Image captioning 데이터와 Visual QA 데이터를 활용한 질문 자동 생성*

이경호○, 최용석, 이공주
충남대학교
gyholee@gmail.com, yongseok.choi.92@gmail.com, kjoolee@cnu.ac.kr

Automatic question generation based on image captioning data & visual QA data

Gyoung Ho Lee○, Yong Seok Choi, Kong Joo Lee
Chungnam national University

요 약

대화형 시스템이 사람의 경청 기술을 모방할 수 있다면 대화 상대방과 더 효과적으로 상호작용 할 수 있을 것이다. 본 논문에서는 시스템이 경청 기술을 모방할 수 있도록 사용자의 발화를 기반으로 질문을 생성하는 것에 대해 연구하였다. 그리고 이러한 연구를 위해 필요한 데이터를 Image captioning과 Visual QA 데이터를 기반으로 생성하고 활용하는 방안에 대해 제안한다. 또한 이러한 데이터를 Attention 메커니즘을 적용한 Sequence to sequence 모델에 적용하여 질문을 생성하고, 생성된 질문의 질문 유형을 분석하였다. 마지막으로 사람이 작성한 질문과 모델의 질문 생성 결과 비교를 BLEU 점수를 이용하여 수행하였다.

주제어: 질문 생성, Sequence to Sequence, Image Captioning, Visual QA

1. 서론

다른 사람의 말을 듣는 방식 중 '공감적 경청'이 있다[1]. 상대방의 말을 집중해서 들으면서 상대로 하여금 자신의 이야기를 더 많이 끌어내도록 격려하는 것이다[2]. 이를 통해 상대방은 자신이 이해 받고 있다는 느낌을 가지게 된다[1]. 이러한 '공감적 경청'은 '더 말하게 하기', '반복하기', '명확히 하는 질문하기', '열린 질문하기', '격려하기' 등을 통해 이루어진다[3]. 대화형 시스템이 이러한 듣기의 기술을 모방할 수 있다면 대화 상대방이 더 깊은 감성 교류를 이루고 있다고 믿게 할 수 있다. 본 논문에서는 시스템이 사용자의 발화를 기반으로 '명확히 하는 질문하기'와 같은 새로운 질문을 생성하여 상대방과 상호 작용할 수 있도록 하는 방안에 대해 연구하였다. 아래 예와 같이, 대화에서 사용자가 어떠한 상황에 대해 설명하는 경우가 있다. 이때, 시스템이 예의 질문과 같이 다시 질문을 할 수 있다면 사용자와 계속 대화를 이어 나갈 수 있다.

사용자: There is a birthday cake
on top of a pink table
시스템: What is being celebrated here?

최근 영상과 자연언어를 결합한 연구가 활발히 진행되고 있다. 이러한 연구에는 이미지 캡션닝(Image Captioning, IC), 비주얼 질의응답(Visual Question Answering, VQA) 등이 있다. 연구를 위해 많은 연구자들이 이미지들을 수집하였고, 수집된 이미지에 대한 설명을 작성한 데이터 셋[4]과 그 이미지로부터 떠오르는 질문과 답변에 대한 데이터 셋[5]들을 만들었다. 본 연구에서는 이러한 데이터들을 '질문하기' 시스템에 활용해보고자 한다.

본 연구를 통해 도달하고자 하는 목표는 상대방이 발화한 정보에 기초한 질문을 생성하는 것이다. 이를 통해 대화 상대의 발화를 이끌어내고 대화를 지속해 나가는 것을 목표로 한다. 이러한 목표를 위해서는 어떤 정보를 담고 있는 발화 또는 문장이 있고, 그 정보의 범위에서 크게 벗어나지 않는 관계를 가진 질문이 필요하다. VQA를 위한 데이터 [5](VQA DATA)는 이미지 캡션을 위한 [4](COCO DATA)을 기반으로 한다. 그렇기 때문에 이 두 데이터를 조합하면 하나의 이미지에 대한 설명과 그 이미지에 대한 질문을 수집할 수 있다. 설명과 질문은 같은 그림을 매개로 하기 때문에 서로 담고 있는 정보 사이의 거리가 멀지 않을 것으로 생각 할 수 있다. 본 논문에서는 이러한 가정하에, 문장에서 질문을 생성하는 시스템에서 이 데이터의 활용 가능성에 대해 살펴보고자 한다.

이전에도 이와 유사한 데이터를 이용하여 문장으로부터 질문을 생성하려는 시도가 있었다[6]. 하지만 [6]의 목적은 이미지로부터 질문을 생성하는 것이었고 그에 대한 비교 모델로 IC, VQA와 유사한 데이터를 사용하였다. 본 논문에서는 이를 좀 더 발전시켜, 이 데이터들을 활용하여 문장에서 질문을 생성하기에 적합하도록 데이터를 정제하는 방법을 연구하였고 이를 통해 생성된 질문을 평가하였다. 이를 통하여 IC, VQA 데이터가 질문 생성에서 활용할 수 있음을 확인한 것이 본 논문의 학술적 기여이다.

2. 관련 연구

* 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단 - 이공분야기초연구-여성과학자 지원을 받아 수행된 연구임(No. 2015051685).

표 1. 질문-설명 예

Image id	Question	Description
78077	Why does the man need the umbrella if it's not raining?	- A thin man slouches on a bench while holding an umbrella. - A sitting man shading himself from the sun with an umbrella. - A person holding an umbrella sitting on the bench. - A man sitting on a bench under an umbrella. - A man sits on a bench, protected by a large umbrella.
457383	What color are the birds on the table?	- A bunch of birds on a table with a glass. - Pigeons fighting over food on a table at an outdoor cafe. - A flock of black birds sitting on top of a table. - Many pigeons gather on a restaurant table. - A group of people sitting at some dining tables having lunch
318857	Where are the magazines in this picture?	- A man sitting on a couch with a cat in his lap playing on a computer - The kitten is interested in what the computer is doing. - A man in glasses using a laptop keyboard - A man with a kitten on his lap uses his laptop. A guy is on his computer with a kitten in his lap.

문장을 기반으로 질문을 생성하는 최근의 연구로 [7]이 있다. 이 연구는 “Kenya is located in Africa.”와 같은 문장을 학생이 읽었다면, 이에 대한 내용 이해 평가를 할 수 있는 “Where is Kenya located?”와 같은 질문을 생성하는 것이 목표이다. 이 연구에서 제안한 시스템은 3가지 단계를 거쳐 질문을 생성한다. 이 연구에서는 평서문에 어휘를 바꾸고, 구문적, 의미론적 변형을 가해 적절한 질문을 만드는 것을 목적으로 한다. 첫 번째 단계에서는 입력된 평서문을 자연언어처리의 다양한 기법을 이용하여 분석하고 구조화 한다.

시스템의 2단계에서는 1단계에서 분석되고 구조화된 정보를 WH-movement, subject-auxiliary inversion 등 미리 정의된 규칙을 기반으로 질문을 생성한다. 다양한 규칙을 적용하여 여러 개의 후보 질문을 만든 후, 3단계에서 Ranking 모델을 적용하여 생성된 후보 질문들의 순위를 결정하고 최종적인 질문을 선택한다. 성능 평가를 위해 학습과 평가 데이터를 구축하고, 모델에 적용하여 생성된 질문을 인간 평가자가 정해진 규정에 따라 평가하도록 하였다.

본 논문도 이 연구와 같이 문장으로부터 질문을 생성하는 것이 목표이다. 하지만 본 논문의 목적은 지식평가보다는 대화를 계속 이어나갈 수 있도록 하는 질문을 생성하는 것이기 때문에 서로 목적하는 바가 다르다. 또한 이 연구는 규칙을 기반으로 하지만 본 연구에서는 데이터 기반의 교사학습 방법으로 질문을 생성하는 방법에 대해 연구하였다.

3. 본론

3.1 데이터

본 논문에서는 VQA DATA와 COCO DATA를 이용한 질문 생성 방법에 대해 연구하였다. VQA DATA는 IC와 관련하여 많이 활용되고 있는 COCO DATA를 기반으로 한다. VQA DATA에는 COCO DATA의 이미지와 관련된 질문과 답변을 가지고 있다. 본 연구에서는 VQA DATA의 Real Image 관련 “Training annotations 2015 v1.0” 데이터의 90%를 training set으로, 10%를 development set으로 사용하였다. 평가를 위해서 VQA DATA의 Real Image 의 “Validation annotations 2015 v1.0” 데이터를 사용하였다. VQA DATA의 질문의 대상 이미지 ID를 이용하여 COCO DATA에서 이미지에 대한 설명들을 찾아 질문과 설명을 연결하였다. 이러한 과정을 통해, 학습 및 개발을 위한 [질문, list of 설명문] 쌍 253,395개와 평가를 위한 데이터 123,900개를 수집하였다. 수집된 질문과 설명에 대한 예는 표 1과 같다.

질문 수집을 위해 사용한 VQA DATA는 이미지와 관련된 질문을 담고 있다. 그렇기 때문에 이 데이터에는 표 1의 마지막 질문과 같이 이미지에서 직접적으로 관련된 답을 찾도록 요구하는 질문들이 존재한다. 본 논문에서는 설명으로부터 일반적인 질문을 생성하는 것을 목표로 한다. 그렇기 때문에 이러한 질문들은 본 논문의 목적과 거리가 멀다. 이러한 질문을 배제시키기 위해 “photo”, “picture” 등의 단어가 들어간 경우 학습과 평가에서 제외하였다.

하나의 이미지에 여러가지 질문과 설명이 있을 수 있다. 동일한 객체에 대한 질문과 설명을 수집하기 위해 두 문장에 같은 명사가 있는 경우를 학습과 평가에 사용하였다. 추가적으로, 동일한 이미지에 대한 [질문, 설명문]이 존재하는 경우 하나의 [질문, 설명문]을 선택하여 학습과 평가에 사용하였다. 또한 질문이나 설명문의 길이가 일정 길이 이상인 경우 학습 및 평가에서 제외시켰다.

본 논문에서 관심 있는 대상은 질문이다. 학습 데이터에 포함된 질문 유형들의 분포는 그림 1과 같다.

본 논문에서는 이중 가장 많은 분포를 나타낸 “What” 질문 유형에 대하여 실험과 평가를 진행하였다.

최종적으로, 학습과 개발, 평가에 사용한 [질문, 설명문] 데이터의 수는 표 2와 같다.

표 2 학습, 개발, 평가 데이터 수

학습	개발	평가
17,579	1,892	9,668

3.2 Seq-to-Seq 모델

이미지에 대한 설명으로부터 질문을 생성하기 위해 [8]의 모델을 수정하여 사용하였다. 본 논문에서 사용한 모델은 그림 2와 같이 구성된다.

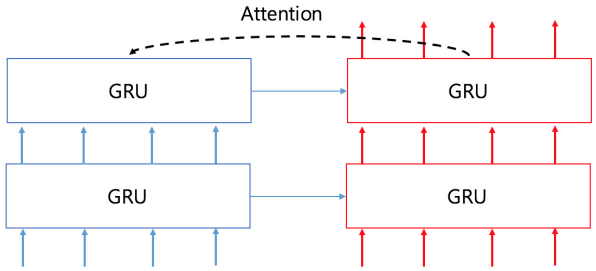

그림 2 모델

이 모델은 Attention 메커니즘을 가지고 있는 Sequence-to-Sequence 모델을 기반으로 한다. 모델은 Encoder와 Decoder에 각각 2개 Recurrent neural network Layer(RNN)을 가진다. 각 RNN 레이어는 Gated Recurrent Unit(GRU)로 구성된다. 모델은 설명문과 질문을 이용하여 학습된다. 문장을 구성하는 각 단어들은 미리 학습된 Word2Vec[9]모델의 임베딩 벡터로 표현된다. 모델에 설명문을 입력하면 각 단어의 자질이 순차적으로 Encoder에 입력되어 각 레이어에서 설명문 전체의 정보를 담고 있는 Hidden State와 각 단어들의 Encoder 출력값이 결정된다. Encoding이 완료되면, Decoder는 Decoding 작업이 시작되었다는 입력과 함께 Encoder의 각 층으로부터 전달 받은 Hidden State 및 Encoder의 출력을 Attention 메커니즘과 결합하여 질문의 시작 단어를 선택한다. Decoding된 첫 단어는 다시 Decoder의 다음 시간 입력으로 들어가고 첫 단계와 같은 과정을 통해 두 번째 단어를 예측한다. 이 과정은 문장의 끝을 나타내는 표지가 나타날 때까지 반복된다.

모델의 효과적인 학습을 위해 Encoder와 Decoder의 길이를 설정하고 설명문과 질문을 이들에 맞게 조정하였다. 문장이 Encoder나 Decoder의 길이보다 짧을 경우, padding을 추가하여 길이를 맞추어 주었다. 또한 질문에 Decoding의 시작을 알려주는 표지와 끝을 알려주는 표지를 문장 앞뒤에 추가해 주었다. 파라미터 학습을 위해서 기울기하강법(stochastic gradient descent)을 이용하였고 초기 학습률로 0.5를 사용하였다. 학습률은 일정한 조건에 따라 0.99의 비율로 감소하도록 하였다. 전체 학습 데이터에 대하여 100번의 iteration을 반복하였다. 기타 파라미터들은 [8]의 설정을 따랐다.

4. 실험 및 평가

질문 생성은 그 결과를 정확히 평가하기에 어려운 점이 많다. 그렇기 때문에 본 논문에서는 평가 데이터에서 이미지에 대한 설명문(Source)과 이미지에 대해 사람이 작성한 질문(Ref), 그리고 Source와 모델을 통해 생성된 질문(Hyp)사이의 관계를 통해 간접적으로 생성된 질문의 유용성에 대해 판단하고자 한다.

4.1 질문 생성 결과

평가 데이터와 그 평가 데이터를 모델에 적용하여 생성한 질문의 예를 표 3에 나타내었다.

표 3의 첫 번째 결과는 Hyp가 Ref와 비슷한 목적을 가지는 질문이 생성된 결과이다. 두 번째 결과는 서로 다른 목적을 가진 질문이지만 Source에 대한 질문으로써 절절한 수준으로 판단된다. 가진 3번째 결과의 경우 Ref와 동일한 문장이 Hyp에서 생성되었다. 4번째 결과에서 Hyp 자체는 문제가 없지만, Source에서 이미 꽃의 색깔을 밝혔기 때문에 의미 있는 질문으로 보기 어렵다. 마지막 예는 Hyp가 올바른 문장을 생성하지 못한 경우이다.

Seq-to-Seq 모델을 통해 생성된 질문과 비교하기 위한 모델로 Near 모델을 정의하였다. Near 모델은 학습데이터에서 사용된 질문을 기반으로 한다. Source에 대한 질문을 생성하기 위해 Near모델은 학습데이터의 질문들과

표 3 질문 생성 예

설명문(Source)	질문(Ref)	생성 질문(Hyp)	Near 모델
a kitchen with a sink a stove and cupboard	what is the item on the stove ?	what kind of appliance are on the stove ?	what color is the kitchen sink ?
a group of three men standing next to each other .	what is the color of the men 's jacket ?	what are the people doing ?	what video game are they playing ?
a group of people who are playing video game .	what game are they playing ?	what game are they playing ?	What video game are they playing?
several pink flower in a large gray vase .	what color is the vase ?	what color is the flower ?	What color vase is the pink flower in?
an intersection with two street sign near a palm tree .	what kind of tree is that ?	what color is the sign sign ?	what number of tree line this sidewalk ?

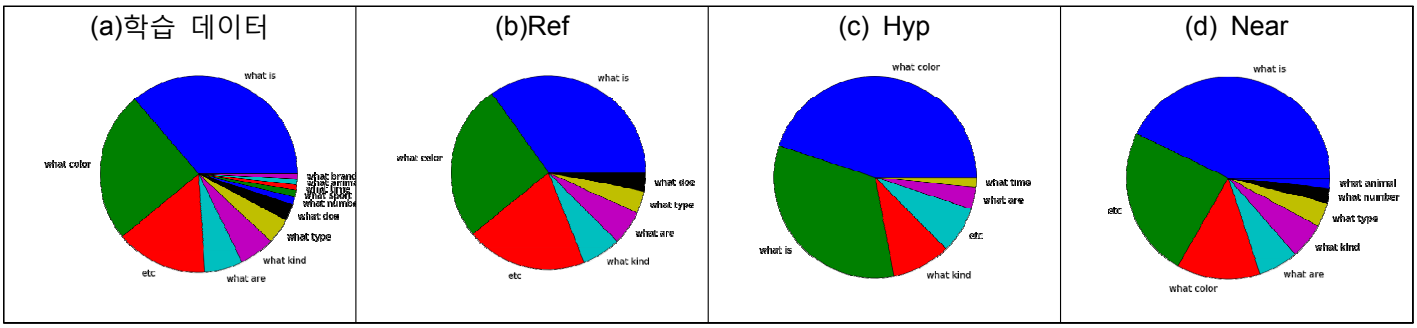

그림 3 질문 유형 분포 2

Source의 유사도를 계산한다. 유사도 계산 결과, Source와 가장 유사한 질문을 Source의 질문으로 결정한다. 질문과 Source를 자질 벡터로 표현하고 이들 사이의 cosine similarity를 계산하여 문장 간 유사도로 사용하였다. 문장의 자질 벡터는 문장을 구성하는 단어들의 단어 임베딩 합으로 계산 된다. Near 모델의 결과 예를 표 3의 Near 모델 열에 나타내었다.

평가 데이터에서 Source, Ref, Hyp, Near의 평균 단어 수는 각각 11.1, 7.3, 6.7, 8.3로 나타났다.

4.2질문 유형 분포

본 논문은 수집된 질문 중 'What' 유형에 대해 질문 생성을 수행하였다. 그림 3은 학습데이터와 평가 데이터, 그리고 Seq-to-Seq모델과 Near 모델에서 생성한 질문에서 어떠한 "what" 질문의 유형이 분포하는지 나타낸 표이다. 학습에 사용한 데이터에서 "What is" 유형이 36.2%로 가장 많았고 "What color"가 24.8%로 그 뒤를 이었다(a). 평가데이터에서도 "What is"가 가장 많은 34.8%, "What color" 유형이 26.2%로 같은 순위가 유지 되었다(b). 하지만 Seq-to-Seq 모델에서 생성한 질문에서는 "What color"가 45.0%로 가장 많았고 "What is"가 33.1%로 두 번째로 많은 분포를 나타냈다(c). Near 모델에서는 "What is" 질문과 "What color" 질문이 42.9%와 13.2%로 나타났다. 전반적으로, 생성된 질문의 질문 유형은 기존의 학습 및 평가 데이터의 질문 유형과 비슷한 분포를 나타냈다. 이를 통해 생성된 질문이 사람이 만든 질문의 유형의 범주와 크게 다르지 않다는 것을 알 수 있다.

질문의 생성 결과에서 "What color"가 좀 더 높게 나타난 이유를 살펴보았다. 학습 데이터에서 설명이 색과 관련된 표현("red", "black", "yellow" 등 16가지 색)을 포함하고 있는 경우가 3,132개, 이중 질문 유형이 "What color"인 경우는 37.5%로 가장 높은 분포를 차지한다. 이에 이어서 "What is" 질문이 29.9%로 나타났다. 평가 데이터에서 Source에 색과 관련 표현이 있고(1,773개), 이때 Ref가 "What color"인 경우는 39.3%, Hyp가 "What color" 유형인 경우는 67.7%로 나타났다. 이 결과를 통해 질문을 생성할 때 Source에 색과 관련이 있는 표현이 있다면, Seq-to-Seq모델의 질문 생성에서 "What color"라고 물어보는 경우가 높다는 것을 알 수 있다. 이는 학습데이터에서 색과 관련된 표현이 설명에 있을 때 "What color" 질문의 유형이 가장 높은 분포를 나타냈고, 이러한 관계를 학습과정에서 모델이 강하게 학습한 결과라고 추측된다.

4.3 BLEU 점수 평가

Ref와 Hyp, Ref와 Near 사이의 유사성을 확인하기 위하여 Smoothed-BLEU 점수[10][11]를 계산하였다. Ref는 Source와 연관된 이미지를 기반으로 사람이 생성한 질문이다. 그렇기 때문에, Ref와 더 유사 할수록 Source와 더 연관된 질문으로 생각 할 수 있다. 표 4는 사람이 생성한 질문과 Seq-to-Seq모델이 생성한 질문과의 BLEU 점수(Ref-Hyp)와, 사람의 질문과 Near 모델의 질문 사이의 점수(Ref-Near)를 나타내었다. 이 결과에서 Near 모델보다 Seq-to-Seq 모델이 더 높은 BLEU 점수를 나타냈다. 이를 통해 모델이 생성한 질문이 단순한 기본 모델보다 더 나은 생성 결과를 나타냈다고 볼 수 있다.

표 4. BLEU Score 결과

유형	Ref-Hyp	Ref-Near
점수	0.322	0.263

본 논문의 방식과 유사한 이미지를 이용하여 이미지로부터 질문 생성하는 [6]의 결과에서 BLEU점수는 0.192를 나타냈다. 본 논문의 모델들이 이 모델에 비해 BLEU 점수가 높게 나온 이유는 본 논문에서 실험한 질문 유형이 제한적이고 그에 따라 나올 수 있는 질문의 폭이 상대적으로 좁았기 때문으로 판단된다.

5. 결론

본 논문에서는 Image captioning과 Visual QA 에 사용되는 데이터를 활용하여 문장에서 질문을 생성하는 것에 대해 제안하고 그 결과에 대해 평가해 보았다. 질문을 생성하는 방법으로는 Attention 메커니즘을 가지는 Sequence-to-Sequence 모델을 사용하였다. 이 데이터와 모델을 이용해 생성된 질문의 질문 유형 분포를 분석하고 사람이 만든 질문과 유사성을 비교하였다. 그 결과, Seq-to-Seq 모델을 통해 생성된 결과가 사람이 만든 질문들과 유사한 질문 유형 분포를 나타냄을 보았고, 사람이 만든 질문과 모델이 만든 질문의 표면적인 유사성 정도를 확인하였다. 이러한 분석을 통해 본 논문에서 제안한 이미지 관련 데이터의 활용 가능성을 확인하였다.

하지만, 학습과 평가를 위해 사용한 질문에 이미지를 확인해야만 답이 가능한 질문 유형이 포함되는 경우가 있었다. 본 논문에서 제안하는 데이터를 통한 질문 생성의 성능을 높이려면, 본 논문에서 추구하는 바를 만족하는 질문을 필터링할 수 있는 패턴에 대한 연구가 좀 더 발전해야 한다고 생각된다. 향후 이러한 데이터와 함께, 질문과 설명문 사이를 연결하는 이미지로부터 더 많은 정보를 추출하고 이를 적절히 질문 생성에 적용 할 수 있다면 양질의 다양한 질문을 생성할 수 있을 것으로 기대되고 향후 계속 연구를 진행할 계획이다.

참고문헌

[1] 백미숙. "종설: 공감적 경청의 자세와 주요 기술." 의료커뮤니케이션 1.1 (2006): 18-26.

[2] 김소은. "말하기와 듣기의 통합적 수업을 위한 서사적 대화 형태로서의 수업 모형 제안." 교양교육연구 8.1 (2014): 375-431

[3] 존 스튜어트・캐런 제디커・사스키아 비테본, 소통: 협력적인 의사소통의 방법-사회구성주의적 접근, CommunicationBooks, 2015. 12. 28., 270p

[4] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European Conference on Computer Vision. Springer International Publishing, 2014.

[5] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[6] Mostafazadeh, Nasrin, et al. "Generating Natural Questions About an Image."arXiv preprint arXiv:1603.06059 (2016).

[7] Heilman, Michael, and Noah A. Smith. Question generation via overgenerating transformations and ranking. No. CMU-LTI-09-013. CARNEGIE-MELLON UNIV PITTSBURGH PA LANGUAGE TECHNOLOGIES INST, 2009.

[8] Vinyals, Oriol, et al. "Grammar as a foreign language." Advances in Neural Information Processing Systems. 2015.

[9]Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[10] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[11] Chen, Boxing, and Colin Cherry. "A systematic comparison of smoothing techniques for sentence-level BLEU." ACL 2014 (2014): 362.