	XLM-RoBERTa-base	0.2441	0.5985	0.5719	0.7072
	XLM-RoBERTa-large	0.5577	0.7621	0.7062	0.8771
	InfoXLM-base	0.2210	0.5313	0.5263	**0.9639**
	InfoXLM-large	0.4901	0.7135	0.6552	0.8870
Sentence + Word-level training	XLM-RoBERTa-base	0.3873	0.6935	0.6889	0.6985
	XLM-RoBERTa-large	0.3135	0.6567	0.6577	0.6558
	InfoXLM-base	0.4824	0.7352	0.7002	0.7905
	InfoXLM-large	0.5688	0.7816	0.7529	0.8199
Sentence + Word-level training + DA	XLM-RoBERTa-base	0.5722	0.7804	0.7419	0.8384
	XLM-RoBERTa-large	0.5848	0.7895	0.7600	0.8289
	InfoXLM-base	0.5780	0.7889	0.7924	0.7856
	InfoXLM-large	**0.5970**	**0.7984**	**0.8056**	0.7916

경우 InfoXLM-base와 InfoXLM-large 모델에 대해 측정한 모든 지표에서 상대적으로 높은 성능을 보여준다. 이는 단어 수준의 정보만을 사용하는 것보다 문장의 전체적인 특징을 함께 고려하는 것이 이 작업에 더 유리하다는 것을 의미한다.

Sentence + Word-level training + DA은 데이터 증강(Data Augmentation, DA) 방법을 적용한 성능이다. 이 경우 XLM-RoBERTa-large 모델은 F1 점수에서 0.4394, Recall에서 0.3867, 그리고 Precision에서 0.5088으로 가장 높은 성능을 보인다. 이 결과는 우리가 제안하는 마스크 예측 기반 번역 오류 데이터 증강 방식이 치명적인 오류의 범위 탐지 작업에서 매우 유용하게 작용하며 성능 향상에 중요한 요소임을 시사한다.

4.4 오류 유무 판별 실험

표 4는 번역 문장에서 치명적인 오류의 유무를 이진 분류하는 작업에 대한 성능을 보여준다. 이 실험을 통해 우리는 다중 작업 학습 방식과 데이터 증강 방식으로 오류 범위 추출 뿐만 아니라 기존 오류 탐지 작업에서도 유용함을 제시하고자 한다. 따라서 이 실험에서 주목할 부분은 Word-level training과의 다중 학습 및 데이터 증강에 따른 모델의 성능 변화이다.

모든 평가 지표에서 Sentence + Word-level training + DA 방법이 가장 높은 성능을 보이는 것을 확인할 수 있다. 특히 InfoXLM-large 모델은 MCC에서 0.5970, F1 스코어에서 0.7984, Recall에서 0.8056으로 가장 높은 성능을 나타낸다. 이는 우리가 제안하는 다중 작업 학습 방법 및 데이터 증강 기반 사전 학습을 함께 고려할 때 모델이 더욱 정확한 분류 성능을 보임을 의미한다. 해당 모델이 비록 Precision에 대해서는 비교적 낮은 점수를 보이지만 다른 평가 지표들까지 종합적으로 고려하는 것이 중요하므로 이러한 부분에서 전반적으로 좋은 성능을 보인다고 설명할 수 있다.

5. 결론

본 논문에서는 기계 번역 시스템에서 발생할 수 있는 치명적인 오류의 범위를 탐지하기 위한 다중 작업 학습 모델과 새로운 데이터셋을 제안하였다. 오류의 범위를 탐지하는 작업과 유무를 판별하는 작업을 동시에 수행하면 단독으로 수행하는 경우보다 좋은 결과를 얻을 수 있음을 확인했다. 또한, 데이터 증강 기술을 통해 더욱 뛰어난 성능을 달성하였다. 이 연구를 통해, 기계 번역의 품질을 높이고 치명적인 오류를 줄일 수 있는 실질적인 방안을 제시하였으며, 이는 향후 번역 시스템의 안정성과 신뢰성을 높이는 데 기여할 것이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2022-2018-0-01405). 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2022R1A2C1007616).

참고문헌

[1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

[2] L. Specia, M. Turchi, N. Cancedda, N. Cristianini, and M. Dymetman, "Estimating the sentence-level quality of machine translation systems," *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, May 14–15 2009. [Online]. Available: https://aclanthology.org/2009.eamt-1.5

[3] L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, and A. F. T. Martins, "Findings of the WMT 2020 shared task on quality estimation," *Proceedings of the Fifth Conference on Machine Translation*, pp. 743–764, Nov. 2020. [Online]. Available: https://aclanthology.org/2020.wmt-1.79

[4] V. Raunak, M. Post, and A. Menezes, "Salted: A framework for salient long-tail translation error detection," *arXiv preprint arXiv:2205.09988*, 2022.

[5] C. Zerva, F. Blain, R. Rei, P. Lertvittayakumjorn, J. G. C. de Souza, S. Eger, D. Kanojia, D. Alves, C. Orăsan, M. Fomicheva, A. F. T. Martins, and L. Specia, "Findings of the WMT 2022 shared task on quality estimation," *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 69–99, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.wmt-1.3

[6] L. Specia, F. Blain, M. Fomicheva, C. Zerva, Z. Li, V. Chaudhary, and A. F. T. Martins, "Findings of the WMT 2021 shared task on quality estimation," *Proceedings of the Sixth Conference on Machine Translation*, pp. 684–725, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.wmt-1.71

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[8] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[9] OpenAI, "Gpt-4 technical report," 2023.

[10] G. Jiang, Z. Li, and L. Specia, "ICL's submission to the WMT21 critical error detection shared task," *Proceedings of the Sixth Conference on Machine Translation*, pp. 928–934, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.wmt-1.97

[11] Y. Chen, C. Su, Y. Zhang, Y. Wang, X. Geng, H. Yang, S. Tao, G. Jiaxin, W. Minghan, M. Zhang, Y. Liu, and S. Huang, "HW-TSC's participation at WMT 2021 quality estimation shared task," *Proceedings of the Sixth Conference on Machine Translation*, pp. 890–896, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.wmt-1.92

[12] S. Eo, C. Park, H. Moon, J. Seo, and H. Lim, "KU X upstage's submission for the WMT22 quality estimation: Critical error detection shared task," *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 606–614, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.wmt-1.56

[13] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. De Souza, T. Glushkova, D. M. Alves, A. Lavie *et al.*, "Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task," *arXiv preprint arXiv:2209.06243*, 2022.

[14] Z. Yang, F. Meng, Y. Yan, and J. Zhou, "Rethink about the word-level quality estimation for machine translation from human judgement," 2022.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020.

[16] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, and M. Zhou, "Infoxlm: An information-theoretic framework for cross-lingual language model pre-training," 2021.

[17] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, Vol. 21, No. 1, pp. 1–13, 2020.