대화 내 상호 참조 해결 모델 성능 분석서

버전 1.0

2024. 11. 06 조병길 (KETI)

Revision History

수정	시작 날짜	끝난 날짜	작성자	설명
v1.0	2024.11.06	2024.11.06	조병길	성능 분석서 초안 작성

목차

1 문서 개요

1.1 개요

본 과제에서는 복합대화 기술의 기반이 되는 대화 내 상호 참조 해결 모델을 개발하였다. 이 문서는 개발한 모델의 성능에 대해 기술한다. 대화 내 상호 참조 해결 문제는 대화에서 등장하는 대명사가 실제로 가리키는 객체를 찾는 것이며 이는 대화 문맥을 이해하는데 중요한 역할을 한다. 이 기능을 구현하기 위한 데이터셋을 구성하고 언어모델을 학습시켰다. 구체적으로 데이터셋을 수집하기 위한 툴을 크롬 브라우저의 확장 프로그램으로 구현하고, 1B, 3B, 8B의 LLAMA 모델로 학습하였다. 최종적으로 학습된 모델의 성능을 4지선다 문제로 평가하였다.

1.2 구성 및 범위

구체적으로 본 문서는 다음과 같은 범위와 내용을 기술한다.

- 상호 참조 해결 문제의 정의 (2. 문제 정의)
- 본 과제에서의 대화 내 상호 참조 해결 문제의 정의
- 학습 데이터 생성을 위한 크롬 확장 프로그램 설명
- 상호 참조 모델을 위한 언어 모델 설명
- 학습 파라미터 설정
- 실험 결과 및 모델 성능 분석
- 향후 보완 계획

2 문제 정의

2.1 상호 참조 해결

상호 참조 해결(Coreference resolution)은 텍스트 내에서 동일한 개체를 지칭하는 다양한 표현들을 식별하고 연결하는 자연어처리 문제이다. 대명사, 지시어, 동의어 등 서로 다른 형태로 언급된 같은 개체들을 파악하여 텍스트의 문맥적 일관성을 이해하는 데 중요한 역할을 한다. 예를 들어, "고양이가 마당에 있었다. 그 동물은 잠을 자고 있었다."라는 문장에서 '고양이'와 '그 동물'이 같은 대상을 가리킨다는 것을 인식하는 것이다. 이러한 상호 참조 해결 능력은 기계 번역, 질의응답 시스템, 텍스트 요약 등 다양한 지식 기반 자연어처리 응용 분야에서 핵심적인 요소로 활용된다.

2.2 멀티 모달 대화 내 상호 참조 해결 – 본 과제에서 설정한 문제

본 과제에서 설정한 멀티 모달 대화 내 상호 참조 해결(Multi-modal Dialogue Coreference Resolution)은 이미지와 텍스트가 함께 제시된 대화에서 대명사나 지시어가 이미지 내의 어떤 객체를 지칭하는지 파악하는 자연어처리 문제이다. 그러므로 이를 해결하기 위해서는 시각 정보와 언어 정보를 동시에 처리하여 대화의 맥락을

이해해야 한다. 예를 들어, "강아지와 고양이가 있는 사진을 보여주며 ' 그 녀석 장난감을 정말 재미있게 가지고 노네 '라고 말할 때, ' 그 녀석'이 사진 속 강아지와 고양이 중 어느 것을 가리키는지 결정하는 것이다. 이러한 멀티 모달 상호 참조 해결은 시각 대화 시스템, 이미지 기반 질의응답, 로봇과의 대화 등 멀티 모달 인터랙션이 필요한 다양한 응용 분야에서 정확한 의사소통을 위한 핵심 기술로 활용될 수 있다.

[이미지] (강아지와 고양이가 함께 장난감을 가지고 노는 사진을 보여주며)

사람 A: "저기 갈색 강아지랑 하얀 고양이가 같이 노는 모습 봐"
사람 B: "그 녀석 장난감을 정말 재미있게 가지고 노네"
사람 A: "응, 우리 강아지가 장난감 가지고 노는 걸 좋아해서 그래"
사람 B: "근데 옆에 있는 애는 좀 무관심해 보이네?"

멀티 모달 대화 내 상호 참조 해결

- "그 녀석" → 갈색 강아지
- "옆에 있는 애" → 하얀 고양이.
  이를 위해 모델은 이미지의 시각 정보(동물의 종류, 색상, 위치, 행동)와 대화의 문맥을 모두 이해하고 연결해야 함.

<멀티 모달 대화 내 상호 참조 해결 예시>

본 성능평가에서는 평가의 용이성을 위해 객관식으로 아래와 같이 문제를 바꾸었다. 또한 아직 초기 단계이므로, 이미지에 대한 설명을 텍스트로 변환하여 문제에 삽입하는 형식으로 구성하였다.

[이미지] (강아지와 고양이가 함께 장난감을 가지고 노는 사진을 보여주며)

사람 A: "저기 갈색 강아지랑 하얀 고양이가 같이 노는 모습 봐"
사람 B: "그 녀석 장난감을 정말 재미있게 가지고 노네"
사람 A: "응, 우리 강아지가 장난감 가지고 노는 걸 좋아해서 그래"
사람 B: "근데 옆에 있는 애는 좀 무관심해 보이네?"

- 문제: 마지막 사람 B 의 “옆에 있는 애”가 가리키는 것은?

  A) 사진 속 고양이
  B) 사진 속 강아지
  C) 사람 속 고양이와 강아지
  D) 사진 속 장난감

- 정답: A) 사진 속 고양이

<객관식 형태의 멀티 모달 대화 내 상호 참조 해결 예시>

3 데이터 수집 방법

3.1 프롬프트 엔지니어링 기반 데이터 생성

멀티 모달 대화 내 상호 참조 해결을 위한 대규모 데이터셋은 현재 존재하지 않으며, 이를 수동으로 구축하기 위해서는 많은 시간과 비용이 필요하다. 데이터셋 구축을 위해서는 다양한 이미지에 대한 여러 사람들의 자연스러운 대화를 수집하고, 각 대화에서 발생하는 대명사나 지시어가 이미지의 어떤 객체를 지칭하는지 일일이 레이블링해야 하는 복잡한 작업이 요구된다. 이러한 한계를 극복하기 위해 GPT-4와 같은 대규모 언어 모델을 활용하여 다양한 이미지 상황에 대한 대화 시나리오와 상호 참조 관계를 자동으로 생성하는 방법을 제안하였다. 이는 데이터 구축 비용을 크게 절감하면서도 풍부하고 다양한 학습 데이터를 확보할 수 있는 효과적인 대안이 된다.

이를 위해 다음과 같이 프롬프트를 작성하였다.

```
# 테스크 설명

------

A: ~(사진을 업로드하며 자연스러운 발화)~

B: ~

A: ~

...(중략)...

B: ~(사진을 업로드하며 자연스러운 발화)~

A: ~

B: ~

...(중략)...

A 혹은 B: (두개의 사진 중 하나에 포함되는 물체를 대명사로 지칭하는 발화)

문제: 마지막 A의 발화에서 대명사가 지칭하는 것은 A와 B중 어느 사진에 포함된 어느 물건인가?

{문제 풀이 과정. 정답에 대한이유와 정답이 아닌것에 대한 이유 포함.}

{선택지}

## Json 출력
```

{

...

}

위와 같은 사진에대한 대화 형식으로 coreference resolution 문제를 만들어줘.

중요: 그리고 최종적으로 Json 형식으로 출력해야함.

예시

설정

A의 사진: 책상 위에 놓인 노트북과 헤드폰

B의 사진: 소파 위에 놓인 책과 커피잔

정답: B의 사진에 있는 커피잔을 지칭하는 발화로 설정

마지막 지칭하는 발화 하는 사람: A

대화:

A: "내가 요즘 집에서 일할 때 찍은 사진을 올릴게."

(사진 1: 책상 위에 노트북과 헤드폰이 놓여있는 사진)

B: "오, 집에서 일할 때 저런 세팅이 편하겠다! 나도 방금 찍은 사진 하나 올릴게."

(사진 2: 소파 위에 책과 커피잔이 놓여있는 사진)

A: "응, 노트북이 있어서 일하기 편해. 헤드폰도 집중할 때 정말 유용해."

B: "나도 책 읽을 때 커피 한 잔이면 더 집중이 잘 되는 것 같아."

A: "맞아, 커피 한 잔이면 에너지도 충전되고 더 오래 일할 수 있지."

B: "그렇지! 이 분위기에서 책을 읽으면 정말 힐링이야."

A: "나도 그런 여유를 즐기고 싶어. 특히 저거 보니까 딱 그런 느낌이 들더라."

문제: 마지막 A의 발화에서 '저거'는 A와 B 중 어느 사진에 포함된 어느 물건인가?

문제 풀이 과정:

정답에 대한 이유: A는 마지막 발화에서 '저거'라고 말하며 B의 사진에 대한 언급을 하고 있습니다. B는 사진 속에서 책과 커피잔을 소개했지만, 이전 발화에서 커피 한 잔을 강조하며 집중에 도움이 된다고 말했습니다. 마지막으로 A가 '저거'라고 한 것은 B가 커피와 관련된 이야기를 했을 때 자연스럽게 연결되는 커피잔입니다.

정답이 아닌 것에 대한 이유:

A의 노트북: A가 올린 사진이지만, '저거'라는 표현은 자신의 물건을 지칭하는 데 적합하지 않으며, 앞서 커피에 대한 대화가 이어졌습니다.

A의 헤드폰: 이 역시 A의 물건이며 대명사 '저거'로 지칭하기에 부자연스럽습니다.

B의 책: B의 사진 속 물건이긴 하지만, 대화 흐름상 커피에 집중하는 대화가 이어졌으므로 '저거'는 책보다는 커피잔을 가리킬 가능성이 더 큽니다.

선택지:

A) A가 올린 사진 속 노트북

B) A가 올린 사진 속 헤드폰

C) B가 올린 사진 속 책

D) B가 올린 사진 속 커피잔

정답: D) B가 올린 사진 속 커피잔

Json 출력

{

"image_descriptions": ["A의 사진: 책상 위에 놓인 노트북과 헤드폰", "B의 사진: 소파 위에 놓인 책과 커피잔"],

"conversation": ["A: "내가 요즘 집에서 일할 때 찍은 사진을 올릴게 (사진 1: 책상 위에 노트북과 헤드폰이 놓여있는 사진)", "B: ~", "A: ~", ...(중략)...],

"problem": "마지막 A의 발화에서 '저거'는 A와 B 중 어느 사진에 포함된 어느 물건인가?",

"choices": ["A) A가 올린 사진 속 노트북", "B) A가 올린 사진 속 헤드폰", "C) B가 올린 사진 속 책", "D) B가 올린 사진 속 커피잔"],

"solution": "D) B가 올린 사진 속 커피잔"

}

주의 사항

- 사진 속 물건과 보기 및 정답을 먼저 설정하고 문제를 만들 것. (이전에 사용된 사진과는 다른 사진으로 설정)

- 또한 누가 그 물체를 지칭할 지 정할 것.

- 이번에는 A가 A의 사진을 "마지막 지칭하는 발화 하는 사람" 으로 정할 것.

- 각 사진에는 두 개 이상의 물체가 포함되어야 함.

- 사진을 올리며 하는 발화에는 사진 속 물체가 등장하면 안됨.

- 보기 중 하나만 명백한 정답이도록 마지막 발화를 설정해야함.

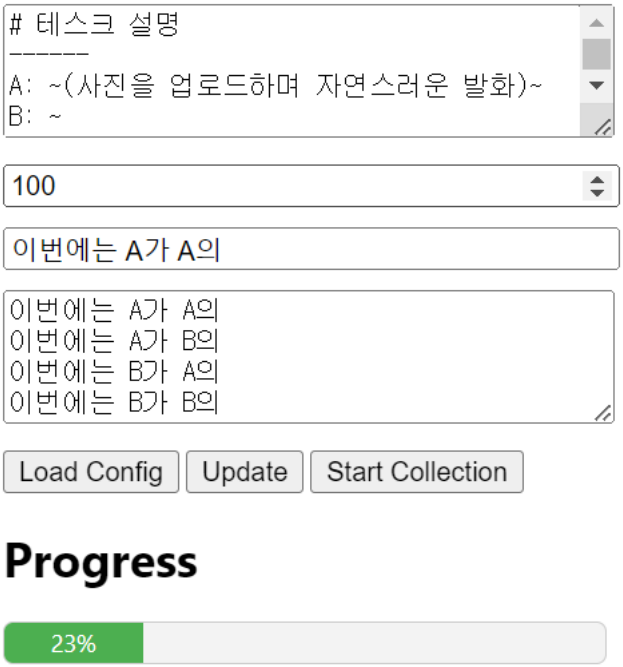
- "(중략)" 에는 몇개의 발화가 더 추가되어야 함.

- 주제는 "패션", "음식", "여행", "취미", "인테리어" 중 하나로 설정할 것

<언어 모델을 위해 데이터를 생성하기위한 프롬프트 엔지니어링>

상호 참조 문제를 효과적으로 생성하기 위해, 먼저 대화 형식과 각 참여자의 발화 패턴을 명확히 정의하는 형식을 제시하였고, 이어서 구체적인 예시를 통해 문제 생성 방식에 대한 이해를 돕도록 one-shot 형태로 구성하였다. 문제의 난이도를 적절히 조절하기 위해 각 사진에 두 개 이상의 물체를 포함하도록 조건을 설정하였으며, 특히 자연스러운 대화 흐름을 위해 사진을 공유할 때의 초기 발화에는 사진 속 물체를 직접적으로 언급하지 않도록 하고, 대화가 진행되면서 자연스럽게 물체에 대한 언급이 이루어지도록 설계하였다. 정답 도출 과정의 설명 가능성과 문제의 퀄리티 향상을 위해 Chain of Thought 방식을 적용하여 상세한 문제 풀이 과정을 포함하도록 하였다. 또한, 문제 해결의 명확성과 평가의 용이성을 위해 4지선다 형태로 선택지를 구성하였고, 생성된 문제들의 일관된 처리와 데이터 활용을 용이하게 하기 위해 최종 출력을 JSON 형식으로 지정하였다.

3.2 데이터 수집을 위한 크롬 확장 프로그램 개발



<데이터 수집을 위한 크롬 확장 프로그램>

이 이미지는 데이터 수집을 위한 크롬 확장 프로그램의 인터페이스다. 첫 번째 텍스트 필드는 프롬프트 내용을 입력하는 곳이다. 두 번째 필드는 프롬프트의 반복 횟수를 설정하는 부분이다. 세 번째 필드는 첫 번째 프롬프트에서 대체할 문자열을 지정하는 칸이다. 네 번째 필드는 변경할 문자열을 줄마다 입력하는 공간으로, 프롬프트가 반복될 때마다 다음 줄의 내용으로 자동 교체된다. 마지막 줄까지 도달하면 다시 첫 번째 줄부터 반복된다. "Start Collection" 버튼을 눌러 데이터 수집을 시작할 수 있다. 하단의 Progress 바를 통해 작업 진행 상황을 확인할 수 있다. 작업이 완료되면 수집 내역을 아래 예시와 같이 txt 파일로 저장할 수 있다.

```
## Conversation

A: "내가 오늘 만든 저녁 사진을 올릴게!"

A의 사진: 식탁 위에 놓인 다양한 음식과 음료.

B: "우와, 정말 맛있어 보인다! 나도 어제 바베큐 한 사진을 올려볼게."

B의 사진: 테라스에서의 바베큐 모습, 그릴 위의 고기와 함께 다양한 야채가 놓여 있음.

A: "그 바베큐도 정말 맛있겠다! 어떤 고기를 구웠어?"

B: "소고기랑 양념한 돼지고기를 구웠어. 야채도 함께 구워서 정말 맛있었어."
```

A: "그렇구나! 나도 그런 고기 구워서 먹고 싶어."

B: "맞아, 여름에는 바베큐가 제일이지!"

A: "그런 분위기에서 저거 먹으면 정말 좋겠어."

B: "그렇지! 특히 저건 완전 맛있었어."

A: "나도 그거 해봐야겠다!"

Problem

마지막 A의 발화에서 '저거'는 A와 B 중 어느 사진에 포함된 어느 물건인가?

Choices

A) A가 올린 사진 속 음식.

B) A가 올린 사진 속 음료.

C) B가 올린 사진 속 야채.

D) B가 올린 사진 속 바베큐 고기.

Solution

D) B가 올린 사진 속 바베큐 고기.

<언어모델로 생성된 멀티 모달 상호 참조 해결 문제 예시>

4 상호 참조 모델을 위한 언어모델

4.1 LLaMA 사전학습 모델 기반 미세조정 학습

LLaMA 사전학습 모델을 기반으로 상호 참조 해결 테스크를 미세조정 학습하였다. LLaMA 모델은 대규모 언어 모델로, 특히 사전 훈련(pretraining)이 되어 있어 적은 데이터로도 미세조정 하여 특정 분야에 활용하는데 효율적으로 사용할 수 있다. 이 모델은 방대한 양의 텍스트 데이터를 활용해 언어 이해와 생성 능력을 강화했으며, 다양한 언어 작업에서 높은 성능을 발휘할 수 있도록 설계되었다. 특히 문장 구조 파악, 의미 이해, 텍스트 일관성 유지 등에 강점을 보이는데, 이를 활용하기 위해 미세조정의 기반 모델로 선택하였다.

LLaMA 모델은 Generative Pretrained Transformer의 일종으로 텍스트 생성형 모델이므로, 문제에서 보기까지 (## Choices ... D) ...) 주어졌을 때 답변을 생성하도록 학습하였다. 구체적으로 Huggingface의 "DataCollatorForCompletionOnlyLM"를 사용하여 "## Solution"을 기준으로 후에 나오는 토큰들에만 손실함수를 계산하고 최소화하는 방식으로 학습하였다. 그런 후 생성된 텍스트에 포함된 알파벳 번호 기준으로 정답과 비교하여 상호 참조 성능을 계산하였다.

5 실험

5.1 학습 설정

학습 설정은 Llama 3.2 모델의 효과적인 훈련을 위해 다음과 같이 선택하였다. A100 GPU 8대에서 각 디바이스에서의 배치 크기는 2로 설정하였고, gradient accumulation 스텝을 4로 추가하여 유효 배치 크기를 64로 설정하였다(8x2x4). 총 학습 epoch는 30으로 설정되어 있어, 데이터가 적더라도 모델이 충분한 반복 학습을 통해 데이터의 패턴을 학습할 수 있게 하였다. 초기 학습 안정화를 위해 100 스텝 동안 워밍업 단계를 두었으며, weight decay를 0.01로 설정하여 과적합을 방지하고 일반화 성능을 높이고자 하였다. Learning rate는 2e-5로 설정되어 모델이 안정적으로 학습할 수 있도록 조정되었다. 또한, 16비트 부동소수점(fp16) 연산을 사용하여 메모리 사용을 줄이고 학습 속도를 높이고자 하였다. 학습 문제 수는 1380개, 테스트 문제 수는 346개로 하였다.

5.2 실험 결과 및 모델 성능 분석

5.2.1 모델 파라미터 별 성능

모델크기	정확도 (물체 레벨)	정확도 (사진 레벨)	정확도 (물체 레벨 exact match)
1B	0.80	0.93	0.70
3B	0.84	0.97	0.65

8B	0.81	0.95	0.68

- 3B 모델에서 가장 높은 정확도 (물체 레벨)를 달성하였음.
- 데이터가 적어서 모델 파라미터가 8B로 늘어나도 성능이 오르지 않고 오히려 0.81로 줄어듦.
- 물체 레벨에 비해 사진 레벨에서 0.1 정확도 이상 정확도가 높았음. 명시적으로 물체를 선택할 수 있는 경우가 아니라 대화의 뉘앙스로 물체를 선택해야하는 경우가 있는데, 사진이 맞았더라도 여기서 성능 하락이 컸을 것으로 예상.
- "정확도 (물체레벨 exact match)"는 보기 번호로만 정답을 계산한 게 아니라 보기를 정확하게 언급하였는 지로 정확도를 계산한 것이다. 그래서 "정확도 (물체 레벨)"에 비해 정확도가 낮다. 같은 보기를 반복하여 생성하는 경우가 가장 많았고 보기 전체를 모두 언급하는 경우도 있었다.

5.2.2 Step 별 정확도 추이

모델크기	step	정확도 (물체 레벨)	정확도 (사진 레벨)	정확도 (물체 레벨 exact match)
1B	0	0.26	0.43	0.14
	152	0.80	0.96	0.79
	326	0.79	0.92	0.67
	478	0.80	0.93	0.69
	630	0.81	0.94	0.71
3B	0	0.24	0.45	0.00
	152	0.85	0.97	0.84
	326	0.82	0.96	0.71
	478	0.85	0.98	0.63
	630	0.84	0.98	0.62
8B	0	0.15	0.45	0.01

	152	0.81	0.97	0.77
	326	0.78	0.95	0.73
	478	0.81	0.96	0.67
	630	0.81	0.96	0.68

- 630 학습 step에 걸쳐서 측정한 정화도를 정리하였다.
- 학습하지 않은 경우 (0 step) 보다 학습을 한 경우 (630 step)가 0.55 이상 정확도가 높았다. Fine-tuning 하는 것이 상호 참조 해결 문제를 푸는데 도움이 되었음을 알 수 있다.
- 전반적으로 152 step에서 가장 좋은 정확도를 보이고 그 후에는 정확도가 감소하는 경향을 보였다. Training data에 대한 over-fitting 문제라고 해석할 수 있다.

6 향후 보완 계획

향후 연구에서는 현재 1,726개에 불과한 데이터 수를 확충하여 보다 신뢰성 있는 모델 성능 평가를 가능하게 할 예정이다. 또한, 텍스트 외에 실제 이미지 모달리티를 통합하여 상호 참조 해결 성능을 향상시킬 계획이다. 이를 위해 이미지에서 텍스트 설명을 생성하여 대화 텍스트에 추가하거나, 이미지와 텍스트를 융합할 수 있는 레이어가 포함된 모델을 적용하는 등 다양한 방안을 모색할 것이다.

7 Appendix: 8B-Instruct

Correct count: 271/346 = 0.78

Correct count (photo level): 325/346 = 0.93

Correct count (exact): 258/346 = 0.74

8B-Instruct chat-template	0	0.00	0.38	0.00
	152	0.79	0.94	0.75
	326	0.80	0.95	0.76
	478	0.78	0.93	0.74
	630	0.78	0.94	0.75

- 정답이지만 문장을 길게 말하는 경우가 있어 오히려 점수가 높지 않았음.

예시1: 마지막 A의 발화에서 '저거'는 B가 올린 사진에 포함된 완성된 요리에 대한 언급입니다. 따라서 정답은 D)

B가 올린 사진 속 완성된 요리입니다.

예시2: 마지막 A의 발화에서 '저거'는 B가 올린 사진에 포함된 손뜨개 작품을 가리키는 것으로 보입니다. 따라서 정답은 D) B가 올린 사진 속 손뜨개 작품입니다.