사전 정보를 활용한 신경망 기계 번역

전현규, 김지윤°, 최승호, 김봉수
(주)와이즈넛
{eddie14,jiyoonkim,csh1019,usgnob}@wisenut.co.kr

Neural Machine Translation with Dictionary Information

Hyun-Kyun Jeon,Ji-Yoon Kim°,Seung-Ho Choi,Bongsu Kim
Wisenut Inc. , Wisenut Research

요 약

최근 생성형 언어 모델이 주목받고 있으며, 이와 관련된 과제 또한 주목받고 있다. 언어 생성과 관련하여 많은 연구가 진행된 분야 중 하나가 '번역'이다. 번역과 관련하여, 최근 인공신경망 기반의 신경망 기계 번역(NMT)가 주로 연구되고 있으며, 뛰어난 성능을 보여주고 있다. 하지만 교착어인 한국어에서 언어유형학 상의 다른 분류에 속한 언어로 번역은 매끄럽게 번역되지 않는다는 한계가 여전하다. 따라서, 본 논문에서는 이러한 문제점을 극복하기 위해 한-영 사전을 통한 번역 품질 향상 방법을 제안한다. 또한 출력과 관련하여 소형 언어모델(sLLM)을 통해 CoT데이터셋을 구축하고 이를 기반으로 조정 학습하여 성능을 평가할 것이다.

주제어: 기계번역, 지식 기반, 생성모델, 전이학습

1. 서론

최근 거대 언어모델(LLM; Large Language Modeling)의 등장으로 생성형 언어모델이 주목받고 있다. 생성형 언어 모델은 자연어 문장을 입력받아 그에 대응하는 텍스트를 생성하는 것을 목표로 하며, 최근에는 연쇄적으로 다음 토큰을 예측하여 문장을 생성하는 GPT[1](Generative Pre-trained Transformer) 구조의 모델 중 GPT-3[2], GPT-4와 같은 모델이 여러 생성 과제에 있어 가능성을 보여주며 연구가 활발히 진행 중이다.

이와 함께 자연어 생성 과제 또한 주목받고 있다. 생성형 모델 발전 이전부터 여러 텍스트 생성 과제들이 있었지만, 효과적인 평가 척도의 부재, 생성 품질의 저하 등으로 인해 연구에 어려움을 겪었다. 하지만 최근의 비약적인 발전으로 인해, 번역, 요약, 질의응답, 이야기 생성 등의 분야에 대한 여러 관점에서의 연구가 재조명받고 있다. 거대 언어모델에 프롬프트 등을 통해서 모델의 파라미터 학습 없이 개선된 품질의 결과를 얻어내거나, 모델이 단순 결과를 추론하는 것이 아닌 생각을 표현한 후 결과를 출력하도록 하는 등의 방법들이 연구되고 있다.

본 논문에서는 여러 생성 과제 중 신경망 기계번역(Neural Machine Translation)의 품질을 향상하는 방법에 관하여 연구하고자 한다. 신경망 기계번역은 인공신경망 모델을 기반으로 번역하고자 하는 문장을 모델에 입력하여 모델의 결과를 번역 결과로 출력하는 것을 의미한다. 기본적인 접근 방법은 입력-출력 간의 표현 공간(representation space)을 엔드-투-엔드(end-to-end)로 학습한 모델을 이용하는 것이다.

하지만 이러한 방법으로 학습된 모델은 유의어나 고유 명사 또는 관용구를 적절한 형태로 번역하는 데에 어려움을 겪는다. 인공지능 관련 문서에 'Generation'이라 표기된 페이지가 있을 때, 이는 '생성'으로 번역되는 것이 적절하다. 하지만, 구글 번역기 등의 상용화된 번역기조차도 이를 '세대'라고 번역하는 등의 오류를 범한다. 또한, 최근 신경망 기계 번역으로 주목받는 '딥엘(DeepL)'의 경우에도, 'Kim'과 같은 영어로 표기된 이름을 '김연아'로 오역하는 등의 번역패턴을 보이는 것을 확인할 수 있다.

따라서, 본 논문에서는 사전 정보를 모델이 직접 활용하여 번역하는 방법을 통해 번역 품질을 높일 것이다. 사전 정보를 별도의 기호화를 하는 것이 아니라 입력과 함께 넣어 모델이 이를 학습할 수 있도록 할 것이다. 또한 사전을 직접 활용하는 것이 아닌, 별도의 과제를 구성하여 학습한 뒤, 번역 과제를 학습하는 방법을 통해 사전 정보를 모델에 내재화할 것이다. 추가로, 사전을 기반으로 최근 생성 모델에서 활용되는 'CoT(Chain-of-Thought)'[3]를 학습하고, 그 결과를 살펴볼 것이다.

2. 관련 연구

사전 정보를 활용한 기계번역 품질 향상을 위한 연구는 특정 도메인에서 실행되었다. 온라인의 어학 사전을 바탕으로 신조어에 대해 번역 품질을 향상한 연구가 있다. [4] 해당 연구에서는 신조어가 포함된 문장의 기계번역 정확도를 높이기 위해 네이버(Naver) 어학사전을 활용하였다. 본 실험은 학습할 수 있는 모델이 아닌 API 형태의 구글 번역 서비스(Google Translation API)를 사용하였다. 위키피디아, 나무위키 등을 통해 신조어를 추출하고 이것의 "이것은 ~ (이)다"라는 템플릿을 활용해 입력을 구성한 뒤 그에 대응되는 레이블을 네이버 어학사전에서 검색한 결과를 사용했다. 이를 통해 번역 성능을 자카드 유사도

(Jaccard similarity) 기준, 0.5245에서 0.5406으로 향상하였다고 기술하였다.

또한 기계번역의 사후 검증(APE)에서 사전을 사용하여 품질을 향상하려는 방법이 연구되었다. 전이 학습을 기반으로 기계번역의 사후 검증 성능을 높인 연구가 있었다.[5] 구글, 아마존, 마이크로소프트의 번역기를 활용해 원문을 번역하고 이를 사후 교정의 입력으로, 기존 레이블을 사후 교정 레이블로 하여 APE 사전을 구축하고 이를 기반으로 사후 교정 모델을 학습시켜 번역 품질의 향상을 실험하였다. mBART[6]를 기반으로 APE 모델과 번역을 학습하고 APE를 학습한 모델을 준비하여 실험하였다. 실험 결과, 구글 번역기의 경우, 기존에는 BLEU 점수가 34.49였지만, mBART(APE), mBART(FT+APE) 각각 39.68, 41.86으로 향상된 결과를 보여주었다.

이처럼 번역 내의 세부 영역 및 과제에 대해 사전을 사용한 연구는 있었지만, 번역 모델 학습에 직접적으로 사용한 연구는 찾아볼 수 없었다. 따라서, 본 연구에서는 사전을 이용하여 번역모델 자체를 학습하고 이에 대한 성능을 확인할 것이다.

3. 방법

본 논문에서는 사전 정보를 활용하여 모델 입력에 검색된 사전 정보를 더한다. 번역에 있어서 문장의 의미는 단어의 의미에 부분적으로 종속되기 때문에, 입력된 한글 문장 내의 단어를 추출하고 이를 활용한다. 기존에는 원문(source text)을 입력받아 대상문(target text)을 출력하는 형식의 과제였다면, 여기에 추가로 사전 정보를 검색해 원문에 붙여주는 것이다.

이를 위해 번역 언어 쌍에 대응되는 사전이 구축되어 있어야 한다. 본 논문에서는 국립국어원의 한-영 외국어 사전을 기반으로 영-한 병렬 사전을 구축하였다. 한국어 표제어가 있으면 그에 대응되는 영어 단어 또는 의미가 추가된 형태이다. 예를 들어, ‘포장’이라는 단어에 대응되는 영어 단어 또는 의미는 ‘wrapping’, ‘gift-wrapping’, ‘package’ 등이 있다. 따라서 이를 병렬로 일대다(one-to-many) 대응하여 구성하였다.

본 논문에서는 사전 정보를 단순히 입력에 추가하는 것뿐만 아니라 다양한 선행 과제를 학습하는 데에 사용한다. 가장 기본적인 방법은 입력 문장의 단어를 사전에 검색하여 그 결과를 입력 문장 뒤에 붙여주는 것이다. 문장 내 단어를 추출하기 위해서는 형태소 분석기를 사용한다. 형태소 분석기를 통해 명사 또는 동사를 추출한 다음, 이를 사전에 검색하여 나온 결과를 차례대로 이어 붙인다. 이때, 동음이의어가 있으면, 이를 전부 표기한다. 이는 동음이의어 중 어떠한 단어가 적절한 의미의 단어 인지 비지도 방법(unsupervised method)으로 판별하기 어렵기 때문이다. 또한 여러 단어를 제공함으로써, 모델이 여러 단어 중 적절한 의미가 있는 단어를 선별하는 분별력을 학습하는 데 도움을 줄 수 있을 것이다.

이러한 방법은 몇 가지 단점이 있다. 우선, 이를 위해서 입출력 언어에 대한 병렬 사전이 필요하다. 이를 위해, 본 논문에서는 국립국어원에서 공개한 ‘한국어-영어 외국어 사전’을 사용하였다. 이 사전은 한국어 단어와 그에 대응되는 영어 단어 또는 의미가 기술되어 있다. 각각의 단어는 워드 파일로 내려받을 수 있도록 서비스를 제공하고 있다. 따라서, 이를 기반으로 단어를 수집하였으며, 추후 검색에 사용될 수 있도록 그 형태를 가공하였다.

하지만, 이러한 사전 정보를 입력 텍스트 뒤에 추가하면 입력 길이가 제한된다. 최근 LLM이 주목받으면서 입력 길이에 대한 제약이 줄어들었다. 그런데도 사전 정보를 입력에 붙여주는 것은 입력 문장의 길이가 길어질수록 그에 따라 제공되는 사전 정보 또한 증가하기 때문에 상대적으로 길이에 대한 불이익이 있을 수밖에 없다.

따라서 본 논문에서는 입력 길이를 제약하지 않고 사전정보를 학습하는 방법에 대하여 실험하고 그 성능을 살펴본다. 입력 길이를 제약하지 않고 사전 정보를 학습하는 방법의 하나는 전이 학습(transfer learning)이다. 여기서 말하는 전이 학습이란, 사전 정보를 학습하는 과제를 먼저 학습한 뒤에 번역 과제를 학습하는 것을 말한다. 번역이 아닌 다른 과제로 사전 정보를 학습할 것이다.

우선 사전 정보를 기반으로 질의응답 데이터셋을 구성하여 학습할 수 있다. 사전을 바탕으로 문장 내의 단어의 의미가 타깃 언어(영어)로 표현하면 무엇인지를 맞추도록 데이터를 구성할 수 있다. 또한 사전에서 검색 시 단어의 의미가 여러 개일 때, 즉 동음이의어 일 또는 다의어일 때, 표적 단어를 모두 입력 텍스트에 추가하여 그중 어떠한 것이 적절한지를 묻는 질의응답 데이터로 변환할 수 있다. 예를 들어 ”홍길동을 등기 이사로 선임했다”라는 문장이 있다고 할 때, 사전 정보를 참고하여, 여기에 ” ‘이사’가 의미하는 것은 무엇이야? 1. moving 2. director” 라는 질의를 생성할 수 있으며 대응되는 레이블은 ” ‘이사’가 의미하는 바는 2. director입니다.”라고 구성할 수 있다.

이처럼 구축된 데이터셋을 기반으로 생성형 질의응답을 먼저 조정 학습한 뒤, 번역 과제를 학습한다. 물론 평가의 정확성을 위해 사전에 대한 학습은 학습(train) 데이터로만 진행한다. 한 문장에 동음이의어 또는 다의어가 많을 경우, 이를 각각의 예제로 간주하여 학습한다.

또한 표적 텍스트를 활용해 동음이의어를 처리한 후, 이를 입력에 추가하여 CoT 데이터셋을 구축 및 학습할 수 있다. CoT은 ‘Chain-of-Thought’의 약자로 생성 모델이 문장을 생성할 때, 과제에서 요구하는 생성뿐만 아니라, 그것의 근거가 되는 생각(thought)을 함께 생성하는 것을 말한다. 최근 거대 언어모델이 주목받으면서, 이를 생성하기 위해 유도하는 프롬프트를 입력에 추가하거나, 이를 기반으로 조정 학습하는 등의 방법

표 1. 사전 데이터 예시

단어	의미
강아지	1. puppy, 2. baby; sweetheart
강연	lecture
강박	1. stifling, 2. obsession

을 통한 생성 품질 향상 기법들이 많이 연구되고 있다. CoT을 조정 학습할 경우, '생각'에 대응되는 텍스트를 구성하여야 하는데, 이를 사람이 직접 구축하기는 어렵다. 따라서, 규칙 또는 모델 기반으로 이를 얻는 방법을 모색하는 것이 효과적이다.

번역 과제는 그 특성상, 사전 정보를 활용하여 생각 텍스트를 비 지도의 방법으로 구축할 수 있다. 입력을 원문 텍스트(source text)로 설정하고, 사전 정보를 생각 텍스트, 결론을 대상 텍스트(target text)로 설정하여 데이터셋을 구성한다. 이를 기반으로 학습된 생성 모델은 번역의 대상이 되는 원문 텍스트를 입력받아 번역의 근거가 되는 사전 정보를 생성하고, 이후 번역 결과를 출력하는 것을 목표한다. 물론 생각 텍스트 앞뒤에 스페셜 토큰('[cot]', '[/cot]')을 추가하여, 추론 시에는 생각 텍스트를 제거할 수 있도록 한다.

이처럼, 사전정보를 이용하여 다양한 관점에서 활용 및 학습하고 그 결과를 공유하고자 한다. 위에서 설명한 방법들을 요약하면 다음과 같다.

1. 사전 정보를 입력에 더하여 번역에 대해 학습한다.
2. 질의응답 데이터를 구축/학습하고 번역에 대해 학습한다.
3. CoT 데이터를 구성하여 번역에 대해 학습한다.

4. 실험 환경 및 설정

4.1 데이터

본 실험에서는 AI 허브(AI-HUB)의 '한국어-영어 번역(병렬) 말뭉치'를 기반으로 학습 및 평가하였다. AI 번역 엔진 개발을 위해 뉴스, 정부/지자체 홈페이지, 간행물, 행정 규칙, 한국 문화, 구어체, 대화체를 기반으로 구축된 데이터셋이다. 원본 데이터셋은 데이터셋의 크기가 크기 때문에, 학습 및 평가 데이터를 각각 10만 개, 1만 개로 표본을 뽑아 사용하였다.

이와 함께 실험을 위해 사전 데이터 또한 구축하였다. 사전 데이터는 국립국어원 한국어-외국어 사전을 기반으로 한국어 단어와 그에 대응되는 영어 의미를 병렬적으로 구성하였다. 한국어에 대응되는 의미가 여러 개인 것은 번호를 추가하여 표기하였다. 총 51,957개의 한국어-영어 사전을 구축하였으며, 그 예시는 아래와 같다.

표1과 같이 구축된 단어 사전을 기반으로 번역 데이터셋 내의 원문 텍스트 내 단어를 검색하여 그 결과를 'resonale'이라는

표 2. 사전 정보가 적용된 번역 데이터 예시

열	내 용
source	산업포장을 받는 임진홍 샬레코리아(주) 대표이사는 국내여행업 분야의 B2B 시장 개척으로 근로자 연차휴가 활성화를 통한 국내관광산업 발전 기여를 인정받았다.
target	Im Jin-hong, CEO of Chalet Korea Co., who is under industrial packaging, was recognized for his contribution to the development of the local tourism industry by revitalizing the employees' annual vacation by opening new B2B market in the domestic travel industry.
reasonale	1. 산업: industry 2. 포장: wrapping; gift-wrapping; package; packaging material, decoration; exaggeration, pavement; surfacing 3. 대표: representative, ... (생략)

표 3. 실험에 사용된 모델 정보

모델	모델 크기 (billion)	단어장 크기	최대 컨텍스트 길이
wisenut-research/KoT5	0.22	32100	relative (pre-trained in 512)
psyche/kollama2-7b	7	32000	4096

칼럼을 추가하여 표기하였으며, 그 예시는 표2와 같다.

4.2 모델

본 실험에 사용된 모델은 T5[7], Llama2[8] 이다. 여러 설정을 바탕으로 성능을 평가하기 위해 비교적 크기가 작은 T5 모델을, CoT등의 지식 추론 가능성을 평가하기 위해 Llama2 모델을 사용하였다. T5는 인코더-디코더 구조의 시퀀스-투-시퀀스 모델(Seq2Seq LM)이며, Llama2는 GPT와 같은 디코더 구조의 인과적 모델(Causal LM)이다. 따라서, 서로 다른 구조와 크기의 모델을 통해 생성형 모델 전반에 대해, 일출력 설정 등을 변화하여 번역 성능을 비교할 것이다.

T5는 (주)와이즈넛에서 공개한 KoT5 모델 중 Base(220M) 모델을 사용하였다. 반면, Llama2 모델은 메타에서 공개한 Llama2 모델 중 7B(70억) 크기의 모델을 한국어 데이터로 추가 학습한 모델인 'psyche/kollama2-7b'모델을 사용하였다. 해당 모델은 허깅페이스 허브(Huggingface Hub)에 공개된 Llama2를 한국어 도메인에서 조정 학습한 모델이다. 각 모델의 크기 및 유형에 대한 자세한 내용은 표3에 기술되어 있다.

4.3 평가 척도

실험을 위한 평가 척도로 BLEU 점수[9]를 사용했다. BLEU(Bilingual Evaluation Understudy Score)는 기계 번역의 결과와 사람이 만든 레이블이 얼마나 유사한지 정밀도(precision)를 기반으로 나타낸 지표이다. BLEU 점수는 평가 단위에 따라 1개 토큰 단위로 계산된 BLEU-1부터, 4개 토큰 단위로 계산된 BLEU-4로 세분되며, 모두를 종합하여 곱한 값의 네제

표 4. T5의 입력 구성에 따른 번역 조정학습 결과

모델	입력 구성	BLEU-1	BLEU-2	BLEU-3	BLEU-4
wisenut-research/KoT5	source	0.3592	0.077	0.0266	0.0111
	source +dictionary	**0.4075**	**0.1335**	**0.0454**	**0.0208**
	source + target dictionary	0.3323	0.1049	0.0412	0.0188

표 5. T5의 학습 방법에 따른 번역 조정학습 결과

모델	조정학습	BLEU-1	BLEU-2	BLEU-3	BLEU-4
wisenut-research/KoT5	NMT	0.3592	0.077	0.0266	0.0111
	QA >NMT	**0.3774**	0.0895	0.0328	0.0140
	NMT(CoT)	0.3488	**0.1266**	**0.0569**	**0.0281**

곱근 BLEU (종합) 점수로 표기한다.

$$BLEU = \min[1, \frac{l(prediction)}{l(reference)}] * (\Pi_{i=1}^{4} precision_i)^{\frac{1}{4}}, \quad where\, l(*)\, is\, the\, length\, of\, tokens\, in\, * . \tag{1}$$

4.4 계산 자원 및 설정

본 실험에 사용된 계산 자원과 그에 따른 설정은 아래와 같다.

- VRAM: RTX-4090 × 2, A100-SXM-40G × 1
- Batch Size(Train/Test): 8/16(T5), 1/4(Llama2)
- Max Sequence Length(source/target): 512/512
- Generation Config: The temperature is 0.1, repetition penalty is 1.0, number of beams is 1, and no sampling.

5. 실험 결과

실험 결과는 다음과 같다. 실험은 단계별로 진행하였으며, 우선 사전 정보의 유용성을 파악하기 위해 조정 학습 실행 시 사전 정보를 입력에 추가하여 그 효과를 확인하였다. 표4는 이를 나타낸 것으로, 같은 사전학습 언어모델(wisenut-research/KoT5)에 대하여 입력 구성을 달리하여 조정 학습하고 평가한 결과이다. BLEU-1 ∼ 4을 사용하여 평가하였으며, 평가 시 토큰 분리는 공백으로 처리하여 계산하였다. 각 입력 구성과 그에 따른 모델 학습은 5 에포크(epoch)만큼 학습하여 평가하였다. 평가 관련해서는 최대 길이를 512로 temperature를 0.1로 설정하였으며, 그 밖의 생성 파라미터는 허깅페이스 라이브러리(transformers.GenerationConfig)의 생성 기본 설정을 따랐다.

실험 결과를 보면 입력에 사전 정보를 추가하여 학습한 것이 좋은 결과를 보여주는 것을 알 수 있다. 표4에서는 입력 구성을 달리한 결과를 나타내는데, 'source'는 입력의 원문을 의미하며, 'dictionary'는 이전 장에서 설명한 사전 정보를 의미한다. 또한 사전의 내용 중에 번역 결과(target)에 등장하는 의미만 따로 선별하여 구성한 사전이 'target dictionary' 이다. 원문 텍스트에 사전정보를 더하여 학습한 모델('source + dictionary')이 BLEU-1 기준 0.4075로 가장 높으며 나머지 BLEU-2 ∼ 4에 대해서도 마찬가지이다.

주목할 점은 'source+target dictionary' 가 'source + dictionary'보다 낮은 점수를 기록했다는 점이다. 기존에는 원문 텍스트에 있는 단어를 사전에서 검색해서 그 결과들을 모두 입력에 더하였다. 이는 사전이 단어뿐만 아니라 의미를 설명하는 내용도 포함되어 있으며, 다의어의 경우 모델이 이를 분별하기 어렵기에 모델 성능이 저하될 것으로 판단하였다. 그래서 번역 결과 레이블(target)을 보고 검색한 사전정보를 추려 입력에 추가하여 학습하였고, 이것이 'source+target dictionary' 이다. 그런데 해당 방법이 기존에 검색한 전체 사전을 제공하는 것보다 낮은 성능을 보인 것이다. 이는 직접적인 사전 정보(exact words)뿐만 아니라 간접적인 사전정보(meaning of words, polysemy)가 모델학습에 도움을 준 것으로 보인다.

또한 표5는 조정학습 방법의 변화에 따른 T5 모델의 성능 결과를 나타낸다. 표에 사용된 모델은 번역 과제를 조정 학습한 'NMT', 번역 데이터를 바탕으로 질의응답 데이터셋을 구축하여 학습하고 이를 기반으로 번역 과제를 조정 학습한 'QA ¿ NMT', 마지막으로 사전정보를 CoT의 근거(생각)로 하여 조정 학습한 'NMT(CoT)'가 있다. 결과를 보면 질의응답 과제를 구성하여 학습한 모델을 기반으로 번역 과제를 조정 학습한 결과가 BLEU-1에서 0.3774로 가장 높은 성능을 보여주었다. 비록 표4에서 표시한 사전정보를 직접 입력하는 것보단 낮지만, 일반적인 NMT학습 대비 높은 성능을 보여주면서 사전정보가 내재 가능하다는 것을 보여주었다. 반면, BLEU-2 ∼ 4는 CoT를 학습한 모델이 높은 수치를 보였다. 이는 단어가 아닌 구 또는 절에 있어서 다른 모델 대비 레퍼런스와 유사하다는 것을 의미하며, 매우 작은 파라미터의 모델(220M)임에도 불구하고 CoT 학습이 일정 부분 효과가 있음을 보여준다.

추가로, 번역에서 CoT 학습이 효과가 있는지 좀 더 살펴보기 위해, 더욱 큰 파라미터 크기를 갖는 모델을 기반으로 실험을 진행하였다. 학습 데이터는 기존 10만 개였으나, 실험 환경의 제약으로 인해 1만 개 표본을 뽑아 1 에포크만 학습하였다. 학습 시에는 전체 파라미터 학습이 아닌 LoRA(Low Rank Adaptation)[10]를 기반으로 한 Adapter를 학습하였다. 평가 데이터는 동일하게 1만 개를 사용하였으며, 그 밖에 실험 설정은 이전과 같다.

Llama2에 대하여, 번역 과제만을 조정 학습한 'NMT'보다

표 6. Llama2 조정학습 및 CoT학습 결과

모델	조정학습	BLEU-1	BLEU-2	BLEU-3	BLEU-4
psyche/kollama2-7b	NMT	0.3922	0.3016	0.2666	0.2439
	NMT(CoT)	**0.4128**	**0.3131**	**0.2884**	**0.2701**

CoT학습을 한 'NMT(COT)'가 더 좋은 결과를 보여주었다. 표6을보면 'NMT(CoT)'는 BLEU-1이 0.4128로 단순 번역과제를 학습한 모델이 기록한 0.3922 대비 좋은 성능을 보여주었다. 두 결과 모두 이전 T5모델과 달리 BLEU-2~ 4가 높게 유지되었으며, 'NMT(CoT)'가 전 지표에서 높은 성능을 기록하였다. 이를 통해 최근 널리 사용되는 sLLM에서 사전 정보를 활용한 CoT 학습이 효과가 있음을 알 수 있다.

6. 결론 및 향후전망

지금까지 사전 정보를 바탕으로 번역 과제의 성능을 높이는 방법들에 대해서 살펴보았다. 입력 텍스트에 번역 대상이 되는 언어에 대한 사전 정보를 추가하는 간단한 방법을 통해서 성능을 향상할 수 있다는 것을 보았다. 또한 실제 번역 레이블과 무관하게, 검색하여 나온 결과를 모두 더해주었을 때 더 높은 성능을 보여주었다는 사실 또한 확인할 수 있었다. 추가로 길이 제약을 극복하기 위해 모델에 질의응답과 같은 부가적인 과제를 학습하여 사전 정보를 내재화하였을 때 유의미한 성능 향상이 있음을 보았다.

또한 사전정보를 CoT에 활용하는 실험을 통해, COT 학습의 효과에 대해서도 살펴보았다. T5에서는 BLEU-1이 아닌 나머지 점수에서 COT 학습이 단순 번역을 학습하는 것 대비 높은 성능이 보임을 확인하였다. 또한, Llama2와 같은 sLLM 모델을 사용하여 보았다. CoT를 학습한 Llama2가 번역과제를 조정 학습한 모델 대비 BLEU-1 ~ 4에서 높은 성능을 보였다.

향후에는 보다 다양한 데이터셋에 대하여 연구가 진행되기를 바란다. 본 연구에서는 번역 과제만을 기반으로 학습 및 평가하였다. 하지만 다른 생성 과제에서도, 사전 정보와 같은 외부 지식을 활용한 언어모델 성능 향상이 가능할 것으로 보인다. 특히 이미 구축된 정보를 비지도로 구축하여 사용하는 것이 가능하며, 이에 따라 적은 비용으로 여러 과제에 대한 지식을 구축할 수 있을 것으로 예상한다.

또한, 크기가 더 큰 모델에 대하여 연구될 필요가 있다. 본 연구에서는 실험 환경의 제약으로 인해, 220만 파라미터 모델과 70억 파라미터 모델에 대해서만 실험을 진행했다. 하지만, 최근 유행하고 있는 Llama2-70B(700억)와 같은 모델에 대해서도 이러한 정보를 프롬프트로 제공하거나 조정 학습하는 것이 효과가 있는지 확인해 볼 필요가 있다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.1711139527,빅데이터 인과 분석을 위한 복잡계 기반 추론 인공지능(REX) 개발 및 실증)

참고문헌

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving Language Understanding by Generative Pre-training," 2018.

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Language Models are Few-Shot Learners," *Advanaces in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020.

[3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advanaces in Neural Information Processing Systems*, 2022.

[4] 이서정, "온라인 어학사전을 활용한 신조어 기계 번역의 정확도 향상 방법," *정보과학회논문지*, Vol. 1, No. 5, pp. 599–602, 2021.

[5] 문현석, "전이학습 기반 기계번역 사후교정 모델 검증," *한국융합학회논문지*, Vol. 12, No. 10, pp. 27–35, 2021.

[6] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 726–742, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.47

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, Vol. 21, pp. 1–67, 2020.

[8] Hugo Touvron, Louis Martin, Kevin Stone et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288,* , 2023.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Jul. 2002. [Online]. Available: https://aclanthology.org/P02-1040

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, ""Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685, 2021*, 2021.