한국어 대화 모델 학습을 위한 디노이징 응답 생성

김태형○, 노윤석, 박성배, 박세영
경북대학교
{thkim, ysnoh, sbpark}@sejong.knu.ac.kr, seyoung@knu.ac.kr

Denoising Response Generation for Learning Korean Conversational Model

Tae-Hyeong Kim○, Yunseok Noh, Seong-Bae Park, Se-Yeong Park
Kyungpook National University

요 약

챗봇 혹은 대화 시스템은 특정 질문이나 발화에 대해 적절한 응답을 해주는 시스템으로 자연어처리 분야에서 활발히 연구되고 있는 주제 중 하나이다. 최근에는 대화 모델 학습에 딥러닝 방식의 시퀀스-투-시퀀스 프레임워크가 많이 이용되고 있다. 하지만 해당 방식을 적용한 모델의 경우 학습 데이터에 나타나지 않은 다양한 형태의 질의문에 대해 응답을 잘 못해주는 문제가 있다. 이 논문에서는 이러한 문제점을 해결하기 위하여 디노이징 응답 생성 모델을 제안한다. 제안하는 방법은 다양한 형태의 노이즈가 임의로 가미된 질의문을 모델 학습 시에 경험시킴으로써 강건한 응답 생성이 가능한 모델을 얻을 수 있게 한다. 제안하는 방법의 우수성을 보이기 위해 9만 건의 질의-응답 쌍으로 구성된 한국어 대화 데이터에 대해 실험을 수행하였다. 실험 결과 제안하는 방법이 비교 모델에 비해 정량 평가인 ROUGE 점수와 사람이 직접 평가한 정성 평가 모두에서 더 우수한 결과를 보이는 것을 확인할 수 있었다.

주제어: 대화 모델, 시퀀스-투-시퀀스 모델, 자연어 생성

1. 서론

챗봇(chatbot) 혹은 대화 시스템은 자연어 질의에 대해 적절한 자연어 응답을 해주는 시스템이다. 이를 위해서는 두 가지 핵심 기술이 필요한데, 하나는 자연어로 된 질의를 정확히 이해할 수 있는 기술이며, 다른 하나는 입력받은 자연어 질의에 적합한 자연어 응답을 생성하는 기술이다. 이 두 기술을 하나의 모델로 구현하는 문제가 최근 널리 사용되는 딥러닝 아키텍처인 시퀀스-투-시퀀스(sequence-to-sequence) 모델 [1]과 구조적으로 잘 맞아 떨어지기 때문에, 대화 응답 모델 개발의 기본 모델로써 시퀀스-투-시퀀스 모델이 널리 사용된다 [2,3,4,5].

대화 시스템을 위한 시퀀스-투-시퀀스 모델은 입력받은 자연어 질의를 벡터 표현으로 변환하는 인코더와 인코딩된 질의 문장에 대한 벡터 표현을 입력받아 적절한 자연어 응답을 생성하는 디코더로 구성된다. 모델의 입력과 출력이 모두 자연어이므로 시퀀스 데이터를 다루기에 적합한 순환 신경망(recurrent neural network)이 인코더와 디코더에 사용된다. 대화 쌍 데이터가 충분한 경우 이러한 기본 형태의 시퀀스-투-시퀀스 모델로도 어느 정도 잘 동작하는 대화 모델을 학습할 수 있다는 것이 최근 학계에 보고되었으며 [2], 따라서 이를 기반으로 여러 추가 연구들이 발표되고 있다 [3,4,5].

본 연구에서는 이러한 시퀀스-투-시퀀스 모델을 통한 대화 응답 생성 모델이 잠재적으로 내포할 수 있는 문제를 다룬다. 표 1에서 볼 수 있는 것처럼 같은 의미를 나타내는 다른 자연어 질의에 대해 시퀀스-투-시퀀스 모델은 강건하지 못한 응답을 생성할 수 있다. 자연어는 그 특징상 같은 의미를 나타내는 수많은 다른 표현이 존재할 수 있다. 특히 한국어의 경우, 문장 내 표현이 동의어로 교체되는 변화 외에도 상대적으로 더 자유로운 어순 변화와 다양한 어미 변화가 가능하다. 이러한 자연어의 본질적인 특성으로 인해 기존 시퀀스-투-시퀀스 모델은 표 1과 같이 학습 시 접하지 못한 수많은 새로운 질의에 대해 제대로 된 응답을 생성할 수 없게 된다. 본 논문에서는 이런 현상을 (*말귀 못 알아듣는*) *사오정 문제* 로 명명한다. *사오정 문제*가 발생하는 근본적인 이유 중 하나는 시퀀스-투-시퀀스 모델의 인코더가 처음 보는 질의 시퀀스에 대해 핵심 의미를 잘 담고 있는 강건한 벡터 표현을 도출하지 못하기 때문이다. 이 문제는 비단 대화 데이터뿐만 아니라 모든 자연어 이해 문제에 공통적으로 나타나는 것이며, 상대적으로 학습 데이터가 작

예시 대화
Q: 공항 안에 커피를 마실 만한 곳이 *있나요?*
A: 네, 많이 있습니다.
어순의 변화
Q: 커피를 마실 만한 곳이 공항 안에 있나요?
A: 어떤 종류를 원하십니까?
단순 어미의 변화
Q: 공항 안에 커피를 마실 만한 곳이 *있습니까?*
A: 네, 바로 준비해 드리겠습니다.

표 1 어순, 어미 변화에 따른 시퀀스-투-시퀀스 모델 응답 생성 결과

고 문장 변화가 심한 한국어 학습 시에는 문제가 더욱 부각될 수 있다. 따라서 다양한 상황에서 나타날 수 있는 다양한 표현에 대해 적절한 응답을 생성하기 위해 *사오정 문제*는 대화 모델 학습 시 꼭 해결해야하는 문제이다.

본 논문에서는 *사오정 문제*를 완화하고 강건한 응답을 생성할 수 있는 디노이징 응답 생성(denoising response generation) 모델을 제안한다. 디노이징 응답 생성 모델은 기존 시퀀스-투-시퀀스 모델의 학습과정에 디노이징 메커니즘을 도입한 모델이다. 단어 순서 변경과 단어 삭제와 같은 노이즈가 가해진 질의 문장을 입력하고 그럼에도 불구하고 본래의 적절한 응답을 생성하도록 질의-응답 쌍을 학습한다. 즉, 같은 의미를 갖지만 다른 문장으로 표현되는 현상을 노이즈가 포함된 질의 문장으로 시뮬레이션하고, 서로 다른 노이즈가 가미된 여러 문장에 대해 모두 적절한 응답을 해내도록 학습하는 것이다. 이를 통해 모델의 인코더를 질의문의 세세한 표현보다는 핵심적인 의미를 포착하도록 학습함으로써 디코더가 보다 의미 적절한 응답을 생성할 수 있도록 한다.

한국어를 딥러닝 모델을 통해 학습할 때 가장 먼저 마주하는 문제는 데이터 부족이다. 대화 모델 학습의 경우 상대적으로 최근에 연구가 시작된 주제로 데이터 부족 문제가 더욱 두드러질 수 있다. 디노이징 응답 생성 모델은 다양한 노이즈가 추가된 질의-응답 쌍을 생성함으로써 데이터를 증강시키는 효과가 있어 데이터 부족 문제를 완화할 수 있다. 또한 한국어는 체언과 조사의 결합, 다양한 형태의 어미 변화 등으로 인해 어절 단위 학습 시 다루어야할 어휘양이 매우 커지며, 이는 시퀀스-투-시퀀스 모델 학습을 어렵게 만드는 요인이 된다. 이 문제를 해소하기 위해 본 논문에서는 형태소 단위와 음절 단위로 문장을 취급하여 디노이징 응답 생성 모델을 학습하는 방법을 소개한다.

제안하는 방법의 우수성을 보이기 위해 약 9만 건의 한국어 질의-응답 쌍 데이터에 대해 실험을 수행했다. 기본적인 시퀀스-투-시퀀스 모델과 제안하는 디노이징 응답 생성 모델을 정량적으로 비교하기 위해 각 모델이 생성한 응답에 대해 ROUGE score를 측정하였다. 또한 각 응답의 적절성 여부를 사람이 직접 {0, 1}로 평가하여 실험 모델들에 대해 정성 평가 역시 수행하였다. 그 결과 제안하는 모델이 비교 모델에 비해 ROUGE 점수에 대해서 최대 24%, 적절한 응답 비율에 대한 정성 평가에 대해 최대 35%, 다양한 질의에 대한 응답 능력에서 최대 34%의 성능 향상을 보임을 확인할 수 있었다.

2. 관련 연구

챗봇 관련 연구는 1966년의 ELIZA [6]로 거슬러 올라간다. ELIZA는 최초의 챗봇으로 알려져 있으며 응답을 생성하기 위해 미리 몇 가지 규칙을 정의하고 그 규칙에 따라 응답을 생성하는 방식으로 설계되었다. 이런 규칙 기반 응답 모델은 결국 제약된 성능을 보일 수밖에 없다. 따라서 최근에는 대용량 대화 코퍼스로부터 질의에 대한 응답 패턴을 학습하는 방법이 주로 연구되고 있다 [2,3,4,5]. 그 중 주목할 만한 연구로 Vinyal과 Le의 연구 [2]를 언급할 수 있다. 이 연구는 기계 번역 분야에서 큰 성과를 보인 시퀀스-투-시퀀스 모델을 사용하여 대화 데이터를 효과적으로 학습할 수 있음을 보였다. 이후 시퀀스-투-시퀀스 모델을 통해 대화의 응답을 생성할 때 발생하는 여러 문제를 해결하기 위한 연구들이 진행되고 있다. [3]와 [4]에서는 모델이 *I don't know*와 같은 학습 데이터에서 빈번히 나타나지만 의미 없는 응답을 자주 생성하는 문제를 해결하기 위해 각각 상호 정보(mutual information)와 데이터 증류(data distillation)를 이용한 방법을 제시하였다. Chen et al. 또한 의미 있는 응답을 생성하기 위해 확률 토픽 모델을 시퀀스-투-시퀀스 모델과 결합시킨 모델을 소개하였다 [5]. 그러나 이러한 연구들은 상대적으로 데이터가 풍부하고 문장 변화가 덜한 영어, 중국어를 대상으로 하여 본 연구에서 해결하고자하는 *사오정 문제*를 직접적으로 다루고 있지 않다.

본 논문에서 제안하는 디노이징 응답 생성 모델은 [7]에서 비지도 학습을 통해 강건한 문장 표현을 얻기 위해 제안한 시퀀셜 디노이징 오토인코더(SDAE)로부터 영감을 얻었다. SDAE는 문장 표현 학습을 위해 임의의 노이즈가 추가된 문장을 입력하여 원래의 문장으로 복원하는 오토인코더 학습 방법이며, SDAE를 통해 얻은 문장 표현이 여러 실험에서 좋은 성능을 보임으로써 그 우수성을 입증하였다. SDAE는 입력 문장에 대해 하나의 강건한 문장 표현을 얻어 다양한 문제에 활용하는 것이 주된 목적인 반면 본 연구에서는 강건한 문장 표현을 얻는 자체보다는 좋은 응답을 생성하는 것이 궁극적인 목적이다. 따라서 SDAE의 디노이징 모델 방식을 따르되 양방향 순환신경망과 어텐션 모델(attention model)을 도입하여 맥락에 따른 문장 표현을 얻을 수 있도록 디노이징 응답 생성 모델을 설계하였다.

조휘열 외는 한국어 대화 데이터에 대해 시퀀스-투-시퀀스 모델을 적용한 연구 [8]를 발표하여 그 가능성을 보였다. 이 연구에서는 아침, 아이돌보기 상황으로 대화의 시나리오를 제약하고 해당 시나리오에 해당하는 한국어 데이터를 활용하여 대화 모델을 학습하였다. 실험을 통해 시퀀스-투-시퀀스 모델이 한국어에 대해서도 제약된 시나리오 내에서 비교적 강건한 응답을 생성할 수 있음을 보였다.

3. 디노이징 응답 생성 모델

그림 1은 본 논문에서 제안하는 디노이징 응답 생성 모델을 도식화한 것이다. 제안하는 디노이징 응답 생성 모델은 시퀀스-투-시퀀스 모델에 디노이징 메커니즘을 도입한 모델이다. 디노이징 메커니즘은 입력으로 주어지는 시퀀스에 노이즈 함수를 이용하여 노이즈를 추가한 후 원래 목표 시퀀스를 생성하도록 학습하는 방법이다. 이를 통해 같은 뜻의 다양한 표현의 질의에도 적절한 응답을 생성하도록 한다.

3.1 시퀀스-투-시퀀스 모델

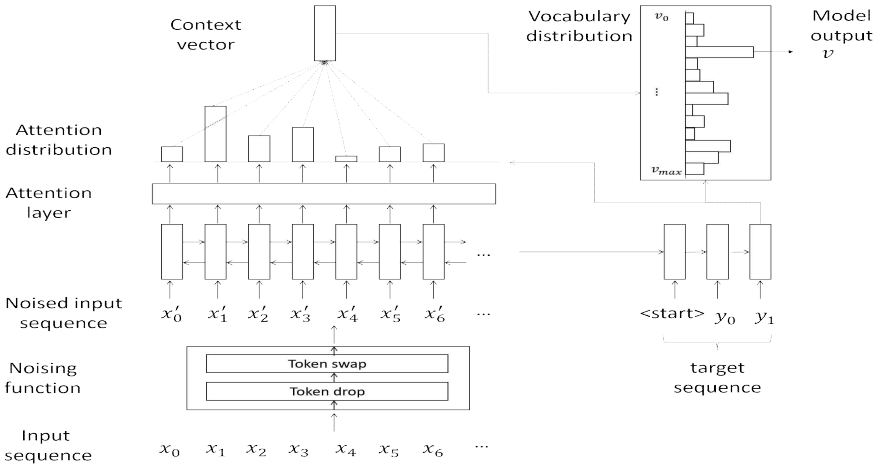


그림 1 디노이징 응답 생성 모델

```
function Noise:
  Input : input sequence S,
          token drop probability p_drop,
          token swap probability p_swap
  output : noised sequence S'

  COPY S to S'
  for i=0 to LENGTH(S):        # token drop
    if RANDOM() <= p_drop:
      REMOVE S'[i] from S'
  for i=0 to LENGTH(S)-1:      # token swap
    if RANDOM() <= p_drop:
      SWAP S'[i], S'[i+1]
  return S'
```

그림 2 노이즈 함수 의사코드

본 모델에서는 [9]에서 기계 번역을 위해 제안한 시퀀스-투-시퀀스 모델을 대화 데이터 학습을 위한 기본 모델로 차용한다. 이 모델은 인코더로 양방향(bi-directoinal) LSTM을 사용하며 디코더로는 LSTM을 사용한다. 양방향 LSTM은 시퀀스를 정방향과 역방향으로 동시 학습한 후, 양방향에서 도출된 두 개의 벡터 표현을 합쳐 입력 시퀀스에 대한 하나의 벡터 표현을 출력한다. 또한 양방향 LSTM에 더해, 어텐션(attention) 메커니즘을 적용하였다. 어텐션 메커니즘은 학습과정에서 중요하다 여겨지는 입력 시퀀스의 특정 부분을 다른 부분보다 집중적으로 반영하기 위한 방법이다. 이를 위해 중요하게 볼 시퀀스의 정보를 가진 맥락 벡터(context vector)를 생성하며, 맥락 벡터 생성을 위해 어텐션 층을 추가로 학습한다. 이렇게 생성된 맥락 벡터를 출력 시퀀스 생성과정에 반영한다.

3.2 디노이징 메커니즘

디노이징 메커니즘을 구현하기 위해서는 입력 시퀀스에 적절한 노이즈를 가할 노이즈 함수 N이 필요하다. 노이즈 함수 N은 노이즈 매개변수 p_{drop}와 p_{swap}를 기반으로 입력 시퀀스 S를 노이즈가 가미된 S'로 변환하는 함수이다.

$$S' = N(S; p_{drop}, p_{swap})$$

여기서 p_{drop}은 S 내에 각 토큰을 제거할 확률이며, p_{swap}은 S 내에 연속된 두 토큰의 위치를 교체할 확률이다. 그림 2는 노이즈 함수 N이 실제로 적용되는 알고리즘에 대한 의사 코드이다. 알고리즘에서 확인할 수 있는 것처럼 제안하는 방법에서는 토큰 제거를 먼저 적용 후 토큰 위치 교체를 적용한다.

한국어에 좀 더 적합한 디노이징 응답 생성 모델 학습을 위해 형태소를 노이즈 함수 N을 위한 토큰 단위로 사용할 것을 제안한다. 한국어 문장 데이터를 학습할 때 고려해야하는 주요 특징 중 하나는 어절의 다양함이다. 한국어는 조사, 어미 변화 등의 이유로 어절 종류가 매우 많아져 어절 단위 학습을 하기 어렵다. 형태소 단위 학습은 한국어의 어절 수 문제를 해결할 뿐만 아니라 한국어에서 빈번히 일어나는 어미, 조사 변화 등을 노이즈 함수를 통해 시뮬레이션하기에 적합하다. 즉, 노이즈 함수를 통해 조사, 어미 부분에도 확률적으로 노이즈를 부여함으로써 모델이 질의문의 본질적인 의미를 좀 더 잘 학습할 수 있게 될 것이다.

4. 실험 및 평가

4.1 데이터 셋

본 논문에서는 스마트 모바일 다국어 언어음성 데이터로부터 대화 쌍을 추출하여 실험 데이터를 구축하였다. 스마트 모바일 다국어 언어음성 데이터는 관광지, 호텔, 공항, 역, 길 등의 장소에서 두 명의 화자에 의해 이루어지는 대화를 가지고 있다. 해당 데이터 셋의 각 대화로부터 연속적인 두 개의 발화에 대해 각각을 선행 발화를 질의 문장, 후행 발화를 응답 문장으로 하여 질의-응답 쌍으로 이루어진 데이터로 만들었다. 이런 과정을 통해 총 90,729개의 대화 쌍으로 이루어진 데이터를 구축하였다. 표 2를 통해 실험에서 사용한 데이터에 대한 간략한 통계 자료를 확인할 수 있다.

표 2 데이터 구성 단위 별 데이터 통계 (단위:개)

구성 단위	대화 쌍 수	어휘 수	전체 등장한 토큰 수
어절	90,729	72,936	914,441
형태소	90,729	14,578	2,253,300
음절	90,729	1,512	3,111,117

표 3 모델 별 ROUGE F1 SCORE 측정 결과

모델	ROUGE-1	ROUGE-2	ROUGE-L
기존 모델 (어절)	0.056	0.014	0.055
기존 모델 (형태소)	0.217	0.080	0.178
기존 모델 (음절)	0.233	0.103	0.185
제안 모델 (형태소)	0.223	0.091	0.183
제안 모델 (음절)	**0.251**	**0.128**	**0.200**

4.2 실험 구성

학습 과정에서는 각 데이터의 구성 단위 어휘 수가 5만개가 넘어 갈 경우 빈도 수가 높은 상위 5만개의 어휘만을 사용하였다. 모든 모델에서 레이어는 한 층만 쌓도록 설정하였다. Hidden state 크기는 모두 1,000을 사용하였다. 모델 학습은 배치 크기를 64로 하는 미니-배치 학습 방법을 사용하였으며 Stochastic gradient descent 알고리즘을 통해 학습하였다. 기본 모델에서 어절 단위 실험의 경우 learning rate를 0.05, learning rate의 감소 비율을 0.95로 하였으며 나머지 네 모델의 경우 learning rate 0.01, 감소 비율은 설정하지 않았다. 학습 횟수는 어절 단위 데이터를 이용한 기본 모델의 경우 200회를 학습하였으며 나머지 모델은 800회를 학습하였다. 제안 모델의 노이즈 함수에서 p_{drop}과 p_{swap}은 0.1로 설정했다. 모든 실험에서 학습 데이터를 8:1:1의 비율로 학습, 검증, 평가 셋으로 나누어 진행하였으며, 검증 오류의 변화를 살펴 모델이 과다 학습(overfitting)되는 것을 막았다.

4.3 비교 모델

실험을 위해 다섯 가지 서로 다른 비교 모델을 설정하였다. 다섯 모델은 디노이징 메커니즘을 적용한 제안 모델 두 가지와 디노이징 메커니즘을 적용하지 않은 세 가지 기존 모델로 나뉜다. 기존 모델 세 가지는 각각 어절 단위, 형태소 단위, 음절 단위로 학습한 모델이며, 제안 모델 두 가지는 형태소 단위와 음절 단위로 학습한 모델이다. 음절 단위 학습은 모델이 다루어야할 어휘 수를 급격히 줄여주는 효과가 있다. 영어권에서는 음절에 해당한다고 볼 수 있는 문자 단위 학습 방법으로 의미 있는 LSTM 언어모델을 학습할 수 있음을 보였다 [10]. 또한 한국어의 경우 어근이 한 음절로 이루어진 용언과 체언이 많아 음절 단위로 시퀀스를 취급하고 노이즈를 적용하는 것이 영어의 문자 단위 학습보다 의미가 있을 수 있다.

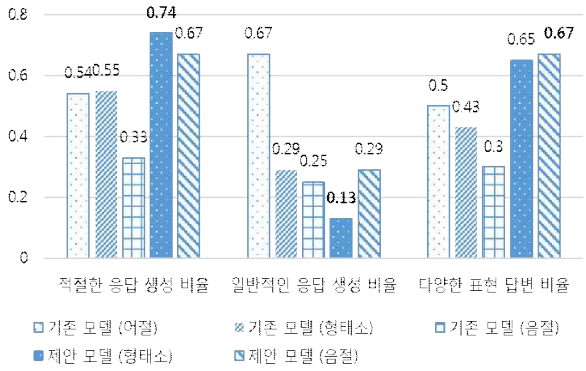


그림 3 각 모델별 정성 평가 성능 비교

4.4 정량 평가

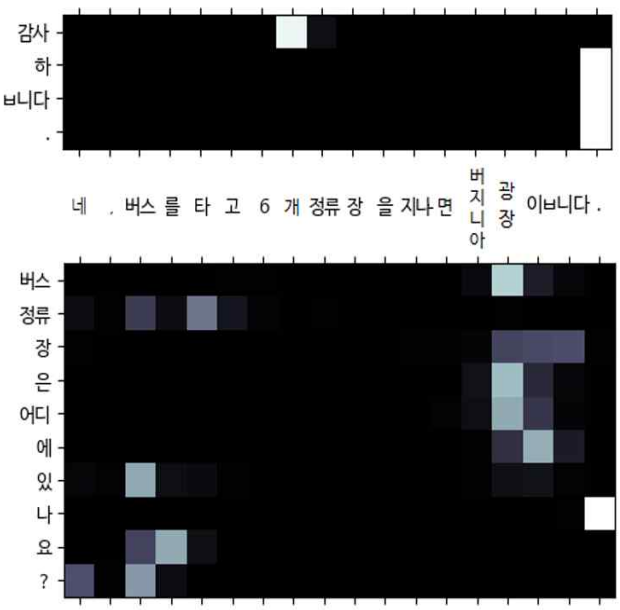
본 논문에서는 각 모델의 성능을 측정하기 위한 지표로 ROUGE F1 점수[11]를 사용했다. ROUGE F1 점수는 모델이 생성한 자연어의 품질을 정량 평가하기 위한 지표로써 모델이 생성한 문장 내의 n-gram 시퀀스들이 실제 정답 문장에 얼마나 포함되어있는지를 수치화한다.

표 3은 테스트 데이터에 대한 각 모델의 ROUGE F1 점수 측정 결과이다. 제안 모델 중 음절을 사용했을 경우가 모든 측정 방식에 대해 가장 좋은 수치를 기록하였다. 그러나 ROUGE 점수의 경우 모델이 다룰 어휘의 수가 적을수록 값이 높아질 수 있으므로 학습 토큰 단위가 같은 모델끼리 비교하는 것이 좀 더 정확하게 성능을 평가하는 방법이 될 수 있다. 이 경우에도 제안하는 방법의 형태소 모델이 기존 방법의 형태소 모델보다 최대 13.7%, 음절 모델에서 최대 24% 가량의 성능 향상을 보였다.

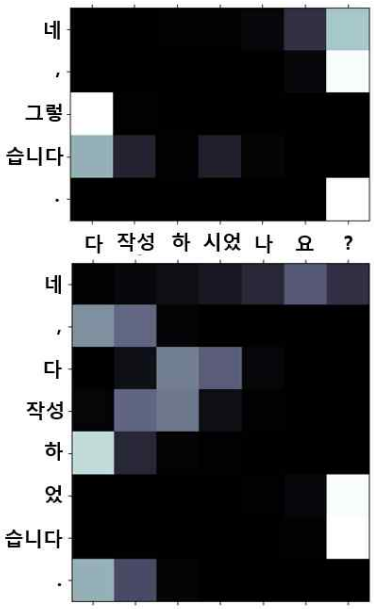
4.5 정성 평가

ROUGE 점수는 요약 등의 자연어 생성 문제에서 정량적 품질 평가 방법으로 널리 사용되고는 있지만, 대화의 경우 같은 의미를 나타내는 굉장히 다양한 자연어 표현이 가능하며 응답의 의미가 다르다 하더라도 충분히 적절한 응답일 수 있기 때문에 평가 도구로써 한계가 명확하다. 따라서 본 연구에서는 사람에 의한 정성 평가를 시행하였다. 정성 평가 방법은 모델이 생성한 응답에 대해 사람이 {적절한 응답, 부적절한 응답}을 {1, 0}으로 평가하여 전체 응답 내에서의 적절한 응답 비율을 살펴보았다. 평가는 테스트 데이터에서 임의로 929개를 추출하여 진행했다.

그림 3의 가장 왼쪽 그래프는 각 모델에서 생성한 응답에 대해 해당 응답이 적절하다고 판단된 경우에 대한 결과가 나타나 있다. 그 결과 제안 모델 중 형태소 모델이 가장 좋은 결과를 보였다. 기존 모델들 간의 결과를 살펴보면 음절 모델이 가장 좋지 않은 성능일 보였는데, 이는 문장을 음절로 쪼개어 학습한 결과 응답 생성 시 문법에 맞는 의미 있는 문장을 만들어내는 데 실패했기 때문이다. 그러나 제안하는 방법의 음절 모델의 경우 기존의 모든 모델보다 더 우수한 성능을 얻었다. 이는 제안하는 모델의 디노이징 메커니즘이 모델을 정규화(regularization)하는 힘이 있기 때문으로 보인다.



가. 질의: 네, 버스를 타고 6개 정류장을 지나면 버지니아 광장입니다.



나. 질의: 다 작성 하셨나요?

그림 4 질의에 대한 응답 생성 시 모델별 어텐션 가중치 차이 예시. 각 그림의 가로축은 질의문이며 세로축은 모델이 생성한 응답. 색이 밝을수록 어텐션 가중치가 높음. 위: 기존 모델(형태소). 아래: 제안 모델(형태소).

제안하는 디노이징 응답 생성 모델은 소위 *I don't know* 문제에도 더 강건한 성능을 보였다. 한국어 대화에 대한 응답 모델 학습 시에도 '네. 그렇습니다.' 나 '네. 알겠습니다.' 등의 지나치게 일반적인 응답을 빈번하게 생성하는 문제가 나타난다. 그림 3의 가운데 그래프는 각 모델 별 적절한 응답 중 이러한 일반적인 응답의 비율을 나타낸 것이다. 이 평가에서도 제안하는 형태소 단위의 디노이징 응답 생성 모델이 가장 적은 일반적인 응답을 생성했으며, 두 번째로 좋은 성능의 모델보다 일반적인 응답의 비율을 절반 가까이 줄였음을 확인할 수 있다. 즉, 그림 3의 왼쪽과 가운데 결과를 통해 형태소 디노이징 응답 생성 모델이 다른 모델보다 더 적절한 응답을 잘 하면서도 더 다양한 응답을 해낸다는 것을 알 수 있다.

본 논문에서 해결하고자 하는 *사오정 문제*에 대한 정성 평가 역시 수행하였다. 이를 위해 테스트 데이터 중 임의의 20개의 대화 쌍을 선택한 후 각 대화 쌍의 질의문을 여러 형태로 변환하였다. 각 질의에 대한 변환은 1. 어순 변경, 2. 특정 단어 제거, 3. 단어 변경, 4. 어미 변경, 5. 혼합의 5가지 방법을 적용하였다. 이를 통해 만들어진 총 120개의 질의에 대해 각 모델에서 생성한 응답의 적절성 평가를 실행하였다. 그림 3의 오른쪽 그래프는 각 모델 별 120개 응답 중 적절하다고 평가된 응답 비율을 나타낸 것이며, 평가한 결과 제안한 디노이징 응답 생성 모델이 기존 모델보다 두드러지게 좋은 성능을 보였다. 그 중 음절 디노이징 응답 생성 모델이 가장 좋은 성능을 보였으나 형태소 모델과 큰 차이를 보이지는 않았다.

표 4은 기존 테스트 발화 외에 앞에서 언급한 5가지 방식의 다양한 표현에 대해 형태소 모델이 생성한 응답을 보인 것으로 밑줄 친 응답은 적절치 못한 응답이다. 첫 예시의 경우 '이 바지 한번 입어봐도 될까요?' 라는 질문의 모든 변형에 대해 제안 모델은 '네, 그렇습니다.' 라는 다소 일반적인 대답을 포함하지만 적절한 응답을 생성함을 알 수 있다. 그러나 기존 모델은 마지막 변형에 대해 '네, 말씀하십시오.' 라는 엉뚱한 대답을 하였다. 두 번째 예시의 경우 제안 모델도 다소 맥락에 맞지 않는 응답을 하기도 하지만 적절한 답변도 해낸 반면, 기존 모델은 모든 질문에 대해 전혀 맥락과 다른 응답을 생성함을 확인할 수 있다.

4.6 어텐션 메커니즘을 통한 분석

제안하는 디노이징 응답 생성 모델이 모든 평가에서 좋은 결과를 보였다. 이는 제안하는 모델이 입력 데이터에 디노이징을 적용하여 학습하는 과정에서 문장의 핵심적인 의미를 학습할 수 있기 때문일 것이다. 이 과정에서 어텐션 메커니즘 또한 영향을 받는다. 그림 4는 같은 질의문에 대해 기존 모델과 제안 모델이 질의의 다른 부분에 주목을 한 결과 전혀 다른 응답을 생성하게 되는 것을 잘 보여준다. 가. 질의에 대해 기존 모델은 '개' 와 '.' 에 주목하고 '감사합니다.' 라는 매우 전형적인 응답을 생성하였다. 반면 제안 모델은 '버스', '타', '광장' 등의 질의 내용 중 중요한 단어에 주목을 함으로써 '버스정류장은 어디에 있나요?' 라는 적절하면서도 뻔하지 않은 응답을 생성할 수 있었다. 나. 질의에서도 기존 모델은 '다' 와 '?' 에만 강력히 주목한 반면 제안 모델은 '다', '작성', '하', '?' 등에 효과적으로 주목함으로써 결과적으로 '네, 그렇습니다.' 와 같은 전형적인 응답이 아닌 '네, 다 작성했습니다.' 로 보다 질문에 특화된 응답을 생성한

표 4 모델 별 같은 뜻의 다양한 표현의 발화에 대한 응답 예시
(변형1 : 어순 변경, 변형2 : 특정 단어 제거, 변형3 : 단어 변경, 변형4 : 어미 변경, 변형5 : 혼합)

예 1) Q: 이 바지 한번 입어봐도 될까요?
A: 네, 탈의실에서 입어보세요.

	기본	변형1	변형2
	Q: 이 바지 한번 입어봐도 될까요?	Q: 한번 이 바지 입어봐도 될까요?	Q: 이 바지 입어봐도 될까요?
기존 모델 (형태소)	네, 입어 보세요.	네, 입어 보세요.	네, 입어 보세요.
제안 모델 (형태소)	네, 입어 보세요.	네, 입어 보세요.	네, 입어 보세요.
	변형3	변형4	변형5
	Q: 이 바지 한번 착용 해봐도 될까요?	Q: 이 바지 한번 입어봐도 됩니까?	Q: 한번 이 바지 착용해봐도 됩니까?
기존 모델 (형태소)	아니요. 마음에 들세요.	네, 입어 보세요.	네, 말씀 하십시오.
제안 모델 (형태소)	네, 그렇습니다.	네, 입어 보세요.	네, 그렇습니다.

예2) Q: 아, 감사합니다. 버스는 몇 분마다 있나요?
A: 자주 있어요. 23분마다 한 대씩 오니까 조금만 기다리시면 될 겁니다.

기본	변형1	변형2
Q: 아, 감사합니다. 버스는 몇 분마다 있나요?	Q: 아, 감사합니다. 몇 분마다 버스가 있나요?	Q: 감사합니다. 버스는 몇 분 마다 있나요?
네 , 바로 알겠습니다.	한 번 더 타세요.	한번에 버스가 있 습니다.
그것은 10 분 마다 한 대씩 있습니다.	네, 3번에 한 번씩 타시면 됩니다.	10분 정도 소요 됩니다.
변형3	변형4	변형5
Q: 아, 감사합니다. 버스는 얼마 마다 있나요?	Q: 아, 감사해요. 버스는 몇 분 마다 있어요?	Q: 감사해요. 얼마마다 버스 는 있어요?
저쪽에 있는 정류장에서 타시면 됩니다.	저도 모르겠어요. 감사해요.	그것만 5천원이요.
3천원 입니다.	그것은 10 분 마다 한대씩 오시면 돼요.	안내원이에요.

것을 알 수 있다.

5. 결론

본 논문에서는 대화에서 나타날 수 있는 다양한 질의에 대해 강건하고 다양한 응답을 생성할 수 있는 디노이징 응답 생성 모델을 제안하였다. 제안하는 모델은 대화 응답 생성을 위해 널리 사용되는 시퀀스-투-시퀀스 모델에 디노이징 메커니즘을 추가하였다. 이를 통해 다양한 질의에 대해 적절한 응답을 생성하도록 설계했다. 실험을 통해 제안한 모델이 같은 뜻의 다양한 형태의 질의에 대해 기존 모델보다 적절한 응답을 생성할 수 있음을 확인 할 수 있었으며 부가적으로 일반적인 응답 생성 비율도 감소함을 보였다. 이를 통해 제안하는 모델의 데이터 증감 효과와 정규화 능력이 질의에 대한 강건한 의미 벡터 표현을 학습하도록 함으로써 일반적인 응답을 생성하는 비율을 줄일 수 있음을 확인하였다.

사사

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2013-0-00109, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," In *Proceedings of NIPS*, pp. 3104-3112, 2014.

[2] O. Vinyals and Q. V. Le, "A Neural Conversational Model," arXiv preprint arXiv:1506.05869, 2015.

[3] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A Diversity-Promoting Objective Function for Neural Conversation Models," In *Proceedings of NAACL-HLT*, pp. 110-119, 2016.

[4] J. Li, W. Monroe, and D. Jurafsky, "Data Distillation for Controlling Specificity in Dialogue Generation," arXiv preprint arXiv:1702.06703, 2017.

[5] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W. Y. Ma, "Topic Aware Neural Response Generation," In *Proceedings of AAAI*, pp. 3351-3357, 2017.

[6] J. Weizenbaum, "ELIZA - A computer program for the study of natural language communication between man and machine," *Communications of the ACM*, Vol. 9, No. 1, pp. 36-45, 1966.

[7] F. Hill, K. Cho, and A. Korhonen, "Learning Distributed Representations of Sentences from Unlabelled Data," In *Proceedings of NAACL-HLT*, pp. 1367-1377, 2016.

[8] 조휘열, 강우영, 한동식, 장병탁, "Konvbot: 한국어 대화 모델," 한국정보과학회 학술발표논문집, pp. 624-626, 2016.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.

[10] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," arXiv preprint arXiv:1506.02078, 2015.

[11] C. Y. Lin, "Rouge: A Package for Automatic Evaluation of Summaries," In *Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out (WAS)*, pp. 74-81, 2004.