FASCODE-EVAL을 위한 복합모달 접근방법

정의석O, 김현우, 박민호, 송화전

한국전자통신연구원
{eschung, kimhw, roger618, songhj}@etri.re.kr

Multi-modal approach for FASCODE-EVAL

Euisok ChungO, Hyun Woo Kim, Minho Park, Hwa Jeon Song
ETRI

요 약

FASCODE-EVAL1은 고객과 시스템간의 의상 추천 대화 문맥과 해당 문맥의 요구사항을 고려한 의상셋 추천 목록으로 구성된다. 의상셋 추천 목록은 3개의 의상셋 후보로 구성되고, 문맥과 관련성이 높은 순서로 정렬된다. 해당 정렬을 찾는 방식으로 의상 추천 시스템 평가를 진행한다. 대화 문맥는 텍스트로 되어 있고, 의상 아이템은 텍스트로 구성된 자질 정보와 의상 이미지 정보로 구성된다. 본 논문은 FASCODE-EVAL 문제를 해결하기 위하여 트랜스포머 기반의 사전학습 언어모델을 이용하고, 텍스트 정보와 이미지 정보를 해당 언어모델에 통합하는 방법을 보여준다. FASCODE-EVAL 실험결과는 기존 공개된 결과들보다 우수한 성능을 보여준다.

주제어: FASCODE, FASCODE-EVAL, 복합모달

1. 서론

FASCODE (FAShion COordination DatasEt / FAShion CODE)는 의상 코디 추천 대화셋, 의상 이미지, 의상 아이템 기술 정보로 구성된 FASCODE-DATA와 의상 추천 문맥과 추천 의상셋과의 관련성 추정을 자동 테스트할 수 있는 두가지 평가셋 FACODE-EVAL1, 2로 구성되어 있다.[1] 기존 연구 [1]은 FASCODE에 대한 자세한 설명과 FASCODE-EVAL1에 대한 베이스라인을 제시하고 있다. 본 연구는 해당 베이스라인의 성능 개선을 시도한다. 기존 연구 [2]는 FASCODE-EVAL1을 pre-trained 모델[3]을 이용하여 접근 하였다. 본 논문은 해당 접근방법을 기반으로 모델 구성을 개선하고, 이미지 자질을 통합하는 복합모달 접근 방법을 제시한다. 성능결과에서 기존 연구들보다 향상된 결과를 보여 본 연구의 타당성을 검증한다. 추가적으로 FASCODE-EVAL2의 평가를 진행하고, 해당 평가셋을 의미에 대하여 논의해 본다.

2. FASCODE

FASCODE-DATA의 대화셋은 평균 8-9턴의 7,236 대화셋으로 구성되고, 의상 아이템 개수는 2,603개이다. FASCODE-EVAL1은 200개의 개발셋과 700여개의 테스트셋으로 구성된다. 본 논문에서는 개발셋만 사용한다. 그리고, FASCODE-EVAL2는 450개의 문제들로 구성된다.

본 논문은 기존 논문 [1]의 후속 연구 결과로 기존 논문의 예제들을 편의상 그대로 활용한다. 표1은 시스템 발화 "<CO>"와 사용자 발화 "US" 그리고 의상 추천 "AC"로 구성된 축약된 대화셋 예제이다. 개발 시스템은 "AC" 생성을 목표로 할 수 있으며, 추가적으로 "CO" 생성을 목표로 할 수 있다. "발화 유형"은 대화셋 발화의 범주에 대하여 기술하며, 응용에 따라 필요 데이터를 추출하거나 데이터셋의 형태를 변경할 때 사용된다. [1]에 전체 유형이 기술되어 있다.

화자	발화	발화 유형
<CO>	안녕하세요 코디봇입니다. 무엇을 도와드릴까요?	INTRO
<US>	오빠가 결혼해요.	
<US>	차분하면서 단정한 코디 보여주세요.	
<CO>	네. 잠시만 기다려주세요.	WAIT
<CO>	차분하고 고급스러운 블라우스와 바지 코디입니다.	EXP_RES_DESCRIPTION
<AC>	JP-076 BL-027 PT-027 SE-004	
<US>	겉옷은 자켓으로 바꿔 주시구요.	USER_FAIL
:	:	:
<CO>	마음에 드신다니 정말 다행입니다.	SUCCESS
<CO>	이용해주셔서 감사합니다.	CLOSING

표1. FASCODE-DATA 대화셋 예제

의상은 의상 분류에 따라 ID를 부여한다. 의상 분류는 크게 OTBS, 겉옷(Outer), 상의(Top), 하의(Bottom), 신발(Shoes)로 구성되고, 각 분류는 하위 유형으로 구성된다. 겉옷(outer)의 경우 코트는 'CT', 가디건은 'CD', 자켓은 'JK'의 ID를 부여한다. ID를 부여 받은 하나의 의상 아이템은 의상 이미지와 형태, 재질, 색상, 감성으로 구성된 기술 텍스트의 쌍으로 표현된다.

표2는 FASCODE-EVAL1과 FASCODE-EVAL2를 설명하기 위한 예제이다. 대화문맥과 의상셋들로 크게 구성되어 있

[1] https://fashion-how.org/ETRI/board.html

고, 대화문맥과 3개의 관련 의상셋(A, B, C)를 한 묶음으로 하고, 3개의 의상셋의 관련성 순위 정보(A=Rank1, B=Rank2, C=Rank3)를 포함하는 것이 FASCODE-EVAL1의 한 평가 예제의 구성이다. 한편 대화문맥과 각각의 의상셋들의 관련성(R)을 정도에 따라 0에서 10점으로 부여한 것이 FASCODE-EVAL2의 한 평가 예제인다. 본 연구는 FASCODE-EVAL1의 평가 척도는 3개의 랭킹 순서만을 대상으로 하므로 가중치 켄달 타우(Weighted Kendall's Tau)를 사용하고, FASCODE-EVAL2의 평가 척도는 이질적 대화문맥과 관련 의상셋 후보들을 관련성의 순위를 평가하는 스피어만 상관계수 (Spearmann's Correlation)을 사용한다. 정리하면, 표5를 평가셋을 구성하는 한 예제의 관점에서 보면, FASCODE-EVAL1의 경우 "문맥-의상셋 3세트와 순위 정보"로 하나의 테스트 아이템인 반면, FASCODE-EVAL2의 관점에서는 "문맥-의상셋-점수"로 세개의 테스트 아이템이 된다.

대화 문맥	
<CO>	안녕하세요. 코디봇입니다. 무엇을 도와드릴까요?
<US>	락 페스티벌에 가는데 독특하고 특이한 룩으로 추천해주세요.
<CO>	치마, 바지, 원피스 중 어떤 옷이 포함된 코디를 추천해드릴까요?
<US>	치마로 추천해주세요.
<CO>	짧은 기장의 치마를 추천해드릴까요?
<US>	네.
의상셋 후보	
(A) R=8.3	 KN-058.jpg SE-060.jpg SK-054.jpg
(B) R=7.1	 SE-060.jpg SH-043.jpg SK-054.jpg
(C) R=4.3	 KN-160.jpg SE-060.jpg SK-054.jpg

표2. FASCODE-EVAL 예제

3. 관련 연구

기존 연구 [1]에서 FASCODE-EVAL1을 고려한 베이스라인을 제시하였다. 개발셋에서 0.529 WKT로 의미있는 결과를 제시하였다. 접근 방법은 서브워드 임베딩을 이용하여 대화문맥을 인코딩하고, 평가셋에서 제시하는 3개의 의상셋을 각각 인코딩하여 다계층의 FCN을 통한 후 통합 후 6개(3!)의 랭킹 순서에 대한 분류기를 학습하는 모델이다. 문제점은 해당 접근 방법으로는 FACODE-EVAL2에 바로 적용하기 어렵다는 점인데, 그 이유는 450!을 분류해야하기 때문이다.

기존 연구 [2]는 대용량 pre-trained model[3]을 이용한 접근 방법을 이용한다. 여기서는 문맥과 의상 아이템 하나의 쌍에 대한 긍정, 부정 이진 분류 형태로 입력 데이터를 재구성 한다. FACODE-EVAL1의 평가를 위해 입력 데이터의 긍정 레이블에 대한 모델의 출력 logit값을 이용하고, 의상셋을 구성하는 의상 아이템들 각각의 logit 값들을 합하여 "문맥-의상셋" 스코어를 계산한다. 해당 스코어값을 평가셋이 제시하는 의상셋 후보들 3개에 각각 적용하여 랭킹을 정한다. 텍스트 정보만을 이용한 실험결과 0.64 WKT의 좋은 성능을 보였다. 문제점은 이미지 자질을 해당 접근 방법에 통합하였을 때 성능이 저하되는 문제점을 보였다는데 있다. 또한, 공개된 접근 방법을 다시 구현하였을 때 해당 성능을 재현하기 어렵다는데 있다.

본 논문은 기존 연구 [2]의 접근 방법과 유사한 접근 방법을 시도하였다. 차이점은 의상셋을 구성하는 의상 아이템 쌍을 모델 입력으로 추가하여 기존 연구[2]의 최고 성능인 앙상블 모델을 넘어서는 결과를 보였다. 또한 이미지 임베딩 결과를 이용한 late fusion 접근방법을 통해 텍스트 데이터만의 성능보다 좋은 결과를 확인하였다. 이러한 평가 데이터 항목의 부분 집합에 대한 스코어링을 접근방법을 통해 FASCODE-EVAL2 평가셋의 타당성을 확인할 수 있었다.

4. 접근 방법

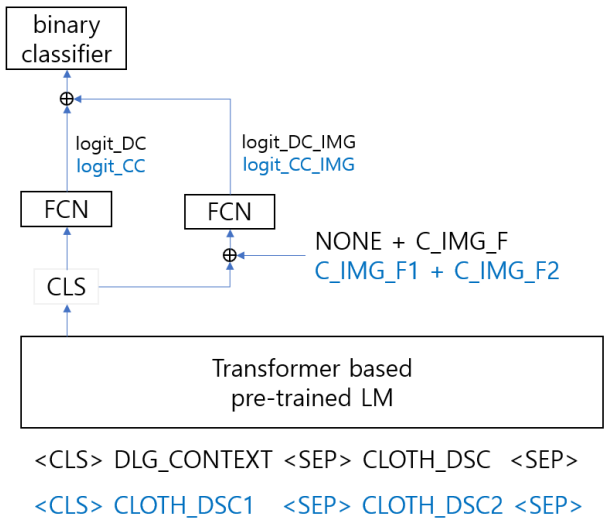

그림1. FASCODE-TEST 네트웍 구조

FASCODE-EVAL을 평가하기 위한 학습 네트웍을 FASCODE-TEST라 하겠다. 그림1은 간략하게 해당 학습 모델을 기술한다. 트랜스포머 기반 사전학습 언어모델을 이용하기 때문에 기존 연구 [2]와 유사성이 있다. 큰 차이점은 단위 학습 데이터의 구조로 볼 수 있다. [2]는

“컨텍스트 + 의상아이템 + 레이블”을 하나의 학습 단위로 한다. 여기서 의상아이템은 의상셋을 구성하는 하나의 의상이다. 의상셋과 컨텍스트의 관계성은 구성 의상아이템들의 독립적인 관계성 평가들의 합으로 표현된다. 그러나 본 연구는 “컨텍스트 + 의상아이템1 + 의상아이템2 + 레이블” 구조로 변경하였다. “의상아이템1 + 의상아이템2”는 의상셋을 구성하는 OTBS의 순서를 고려한 2개를 선택하는 조합을 의미한다. 이를 통해 본 연구는 컨텍스트와 의상과의 관계정보 뿐 아니라, 의상셋을 구성하는 의상들간의 관계정보를 활용할 수 있게 되었다. 이 지점이 본 연구의 성능개선에 가장 큰 영향을 미친 점이다. 이는 이진 분류 레이블을 고려한 데이터 증강에 있어도 큰 역할을 하였다.

입력 데이터의 레이블은 입력의 Positive, Negative를 기술한다. 즉, 대화문맥과 의상쌍이 어울리면 Positive 아니면 Negative가 한예가 될 수 있다. 이러한 학습 데이터 생성을 위해 FASCODE-DATA의 재구성이 필요하다. 발화유형에서 의상 추천 SUCCESS와 FAIL정보가 기술되어 있어 이를 활용한다. FAIL에 대한 데이터 증강은 기존연구 [2]와 같이 동일 의상 하위 유형에 한하여 생성한다.

그림1에서 “컨텍스트 + 의상아이템1 + 의상아이템2 + 레이블”이 두개의 스트링 시퀀스로 변환되는 것을 확인할 수 있다. DLG_CONTEXT는 컨텍스트 스트링을 의미하고, CLOTH_DSC는 의상아이템 기술정보들을 연결한 스트링이다. 각 스트링 시퀀스에 부합되게 이미지 자질벡터 C_IMG_F가 입력된다. 두개의 스트링 시퀀스에 대한 pretrained model의 CLS 값을 FCN(fully connected network)를 통해 logit_DC나 logit_CC를 생성한다. CLS와 통합되는 이미지 자질 정보는 FCN을 통해 logit_DC_IMG나 logit_CC_IMG를 생성한다. 서로 연결되는 값들은 색상으로 분리하여 기술했다. 두개의 logit들은 가중치를 통해 하나의 logit으로 생성되고, binary classifier에 전달된다. (1)은 입력데이터 i의 통합 logit생성을 기술한다. 여기서 β는 이미지 자질의 가중치가 된다. 학습은 (1)의 logit을 이용한 크로스 엔트로피 손실값을 계산하여 최소화하는 방식으로 진행한다.

$$logit_i = \begin{cases} logit_i^{DC} + \beta \times logit_i^{DC_{IMG}} & or \\ logit_i^{CC} + \beta \times logit_i^{CC_{IMG}} \end{cases} \quad (1)$$

위의 (1)은 학습에 사용되는 방식이고, 일부 수정을 통해 FASCODE-EVAL에 활용한다. 학습의 경우와 달리, 평가셋은 “컨텍스트 + 의상셋”을 단위로 진행되기 때문에 의상셋은 OTBS의 부분집합으로 유형을 갖고, 각 유형에 따라 의상 아이템을 갖는다. 이를 s(i)라 하자. 그리고, 두개의 의상쌍의 조합들의 집합을 sp(i)라 하자. 그러면 “컨텍스트 + 의상셋”으로 구성된 입력 i에 대한 score값은 다음과 같다. (2)에서 의상쌍에 대한 입력 가중치 α 를 적용하여 하이퍼 패러미터로 사용하였다. 이미지는 단일 이미지와 이미지 쌍에 대한 logit값은 동일하게 적용하고 전체 score에 이미지 가중치 β를 학습과 동일하게 사용하였다. 여기서 (2)에서 구해지는 score값을 이용하여 FASCODE-EVAL1,2에 사용한다. 즉, 컨텍스트와 의상셋 후보 3개가 제시되었을 때 컨텍스트와 의상셋 3쌍을 생성하고, 해당 쌍 각각에 대하여 score를 연산하여 정렬하는 접근 방법으로 평가를 진행한다.

$$score_i = \sum_{j \in s(i)} logit_{i,j}^{DC} + \alpha \times \sum_{j \in sp(i)} logit_{i,j}^{CC} + \beta \times logit_i^{IMG} \quad (2)$$

$$logit_i^{IMG} = \sum_{j \in s(i)} logit_{i,j}^{DC_{IMG}} + \sum_{j \in sp(i)} logit_{i,j}^{CC_{IMG}} \quad (3)$$

FASCODE-TEST에서 이미지 자질은 의상ID 하나에 대한 벡터값으로 표현된다. 해당 벡터값을 사전학습을 통해 의상 이미지에 대한 임베딩 벡터값의 역할을 한다. FASCODE-TEST 학습 과정에서는 해당 벡터값을 고정해 놓고 사용한다. 해당 이미지 벡터의 사전 학습 과정은 본 논문의 범위에 포함하지 않는다. 하나의 의상 ID는 Daily, Gender, Embellishment의 세가지 유형의 감성특징 벡터로 구성하여 사용하였다.

5. 실험 결과

1. FASCODE-EVAL1 실험

실험은 우선 FASCODE-DATA를 FASCODE-TEST 학습 데이터 유형으로 변경한다. 학습 데이터 유형은 “컨텍스트 + 의상아이템1 + 의상아이템2 + 레이블”이다. 발화 유형 SUCCESS만을 대상으로 추출하고, 데이터 확장 과정을 통해 FAIL 데이터를 생성하였다. FACODE-TEST의 input 구조는 그림1에 기술되어 있다.

학습은 트랜스포머 기반 사전학습 언어모델[4]를 이용하였다. 배치 크기 32, 시퀀스 길이 512를 사용했다. 실험은 공개된 기존연구[2]를 그대로 구현한 “FASCODE-TEST-BASE”와 의상쌍을 추가한 접근 방법 “+의상셋관계모델”, 이미지 자질을 추가한 “+이미지자질모델”을 텍스트상에 공개된 “기존연구1”[1], “기존연구2” [2]와 비교한다. 평가는 FASCODE-EVAL1에 대하여 WKT로 진행하였다. 표3은 해당 실험 결과를 기술한다. 실험결과 “기존연구2”는 재현이 쉽지 않았다. 그러나 의상셋 관관계모델 추가하였을 때 기존 공개 성능을 넘어설 수 있었다. 또한 이미지 정보를 추가한 복합모달을 적용하였을때 기존연구[2]가 0.578로 성능저하를 보인 반면 본 연구는 이미지 자질모델을 late fusion 방식으로 접근하였을 때 WKT 0.703으로 최고의 성능을 얻을 수 있었다.

실험	결과(WKT)	비고
기존연구1	0.529	
기존연구2	0.656	Ensemble
FASCODE-TEST-BASE	0.42	
+의상셋관계모델	0.681	
+이미지자질모델	0.703	

표3. FASCODE-EVAL1에 대한 실험결과

2. FASCODE-EVAL2 실험

FASCODE-EVAL1의 경우 단일 컨텍스트와 관계된 3개의 의상셋에 대한 관계성 랭킹에 대한 평가인 반면, FASCODE-EVAL2는 서로 이질적인 컨텍스트와 의상셋에 대한 confidence score에 대한 평가이다. 즉 score 자체가 local의 성격이 아닌 global한 성격을 지닌다. 그림2는 x축 학습 단계에 따라 y축 스피어만 상관계수 값을 보여준다. 여기서 실험들은 수식(1)의 β 값을 변화한 실험결과이다. 전반적인 실험 결과들은 학습 단계가 진행됨에 따라 성능이 개선되는 것을 보여준다. 여기서 β 값이 0.0인 것은 이미지 정보를 배제한 학습과 평가 결과를 의미한다. 실험결과 0.25값을 적용한 것의 성능이 제일 좋게 나왔고, 0.0의 경우, 성능개선에 학습시간이 많이 요구되었다. 정리하면, FASCODE-EVAL2의 실험 결과는 FASCODE-TEST가 의상 추천 시스템에서 confidence score를 계산하는 독립 모듈 역할이 가능함을 시사하므로 다양한 활용 가능성을 보여줬다고 할 수 있다.

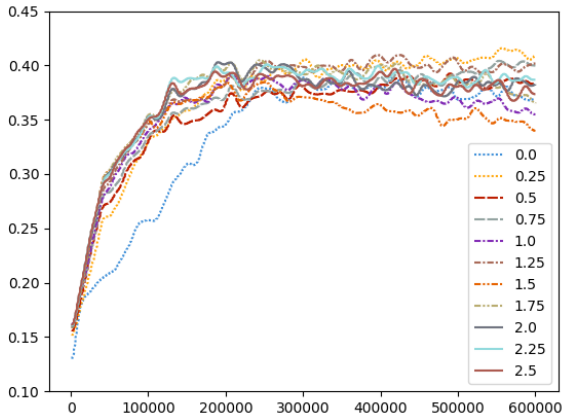

그림2. FASCODE-EVAL2에 대한 실험결과

6. 결론

본 논문은 인터랙션 기반 의상 추천 영역 데이터인 FASCODE를 이용한 복합 모달 학습과 평가 진행에 대하여 서술하였다. 대화 텍스트와 의상 기술 텍스트는 트랜스포머 기반 사전학습 모델의 입력 유형으로 변경하였고, 이미지 정보는 독립된 사전학습 과정을 통한 해당 이미지의 벡터들로 고정하여 학습에 반영하였다. 이미지와 텍스트의 통합은 late fusion 접근 방법을 사용하였고, 실험을 통해 FASCODE-EVAL1에서 기존 연구들의 성능을 넘어 서는 결과를 보였다. 또한 제시하는 FASCODE-TEST 모델이 FASCODE-EVAL2의 평가에서 의상 추천 시스템의 독립적인 confidence score 모듈 역할로 활용 가능함을 확인할 수 있었다. 향후 본 연구를 활용하여 실질적인 상용 의상 추천 시스템 개발을 진행할 예정이다.

감사의 글

본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [21ZS1100, 자율성장형 복합인공지능 원천기술 연구]

참고문헌

[1] 정의석, 김현우, 오효정, 송화전, "인터랙션 기반 추천 시스템을 위한 데이터셋 연구", HCLT. 2020.
[2] 박영준, 조병철, 김경선, "Pre-trained model기반의 의상 추천 시스템", ETRI 자율성장 인공지능 경진대회 (fashion-how.org). 2020.
[3] Clark, K., Luong, M. T., Le, Q. V., and Manning, C. D., "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", ICLR. 2020.
[4] J. Park, "KoELECTRA: Pretrained ELECTRA Model for Korean", GitHub. 2020.