1) background : chit-chat모델, 이전 모델

2) intro : 논문에서 제시한 dataset(original&revised persona), 모델소개(이전 모델과 다른점)

3) experiement : 실험 방법, 어떤 의미의 실험인지

automated evalution metrics(자동화된 평가 지표)

크라우드 소싱된 작업자가 모델에 대한 인간 평가를 수행하는 extrinsic evaluation

4) 앞으로의 연구주제는 무엇이 있을지 제시 (논문에 있는 거+우리 팀이 생각하는 거)

Ⅴ. Experiments

automated evalution metrics(자동화된 평가 지표)

크라우드 소싱된 작업자가 모델에 대한 인간 평가를 수행하는 extrinsic evaluation

1. Automated metrics

1) Persona Conditioning

대부분의 모델은 최소한 원래(수정되지 않은) 버전에 대해 자신의 페르소나에 대한 예측을 conditioning(조건화)할 때 크게 향상된다. 이는 단어 overlap이 없는 수정된 버전보다 쉬운 작업이다. 예를 들어, profile memory generation 모델은 seq2seq에 비해 복잡도와 hits@1이 개선되었으며 모든 순위 알고리즘(IR baseline, starspace and profile memory networks)은 향상된 hits@1을 얻는다.

2) Ranking vs Generative

Ranking model은 generative model보다 순위에서 훨씬 낫다. 이것은 아마 그들이 최적화하는 지표이기 때문에 분명할 수 있지만 여전히 성능 차이는 극명하다. Generative model이 사용하는 단어 기반 확률은 잘 작동하지만 순위가 요구하는 문장 기반 확률을 제공할 만큼 충분히 보정되지 않았다. Human evaluation은 이러한 방법을 비교하기 위해 사용된다.

3) Ranking Models

IR baseline은 학습된 유사성 metric을 인해 starspace보다 성능이 우수하고 profiled 대한 attention 메커니즘으로 인해 profile memory network가 성능이 우수하다. (모델의 다른 모든 부분이 동일하기 때문에) 마지막으로 KV profile memory network는 현재 대화와 유사한 train set에서 이웃 대화 기록을 다음 발화 쌍을 고려할 수 있기 때문에 페르소나가 없는 경우 profile memory network보다 성능이 우수하지만 페르소나

정보를 사용할 때 성능이 비슷하다.

4) Revised Personas

수정된 페르소나는 사용하기가 훨씬 더 어렵다. 그러나 profile memory networks가 없는 것에 비해 여전히 약간의 이득이 있다. (0.354vs0.318hits @1) 우리는 또한 두 가지 훈련 변형을 시도했다. Train set의 원래 페르소나 또는 수정된 페르소나와 함께 두 가지 훈련 변형을 시도했다.(Table6) 수정된 페르소나에 대한 교육은 모델이 단순한 단어 겹침 이상을 학습하도록 강요되어 모델이 더 많이 일반화되도록(즉, 다른 구문의 의미론적 유사성 학습) 원래 형식 또는 수정된 형식의 테스트 예제 모두에 도움이 된다.

5) Their Persona

또한 다른 화자의 페르소나 또는 두 페르소나를 동시에 모델에 conditioning할 수 있다.(Table 5, 6) "자신의 페르소나"를 사용하면 data set에 미치는 영향이 적다. 대부분의 speakers는 자신의 관심사에 초점을 맞추는 경향이 있기 때문이라고 생각한다. 이것이 다른 data set에서 얼마나 자주 발생하는지 흥미로울 것이다. 분명히 이것은 군중 작업자에게 줄 수 있는 특정 지침에 의해 왜곡된다. 예를 들어, 자신에 대해 이야기하지 말고 다른 사람의 관심사에 대해 이야기하지 마십시오. 라고 하면 이러한 측정 항목이 변경될 가능성이 있다.

2. Human evaluation

Automated metrics는 대화를 평가하는데 악명이 높기 때문에 크라우드 소싱 작업자를 사용하여 인간 평가도 수행한다. 절차는 3.3절에서와 같이 data set 수집 프로세스 자체에서와 거의 똑같은 설정을 수행한다. 이 설정에서 우리는 Turker 두 명을 짝지어 수집된 pool에서 무작위 원래 페르소나를 각각 할당하고 채팅을 요청한다. 여기에서 Turker의 관점에서 볼 때 Turker와 페어링되는 대신 우리 모델 중 하나와 페어링된다는 점을 제외하면 모든 것이 동일하게 보인다.(이를 알지 못함) 이 설정에서 Turker와 모델 모두에 대해 페르소나는 test set pool에서 나온다.

대화 후, 모델의 품질을 평가하기 위해 Turker에게 몇가지 추가 질문을 한다. 유창함, 참여도 및 일관성을 평가하도록 요청한다. (1-5점) 마지막으로 두가지 가능한 프로필을 표시하여 다른 화자의 프로필을 감지하는 능력을 측정하고 Turker가 방금 말한 사람의 프로필일 가능성이 더 높은 프로필을 묻는다. (부록)

결과는 No Persona 및 Self Persona 범주에서 각각 100개의 대화에서 가장 성과가 좋은 generative 및 ranking model에 대해 보고된다.(Table4) 또한 챗봇을 인간(또 다른 Turker)으로 대체하여 인간의 성과 점수를 평가한다. 이것은 우리가 모델로 목표할 수 있는 상한 점수를 효과적으로 제공한다. 마지막으로 중요한 것은 PERSONA CHAT에서 훈련된 모델

을 vinyals와 Le(2015)에 이어 Twitter 및 OpenSubtitles data set(2009 and 2018 versions)로 훈련된 잡담 모델과 비교하는 것이다. (일부 모델 채팅 예시 : 표 7, 8, 9, 10, 11, 12)

첫째, 모든 PERSONA CHAT 모델과 OpenSubtitles 및 Twitter에서 훈련된 모델 간의 유창성, 참여도 및 일관성의 차이를 확인한다. PERSONA CHAT는 두 화자가 서로를 모르는 경우 다른 리소스와 달리 질문 및 답변에 중점을 두는 대화 시작을 위한 교육 데이터를 제공하는데 특히 강력한 리소스이다. 또한 모델 간의 미묘한 차이에 대한 제안도 볼 수 있지만 이러한 차이는 인간 평가자의 평가의 높은 분산으로 인해 모호해진다. 예를 들어, generative 및 ranking 모델 사례 모두에서 페르소나가 부여된 모델은 페르소나 감지 정확도에 의해 입증 된대로 인간 대화 파트너가 이 모델을 할 수 있는 동시에 비 페르소나 기반 대응 모델과 비교하여 우창성과 일관성을 유지할 수 있다.

유창함, 참여도, 일관성 및 지속적인 성격 사이의 균형을 찾는 것은 향후 연구에서 여전히 어려운 과제이다.

3. Profile Prediction

PERSONA CHAT를 사용하여 자연스럽게 대화 중 다음 발화 예측(태스크1), 대화 내역에 따른 프로필 예측(태스크2)을 고려할 수 있다. 이 작품의 주요 연구는 프로필 정보의 사용을 보여주는 태스크1이다. 그러나 태스크2는 이러한 정보를 추출하는데 사용할 수 있다. 전체 연구는 이 논문의 범위를 벗어나는 동안 몇 가지 예비 실험을 수행했다.(부록 D) 인간 화자의 프로필은 높은 정확도로 대화(표4, 94.3%, 인간 성과와 유사) 또는 모델이 인간의 관심사에 주의를 기울이고 있음을 보여주는 모델의 대화(KV 프로필 메모리 23%)에서 예측할 수 있다. 또한 table14에 표시된대로 추가 대화를 통해 정확도가 명확하게 향상된다. 태스크1과 태스크2를 전체 시스템으로 결합하는 것은 미래 연구의 흥미로운 영역이다.

VI. Conclusion & Discussion

이 작업에서 우리는 각 참가자가 할당된 페르소나의 역할을 수행하는 군중소스 대화로 구성된 PERSONA CHAT dataset를 도입했다. 그리고 각 (군중 소스)페르소나는 단어 고유의 의역을 가지고 있다. 이 data set에서 다양한 기준 모델을 테스트하고 대화 상태 외에 자체 페르소나에 접근할 수 있는 모델이 더 매력이지는 않지만 annotator에 의해 더 일관된 것으로 점수가 매겨짐을 보여준다. 반면에, 우리는 PERSONA CHAT(페르소나 유무에 관계없이)에서 훈련된 모델이 다른 리소스(영화, 트위터)의 대화로 훈련된 모델보다 더 매력적이라는 것을 보여준다.

PERSONA CHAT은 미래 대화 시스템의 구성요소를 교육하는데 유용한 리소스가 될 것이다. 인간이 생성한 프로필과 대화를 짝지었기 때문에 데이터는 일관된 성격과 관점을 가진 에이전트를 구성하는데 도움이 된다. 또한 대화에서 프로필을 예측하면 성공에 대한 지표가 있는 목표 지향적 대화 방향으로 잡담 작업이 이동한다. 프로필의 의역을 수집하기 때문에 사소하게 일치시킬 수

없다. 실제로 우리는 원래의 프로필과 다시 표현된 프로필이 그 자체로 의미론적 유사성 data set 로서 흥미롭다고 믿는다. 데이터가 사용자 프로필에 대해 질문하고 답변을 기억하고 대화에서 자연스럽게 사용할 수 있는 교육 에이전트를 지원하는데 도움이 되기를 원한다.

A. Next Utterance Prediction Additional Evaluation Metrics

(Table 5,6)다음 발화 예측에 대한 추가 결과. 우리는 one's own의 ("Self") 또는 조합 ("둘다") 대신 다른 사람의 페르소나 ("Their")를 조건화하는 결과를 제공한다. 특히 다른 사람의 페르소나를 아는 것은 이러한 모델을 사용하여 이 데이터의 정확성에 도움이 되지 않는다는 것을 알 수 있다. 마지막으로, ranking 모델에서 (Table6) 원래 페르소나와 수정된 페르소나에 대한 훈련의 성능 차이를 보여준다. 수정된 페르소나는 개선된 결과를 제공한다. 아마도 모델이 사소한 단어 overlap(즉, 서로 다른 구문의 의미적 유사성) 이상을 학습해야하기 때문일 수도 있다.

B. Example Dialogs between Humans and Models

(표 7, 8, 9, 11, 12)5.2절의 human evaluation의 일부로 수집된 다양한 모델과 Turker 간의 대화 예를 보여준다.

C. Human Evaluation Measures

인간과 모델 간의 대화 후 Turker에게 모델의 품질을 평가하기 위해 몇가지 추가 질문을 한다. 순서는 아래와 같다.

1) Fluency : 1~5점의 점수로 다른 화자의 유창성을 평가하도록 한다. 1은 전혀 유창하지 않음, 5는 매우 유창함, 3은 OK

2) Engagingness : 유창함을 무시하고 다른 화자의 참여도를 판단하도록 함.

3) Consistency : 대화의 일관성

4) Profile Detection : 두가지 가능한 프로필을 표시하고 Turker가 방금 말한 사람의 프로필일 가능성이 더 높은 프로필을 묻는다. 하나의 프로필은 무작위로 선택되고 다른 하나는 모델에 주어진 진정한 페르소나이다.

D. Profile Prediction

이 논문의 주요 연구는 페르소나를 조건화하여 다음 발화 분류를 향상시키는 능력이지만 자연스럽게 대화 중 다음 발화 예측 및 대화 이력에 따른 프로필 예측을 고려할 수 있다. 주요 논문에서 프로필 정보를 사용하여 과제1을 개선할 수 있다. 그러나 과제2는 이러한 정보를 추출하는데 사용할 수 있다.

여기에서는 일련의 대화 발화에서 화자의 페르소나를 예측하는 능력에 대한 예비 연구를 수행한다. 인간 (페르소나0)과 가장 성능이 좋은 모델인 5.2절의 검색 기반 KeyValue Profile Memory Network(페르소나1)간의 대화를 고려한다. 네가지 조합을 모두 고려하여 각 화자의 대화 발화에서 두 화자의 프로필 정보를 예측하는 기능을 테스트했다. 주요 논문에서 사용된 것과 동일한 IR 기준 모델을 사용하여 프로필을 예측한다. 전체 프로필 수준(프로필을 구성하는 모든 문장을 가방으로 고려)또는 문장 수준 (각 문장을 개별적으로 고려)에서 프로필 후보의 순위를 매긴다. 각 긍정적인 프로필에 대해 100개의 부정적인 프로필 후보를 고려하고 모든 대화 및 후보에 대해 평균의 실제 프로필을 예측하는 오류율을 계산한다. (표 13) 프로필 정보에 조건이 지정된 모델과 그렇지 않은 동일한 KV 메모리 모델에 대해 제공된다. 결과는 아래와 같다.

1) 대화 발화(페르소나0, 프로필0)에서 말하는 모델에 관계없이 프로필과 문장 수준 모두에서 높은 정확도로 인간 프로필을 예측할 수 있다.

2) 마찬가지로 모델의 프로필은 프로필에서 조건이 지정될 때 발화(페르소나1, 프로필1)에서 높은 정확도로 예측할 수 있다. 그렇지 않으면 이는 기회 수준이다.(프로필 없음)

3) 인간의 대화에서 모델의 프로필을 예측할 수 있지만 모델이 자체 프로필에 조건을 지정하는 한 정확도가 낮다.(페르소나0, 프로필1) 이것은 인간이 모델의 발화에 반응하고 모델의 관심사에 주의를 기울임을 나타낸다.

4) 마찬가지로 인간의 프로필은 모델의 대화에서 예측할 수 있지만 정확도는 낮다. 흥미롭게도 프로필 컨디셔닝이 없는 모델이 더 낫다. 아마도 자신에 대해 이야기하는데 집중하지 않고 인간의 관심사에 더 많은 관심을 기울이기 때문일거다. 여기에서 탐구하고 이해해야 하는 절충안이 있는 것 같다.

또한 대화 길이 1~8(이 작업에서 가장 긴 길이)에 대한 오류율을 계산하여 대화가 진행됨에 따라 프로필 예측 성능을 연구한다. (표14)대화 길이가 증가함에 따라 모든 경우에서 페르소나 예측 오류율이 감소함을 보여준다.

전반적으로 이 섹션의 결과는 중요한 추출 작업인 대화 발화가 주어진 프로필을 예측하는 것이 타당함을 보여준다. 더 정교한 모델을 사용하면 더 나은 결과를 얻을 수 있다.