대화에서 멀티태스크 학습을 이용한 감정 및 화행 분류

신창욱○, 차정원
창원대학교
{papower1, jcha}@changwon.ac.kr

Emotion and Speech Act classification in Dialogue using Multitask Learning

Chang-Uk Shin○, Jeong-Won Cha
Changwon National University

요 약

심층인공신경망을 이용한 대화 모델링 연구가 활발하게 진행되고 있다. 본 논문에서는 대화에서 발화의 감정과 화행을 분류하기 위해 멀티태스크(multitask) 학습을 이용한 End-to-End 시스템을 제안한다. 우리는 감정과 화행을 동시에 분류하는 시스템을 개발하기 위해 멀티태스크 학습을 수행한다. 또한 불균형 범주 분류를 위해 계단식분류(cascaded classification) 구조를 사용하였다. 일상대화 데이터셋을 사용하여 실험을 수행하였고 macro average precision으로 성능을 측정하여 감정 분류 60.43%, 화행 분류 74.29%를 각각 달성하였다. 이는 baseline 모델 대비 각각 29.00%, 1.54% 향상된 성능이다. 본 논문에서는 제안하는 구조를 이용하여, 발화의 감정 및 화행 분류가 End-to-End 방식으로 모델링 가능함을 보였다. 그리고, 두 분류 문제를 하나의 구조로 적절히 학습하기 위한 방법과 분류 문제에서의 범주 불균형 문제를 해결하기 위한 분류 방법을 제시하였다.

주제어: 발화 감정 분류, 발화 화행 분류, 계단식 분류 모형, 멀티태스크 학습

1. 서론

대화 시스템은 사용자와 대화를 수행하며, 입력된 사용자의 발화에 적절한 발화 생성을 목표로 한다. 따라서 대화 시스템은 적절한 발화를 출력하기 위해 입력된 사용자 발화의 의미를 이해할 수 있어야 한다. 그 중 일상대화 시스템은 주어진 목표 없이 사용자의 입력 발화와 현재 주제에 알맞은 적절한 발화를 출력하여야 한다. 이는 기계학습 모델의 입장에서 많은 어휘와 문맥의 흐름을 이해하도록 요구하여 결과적으로 학습을 어렵게 한다.

지도학습 기반 분류 모형을 작성하기 위해서는 고품질의 학습 데이터셋이 필요하다. 발화 감정 인식은 고품질 대용량 학습 데이터셋을 구축하는 데에 비용과 시간이 많이 소요된다. 그리고 적은 양의 데이터셋은 범주 불균형 문제(class imbalanced problem)를 가지고 있을 가능성이 높다.

본 논문에서는 넓은 도메인의 일상 대화에 대하여 사용자의 입력 발화의 감정과 화행을 동시 추론하는 End-to-End 대화 모델링 구조를 연구한다. 설정한 분류 문제를 해결하기 위해 실험을 수행하는 중 발생하였던 문제를 해결 시도하고, 그 과정과 결과를 공유한다. 그 중 첫 번째는 데이터 불균형(imbalance)으로 인해 발생한 softmax 층의 편향(bias)이다. 감정과 화행 중 더 심한 불균형을 가진 것은 감정 분류인데, 제 1 고빈도가 '감정 없음' 으로 전체 데이터셋의 총 83.10%를 차지하고 있다. 이로 인하여, 학습된 모델은 모든 평가 샘플에 대해 '감정 없음' 을 추론하는 문제가 발생하였다. 이를 해결하기 위한 과정을 이어서 기술한다. 또한, 감정과 화행 두 문제가 공유하는 계층을 더하고 빼면서 공유 계층(shared layer)의 효과를 확인하고자 하였다.

우리는 발화당 감정과 화행이 부착된 대화 데이터셋을 이용하여, 매 턴 입력된 발화의 감정과 화행을 추론하는 모델을 설계하였다. 이 때, 입력된 발화만을 사용하는 것이 아니라 주어진 대화의 기록을 모두 입력으로 하여 문맥에 기반하는 감정과 화행 추론을 수행한다.

우리의 구조에서 두 분류 문제는 같은 단어 임베딩(word embedding)을 공유하고, 일부 계층(layer)도 공유하도록 설정되어 있다. 따라서 단어 임베딩과 일부 공유 계층들은 두 문제로부터 함께 학습되어, 결국 모델의 성능 향상에 기여한다. 또한 데이터 불균형 분류 문제를 해결하기 위한 계단식 분류 모듈을 제안한다. 제안하는 구조와 주어진 데이터셋을 이용하여 모델을 학습하고, 그 성능을 비교하였다.

본 연구의 기여점을 정리해 본다면 다음과 같다.

- 대화 내 발화의 감정과 화행을 동시에 분류할 수 있는 인공신경망 구조를 제안한다.
- 강한 학습 데이터 불균형 상태의 분류 데이터셋에서 강건하게 동작하는 분류 모듈을 제안한다.
- 두 개의 분류 문제로부터 동시에 학습되는 공유 계층을 둠으로써 두 분류 문제의 성능을 향상시킬 수 있음을 확인하였다.

2. 관련 연구

인공신경망을 이용한 발화 내 감정 분류/화행 분류는 여러 연구자에 의해 시도된 바 있다.

[1]에서는 대화의 문맥을 반영한 감정 분류를 수행하

기 위해 CNN-LSTM(Convolutional Neural Network - Long Short-term Memory) 구조를 제안하였다. CNN으로 하나의 발화를 입력받아 자질을 추출하고, 그렇게 추출된 자질을 LSTM에 입력하여 문맥을 반영하고자 하였다. 대용량의 뉴스, 드라마 데이터셋에 대해 skip-gram으로 비지도 단어 임베딩 학습을 수행하여 단어 임베딩을 작성하였다. 본 논문에서 사용하는 데이터셋과 같이, 감정 없음 지표의 분포가 전체의 68.9%를 차지하는 등 강한 불균형이 존재하는 데이터셋을 사용하였다. CNN-LSTM을 이용한 감정 분류 성능은 정확도로 82.93%, macro F1 score로 77.56%를 달성하였다.

[2]에서는 화행을 분류하기 위해 지지벡터기계(Support Vector Machine; SVM)를 이용한 모델을 제안하였다. 지지벡터기계로 분류를 수행할 때, 저빈도 클래스들에 대해서 우선 분류를 수행하고, 저빈도 클래스 분류기에서 모두 분류되지 않은 경우에만 고빈도 분류기를 동작시킬 수 있도록 단계를 나누어 수행하였다. 또, 분류가 수행된 후, 분류기의 모호성 수치를 추산하고, 그 수치가 임계값 이하인 경우에는 후보정 규칙으로 보정하여 성능을 향상시켰다. 예약 도메인의 대화 데이터셋에 17개의 화행 지표가 부착된 데이터셋을 사용하였다. 본 논문과는 사용한 대화 데이터셋의 도메인과 양, 그리고 화행의 지표의 수에 차이가 있다. 실험 결과로 최종 micro F1 score 86.18%의 성능을 달성하여, 분류 우선순위 조정과 후보정 규칙을 이용하여 화행 분류 성능을 개선할 수 있음을 보였다.

3. 일상대화 데이터셋

본 논문에서 실험을 위해 사용한 데이터셋은 [3]에서 제안한 Daily Dialog 데이터셋이다. 10개의 도메인이 대화마다 부착되어 있고, 감정 없음을 포함한 7개의 감정 그리고 4개의 화행 지표가 발화마다 부착되어 있다. 이 데이터셋은 영어 대화 데이터셋이므로, 번역하여 실험에 사용하였다. 표 1에 통계량을 정리하였다.

표 1. 일상대화 데이터셋의 통계량

구분	수량	단위
전체 대화의 수	13,112	대화
전체 발화의 수	102,980	발화
대화당 평균 발화의 수	7.85	발화/대화
발화당 평균 어절의 수	8.03	어절/발화
발화당 평균 음절의 수	24.48	음절/발화

이 데이터셋은 13,000여 대화의 작은 데이터셋으로, 이 데이터셋을 이용해 넓은 도메인의 일상대화를 모델링하기에 어려움이 있을 것으로 예상하였다. 따라서 [4]에서 제안된 바 있는 denoising을 이용하여 대화 데이터셋의 양을 증축하였다. 또한, 해결하고자 하는 문제가 마지막 발화의 감정과 화행 분류이므로, 일정 턴 이상의 대화는 마지막 턴을 잘라가며 증축할 수 있다. 이에, 3턴 이상의 대화는 3턴이 될 때까지 마지막 턴을 제거하면서 대화를 증축하였다. 그림 1은 우리의 증축 과정의 의사코드이다. 이 증축 과정을 수행한 이후의 통계량을 표 2에 정리하였다.

```
function Augment:
  Input : input dialogue D
  Output : dialogue list L

  APPEND D to L
  if TURN_LENGTH(D) <= 3: return L
  else :
    for len=3 to TURN_LENGTH(D)-1:
      COPY D to D'
      SLICE D' from 0 to len
      APPEND D' to L
  return L
```

그림 1. 대화 증축 알고리즘의 의사코드

표 2. 증축 후 일상대화 데이터셋의 통계량

구분	수량	단위
전체 대화의 수	86,664	대화
전체 발화의 수	570,171	발화
대화당 평균 발화의 수	6.58	발화/대화
발화당 평균 어절의 수	7.89	어절/발화
발화당 평균 음절의 수	24.07	음절/발화

4. 제안 방법

본 논문에서 제안하는 인공신경망 구조는 총 4개의 모듈로 구성되어 있다. 이는 history module, last utterance module, short-term memory module, 그리고 classification module이다[그림 2].

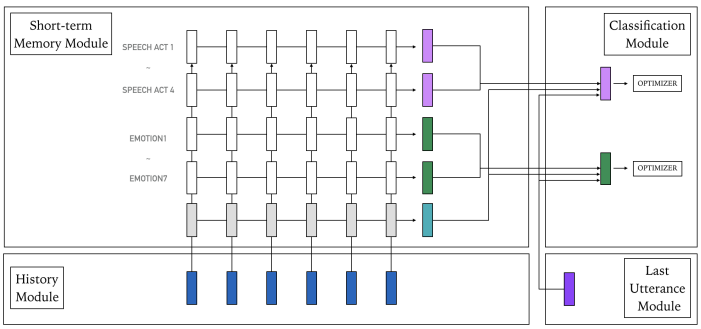

그림 2. 제안 구조

History module은 현재까지 진행된 발화 기록에서 마지막 발화를 제외한 나머지 발화가 입력되는 모듈이다. 입력된 발화들을 각각 음절 단위 CNN encoder에 입력하여 발화 분산 표현을 작성한다. Last utterance module은 마지막 발화를 입력으로 받아 마지막 발화의 분산 표현을 작성한다. 본 구조에서 감정과 화행을 분류하고자 하는 발화가 이 마지막 발화이므로, history module에 함께 입력하지 않고, 따로 떼어서 encoding을 수행한다.

발화 분산 표현 작성에 RNN(Recurrent Neural Network)과 attention을 이용할 수 있다. RNN은 비교적 길이가 긴 발화의 앞 쪽 의미 정보를 유실할 수 있다. 또한 경사도 소실(vanishing gradient) 혹은 경사도 폭발(exploding gradient) 문제를 겪을 수 있다. 그러나, CNN의 경우, 길이와 무관하게 정보를 추출할 수 있으며, 경사도 소실/폭발 문제로부터 자유롭다.

위에서 작성된 분산표현들은 short-term memory module에 입력된다. Short-term memory module은 입력된 분산표현들을 [5]의 slot recurrent cell로 인코딩하여 n개의 slot vector를 작성한다. Slot recurrent cell의 entity로 7개의 감정 지표와 4개의 화행 지표를 부여하였다. 모든 발화 기록에 대해 RNN의 연산이 수행 완료되면, 7개의 감정 지표가 부착된 RNN과 4개의 화행 지표가 부착된 RNN의 연산 결과를 각각 fully-connected layer에 입력하여 두 개의 분산 표현을 작성한다. 어떤 entity도 부여되어 있지 않은 RNN을 하나 추가로 두어, shared layer로 동작할 수 있도록 하였다. 따라서 총 3개의 분산 표현이 각각의 classification module로 보내어진다.

그리고 마지막 단계인 classification module은 입력받은 3개의 분산 표현을 연결(concatenation)하고, 그것을 자질로 하는 softmax classifier를 동작시켜 결과를 도출한다.

그러나 최초 설정한 일반적인 softmax classifier는 심한 데이터 불균형으로 학습이 되지 않았다. 이에 cascaded classification module을 제안한다. 제안하는 cascaded classification module은 데이터의 분포 불균형을 가능한 완화하기 위하여 다음과 같이 설계하였다.

첫 번째 분류기는 최다빈도 클래스와 나머지를 분류한다. 그리고 첫 번째 분류기에서 '나머지'로 분류된 경우 두 번째 분류기에 입력하여 두 번째 클래스와 나머지를 분류한다. 다시 두 번째 분류기에서 '나머지'로 분류된 경우만 세 번째 분류기에 입력하여 나머지 클래스의 분류를 수행한다. 감정의 경우 하위 5개 클래스의 비율이 낮아 세 번째 분류기에서 5개 클래스를 분류할 수 있도록 하였다. 화행은 4개의 클래스를 가지고 있으므로, 세 번째 분류기에서 하위 2개 클래스를 분류할 수 있도록 하였다.

5. 실험

5.1. 실험 설정

실험에는 3장에 기술한 Daily Dialog 번역 데이터셋을 사용하였다. 3장에 기술한 증축 기법으로 86,664 대화를 획득할 수 있었으나, 분류 지표의 분포가 강하게 불균형 상태이기에 그 중 마지막 발화가 '감정 없음'인 대화를 임의 추출하여 일부 제거하고 학습에 사용하였다. 실험에는 데이터셋 중 'relationship' 도메인의 대화만을 사용하였다. [표 3, 4]에 실험에 사용한 데이터셋의 통계를 기록하였다.

실험의 baseline으로, [6]에서 제안된 DMN(Dynamic Memory Network)을 문제에 맞도록 수정하여 사용하였다. DMN의 input module에 대화의 기록을, question module에 사용자의 마지막 발화를 입력하였다. input module은 각 발화의 표현을 얻기 위해 음절 단위 position encoding[7]을 사용하였고, 각 발화 간의 상관관계를 연산하기 위해 bidirectional GRU를 사용하였다. question module은 입력된 사용자 발화를 음절 단위 unidirectional GRU에 입력하여 발화 분산 표현을 작성하였다. hop은 3으로 설정하였고, 마지막으로 answer module에서는 memory module의 결과와 question module의 결과를 연결한 후, softmax classifier로 분류를 수행하였다.

표 3. 실험 데이터셋의 화행 지표 분포

화행	발화의 수	비율(%)
inform	9918	53.63
question	5177	28.00
directive	2163	11.70
commissive	1233	6.67
계	18,491	100.00

표 4. 실험 데이터셋의 감정 지표 분포

마지막 발화의 감정	수량	비율(%)
감정 없음	10,301	55.71
행복함	6,876	37.18
놀람	761	4.12
슬픔	210	1.14
화남	195	1.05
역겨움	118	0.64
두려움	30	0.16
계	18,491	100.00

모든 실험에서 hyper-parameter는 다음과 같이 설정하였다. minibatch의 크기는 100, 음절 임베딩의 크기는 100, 모든 RNN의 hidden size는 500, validation-based early-stopping의 patience는 20, learning rate는 0.00001로 설정하였다. gradient clipping을 10으로, noisy gradient[8]를 0.001로 설정하였다. 그리고 optimizer로 Adam을 사용하였다. 멀티태스크 학습을 위하여 두 task를 위한 optimizer 두 개를 설정하였다. 매 step에 하나의 optimizer만을 50%의 확률로 동작시켜 alternate training을 수행하였다.

5.2. 실험 결과 및 토의

[표 5]는 baseline으로 설정한 DMN과 제안 구조의 실험 평가 성능을 정리한 것이다. 분류 데이터셋의 불균형이 강하게 나타나기 때문에, macro average precision과 micro average precision을 함께 실었다[9].

실험 1은 baseline과의 비교를 위해 제안하는 계단식 분류 모듈과 공유 계층을 모두 적용하지 않은 것이다. 음절 CNN을 이용한 발화 분산 표현 획득이 baseline의 음절 단위 position encoding 이후 bidirectional GRU를 이용한 문맥 반영보다 더욱 좋은 성능을 내었다. 특히,

표 5. Baseline과 제안 구조의 실험 결과. 실험 3과 4는 5회 수행하여 최고와 최저 성능을 제외한 평균 성능을 실었다. PE: position encoding, bi-GRU: bidirectional GRU.

구분	입력 발화 인코딩 방법	계단식 분류 모듈	공유 계층	Micro Average Precision		Macro Average Precision	
				감정 분류	화행 분류	감정 분류	화행 분류
Baseline	음절 PE + bi-GRU	X	X	71.28	84.43	31.43	72.75
제안 - 실험 1	음절 CNN	X	X	**82.79**	87.58	54.44	77.48
제안 - 실험 2		X	O	81.81	**88.25**	51.88	**81.10**
제안 - 실험 3		O	X	73.94	85.53	54.96	73.31
제안 - 실험 4		O	O	76.61	85.73	**60.43**	74.29

감정 분류의 macro average precision에서 약 23% 가량의 성능 향상은 제안 구조가 imbalanced dataset에서도 강건히 동작하였음을 시사한다.

실험 1과 2는 제안하는 계단식 분류 모듈을 적용하지 않은 실험이다. 그리고, 실험 3과 4는 제안하는 계단식 분류 모듈을 적용한 것이다. 실험 1과 2의 비교, 그리고 실험 3과 4의 비교로 공유 계층의 효과를 검증할 수 있다. 실험 1와 2의 성능 차이는 감정 분류에서 약 2.5% 하락, 화행 분류에서 약 3.6% 상승으로 두 문제의 성능 상승과 하락이 대조되는 것을 알 수 있다. 그러나, 실험 3과 4의 차이는 감정 분류에서 약 5.5% 상승, 화행 분류에서 약 0.98% 상승으로 일관된 성능 향상을 확인하였다. 이 두 쌍의 차이는 계단식 분류 모듈의 유무이기에 이로부터 제안하는 계단식 분류 모듈이 공유계층과 함께 적용되어 시너지 효과를 내었다고 분석할 수 있다.

실험에 사용한 데이터셋에서, 감정과 화행 중 데이터 불균형이 더욱 강한 것은 감정 분류이다. 그러한 감정 분류에서 최고 성능을 달성한 모델은 제안 구조와 제안 계단식 분류 모듈, 그리고 공유 계층을 모두 적용한 실험 4로, 그 성능은 baseline 대비 29.00%의 성능 향상이 있었다.

6. 결론

대화를 모델링하고자 하는 시도가 계속해서 수행되고 있다. 그 중에서도 사용자의 입력과 사전 지식, 프로파일(profile) 정보 등을 이용하여 발화 생성을 위한 자질을 작성하고, 그것을 바탕으로 다음 시스템 발화를 추론하는 방식의 연구가 주로 진행되고 있다.

우리는 일상 대화 내 발화의 감정과 화행을 추론하는 인공신경망 구조를 제안하였다. 발화의 감정과 화행은 시스템 발화 생성을 위한 자질로써 그 역할을 수행할 수 있다. 우리의 구조는 입력된 발화 외에도 진행된 대화 기록에 근거하여 마지막 사용자 발화의 감정과 화행을 추론한다. 제안하는 구조를 이용한 실험을 통해 감정 분류 기준 baseline의 성능보다 macro average precision 기준 29.00% 높은 성능을 달성하였다.

실험 결과로부터, 제안하는 구조가 발화의 감정과 화행을 분류하는 데에 도움이 됨을 보였다. 또한, 둘 이상의 문제로부터 학습되는 공유 계층을 두고, 그 공유 계층이 분류 문제 성능 향상에 어떠한 영향을 미치는지 확인하였다.

본 연구에서는 감정과 화행, 도메인 주석이 부착된 대화 코퍼스를 이용하여 발화 내 감정과 화행을 분류하는 모델을 작성하였다. 실제 감정과 화행, 도메인이 부착된 코퍼스 작성은 상당한 시간과 노력이 필요한 작업으로, 대용량 데이터셋을 구축하는 데에 제한이 있다. 이에, 더욱 적은 데이터셋을 이용하거나 혹은 반지도 학습, 혹은 비지도 학습 방식을 결합하여 발화 내 자질을 추출할 수 있는 방법에 대해 연구가 필요하다.

Acknowledgement

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1D1A1B03033534).

참고문헌

[1] 신동원, 이연수, 장정선, 임해창, “CNN-LSTM을 이용한 대화 문맥 반영과 감정 분류”, 제 28회 한글 및 한국어 학술대회 논문집, 2016.

[2] 송남훈, 배경만, 고영중, “분류 우선순위 적용과 후보정 규칙을 이용한 효과적인 한국어 화행 분류”, 정보과학회 논문지, 제43권 제1호, pp. 80-86, 2016.

[3] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, Shuzi Niu, “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”, arXiv:1710.03957v1, 2017.

[4] 김태형, 노윤석, 박성배, 박세영, “한국어 대화 모델 학습을 위한 디노이징 응답 생성”, 제 29회 한글 및 한국어 정보처리 학술대회 논문집, 2017.

[5] Chang-Uk Shin, Jeong-Won Cha, “End-to-end task dependent recurrent entity network for goal-oriented dialog learning”, Computer Speech & Language, doi:10.1016/j.csl.2018.06.004

[6] A. Kumar, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, “Ask Me Anything: Dynamic Memory Networks for Natural Language Processing”, arXiv:1506.07285v5, 2016.

[7] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus, “End-To-End Memory

Networks”, arXiv:1503.08895v5, 2015.
[8] A. Neelakantan, L. Vilnis, Q. Le, I. Sutskever, L. Kaiser, K. Kurach, J. Martens, “Adding Gradient Noise Improves Learning for Very Deep Networks”, arXiv:1511.06807v1, 2015.
[9] Marina Sokolova, Guy Lapalme, “A systematic analysis of performance measures for classification tasks”, Information Processing & Management, doi:10.1016/j.ipm.2009.03.002