A Knowledge Grounded Neural Conversation Model

Review

1. Introduction

a. 이전 Conversation neural network model 은 실제 세상에 적용하기 어렵고, 다양한 하드코딩된 패턴이 필요해서 비효율 적이다(패턴이 무한대). 또한 외부데이터 접근에 어려움을 겪었다
b. 이 논문이 제안하는 건 Fully data driven Seq2seq 모델을 베이스라인으로 사용해서 다양한 영역에서 conversation 모델을 사용할 수 있게 하는 것이다
c. 주요 포인트
    i. 대화 sequence 의 history 를 사용해서 더 자연스러운 대화를 생성
    ii. External data 에서 fact 를 찾는 것
        1. External data 는 23M 일반적인 도메인 conversation data(Twitter)
        2. 1.1M Foursquare Data

2. Related work

a. 이전 conversation generation 방식
    i. Text data 활용해서 machine translate 와 Q&A 등 neural network 확장과 conversation system 구축
    ii. Task based conversation 발전
        1. example
            a. 식당 예약
            b. 비행기 예약
    iii. 이 논문 approach 는 반대로 general 한 대화가 목적
        1. 다양한 도메인에 쓸 수 있도록 만드는 것
        2. 방대한 양의 데이터 필요

3. Grounded Response Generation

a. Conversation data set 문제점

i. 다양한 데이터 셋이 존재하지만 그 데이터 셋의 대화에서 모든 지식이 대화에 나타나지 않는 것
ii. 그리고 대화 주제에 bias 된 것
iii. 중복되는 데이터 패턴이 나온다는 것

b. Knowledge-grounded model architecture

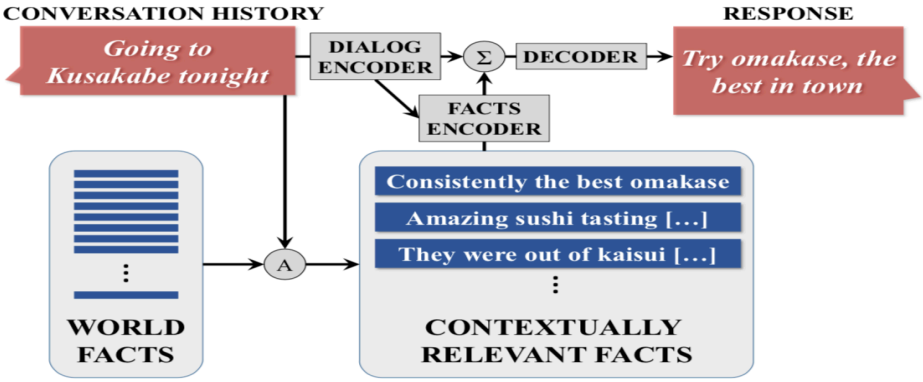


Figure 3: Knowledge-grounded model architecture.

i.

ii. Word Fact (Foursquare, Wikipedia, Amazon)

1. Indexed NER 데이터

iii. Source Sequence S :

1. fact word focusing 하고 keyword matching -> entity linking or NER 이후 focusing 된 Sentence 를 가지고 모든 contextually relevant fact $F = \{f_1,...,f_k\}$ 에서 Query 실행

iv. 마지막으로 (1) Conversation history 와 (2) relevant fact 가 2 개의 Seq2seq neural network 에 fed 로 들어감

c. 장점:

i. Knowledge-grounded approach 는 Seq2seq response generation 보다 훨씬 일반적인데, 관심을 가지는 각 개별 entity 에 대한 동일한 conversation pattern 을 학습하는 것을 피할 수 있기 때문
ii. vocabulary 에 단어가 없어도 최대한 매칭되는 Contextually Relevant Facts 를 사용해서 처리 가능

d. Encoder multi-task learning

i. 2Type
   1. (S, R)
      a. **S: representation the conversation history**
      b. **R: response**

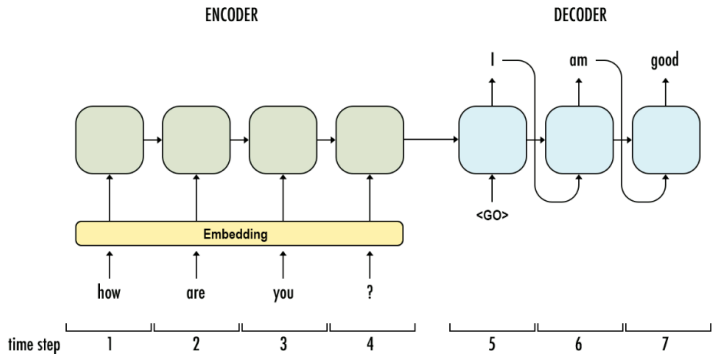
2. Full model
    a. ({ F , S }, R)
        i. F-> Contextually Relevant Facts set

ii. 2 개의 Training type 을 가지면서 장점:

1. 대화 전용 데이터세트를 미리 훈련시키고, 이미 대화의 근본을 학습한 대화(Free training) 인코더와 디코더로 멀티 태스킹**(웜 스타트)**을 시작할 수 있다.
2. 두 가지 작업에서 서로 다른 종류의 데이터를 사용할 수 있는 유연성을 제공한다.
3. (R = fi) 중 하나로 대체 가능
    a. 자동 인코더와 유사하게 만들고 훨씬 더 만족스러운 응답을 생성하는데 도움을 준다.

e. Dialog Encoder-Decoder

i. Seq2seq model 사용

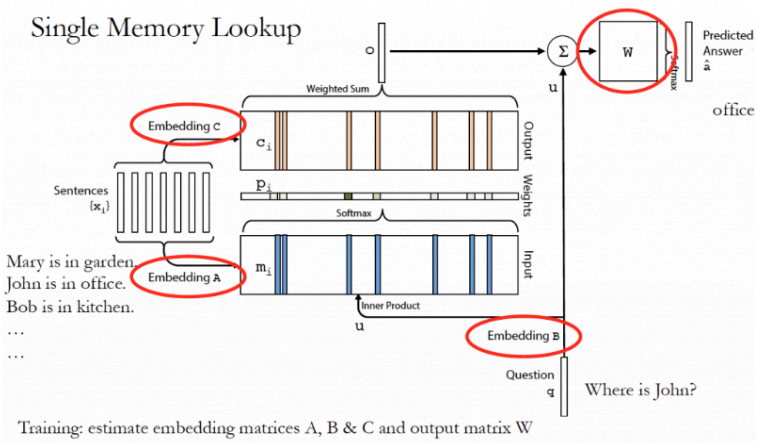


1.
2. No share weight or word embedding

f. Fact Encoder

i. Memory network 기반

1. http://solarisailab.com/archives/690



2.

3. Memory network 에 사용되는 Bag of word 대신 RNN representation vector 를 사용
4. **u is the summary of the input sentence.**
5. **ri is the bag of words representation of fi**
6.

$$m_i = Ar_i \quad (1)$$

$$c_i = Cr_i \quad (2)$$

$$p_i = \text{softmax}(u^T m_i) \quad (3)$$

$$o = \sum_{i=1}^{k} p_i c_i \quad (4)$$

7. $$\hat{u} = o + u \quad (5)$$

g. Data set

i. Foursquare
   1. 레스토랑 및 다른 상업적 공간에 대한 의견을 모은 데이터
   2. Twitter 대화에서 사용된 데이터로 한정

ii. Twitter
   1. 23M general dataset 3-turn 대화 -> background data set
   2. 1 million two-turn conversations 을 Twitter data 에서 모았는데 여기에 Foursquare entity 포함 -> 1M grounded dataset 이라 부름
   3. 첫 번째 대화턴에 entity tag 를 달아서 어떤 형태의 Business 인지 확인하고, 이를 통해서 Foursquare data 의 service agent 가 대답한 것을 제거
      a. 왜냐 Real conversation 을 찾고 싶어서

h. Grounded Conversation Dataset

i. 1M grounded dataset
ii. 입력 단어와 Tip 사이의 cosine similarity 를 적용해서 가장 높은 10 개의 Tip 을 유지
iii. 2 개의 Score function 사용
   1. 학습된 1-gram LM

2. χ-square score : 얼마나 많은 콘텐트가 들어 있는지 평균점수

iv. 크라우드 소싱: 사람이 평가한 4k validation 과 test set 구성

4. Experimental setup

a. Multi-Task Learning(3 개)
   i. Facts-task
   ii. NoFacts-task
   iii. Autoencoder task
      1. Facts-task -> 에서 R 을 F 로 바꾸는 것

b. Learning system
   i. Seq2seq : NoFact : 23M general conversation dataset
      1. Only one task
   ii. MTask : Nofact 2 개 23M general dataset and 1M grounded dataset
   iii. MTask-R
      1. NoFact, Fact task
   iv. MTask-F
      1. NoFact, Autoencoder
   v. MTask-RF
      1. Task 3 개다

c. Network structure
   i. One-layer memory network
   ii. Two-layer Seq2eq model

d. Decoding and Reranking
   i. $\log P(R|S,F)+\lambda \log P(S|R)+\gamma|R|$
   ii. Response 의 점수 계산

e. Evaluation Metrics
   i. BELU

5. Result

a. Automatic Evaluation:
   i. BELU and Perplexity
      1. Perplexity: 어떤 인식 태스크의 인식 난이도를 나타내는 지표는 인식 대상 어휘수와 문법의 복잡도를 고려한 perplexity 가 있다.

Perplexity 는 어떤 시점에서 선택할 수 있는 인식 대상 후보의 평균 갯수의 의미를 갖는다

Model	BLEU	Diversity 1-gram	Diversity 2-gram
SEQ2SEQ	0.55	4.14%	14.4%
MTASK	0.80	2.35%	5.9%
MTASK-F	0.48	9.23%	26.6%
MTASK-R	**1.08**	7.08%	21.9%
MTASK-RF	0.58	**8.71%**	**26.0%**

ii. Table 2: BLEU-4 and lexical diversity.

1. 주목할 점 :
    a. MTask-R: Fact data injection 의 효과가 있다 !!
    b. Diversity : MTask-RF 가 3 가지 데이터(NoFact, Fact, Autoencoder) 사용해서 Diversity 가 높다

b. Human Evaluation :

i. 문맥의 (1)적절함, (2)유익함 기준 평가

ii. MTask-R system 이 두가지 모두 좋게 평가됨. 나머지는 두가지 Task 평가 모두 적절함과 유익함에서 균형을 이루지 못함

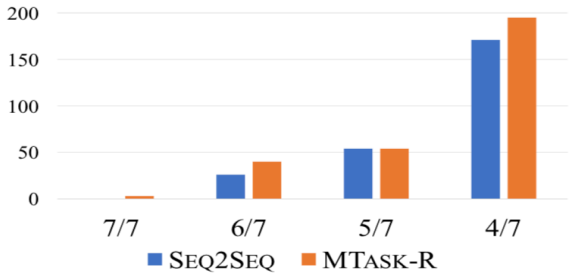


Figure 4: Judge preference counts (appropriateness) for MTASK-R versus SEQ2SEQ.

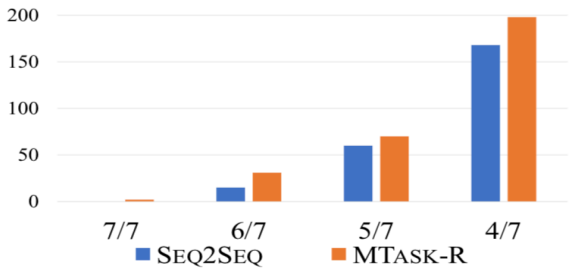


Figure 5: Judge preference counts (informativeness) for MTASK-R versus SEQ2SEQ.

iii.

1. 사람들이 봤을 때 적절한 대화라고 판단된 것 들을 투표 7/7 은 7 명중에 7 명 모두 유익한 대화라고 생각되는 문장들을 평가
2. MTask-R 이 Seq2seq 모델보다 좋은 평가

6. Discussion:

a. 실재 생성된 대화 패턴

1	A: I would give ALMOST anything for some [*Mexican restaurant*] right now. B: Me too. Creamy jalapeno sauce is really good.
2	A: Visiting the celebs at Los Angeles International Airport (LAX) - [...] w/ 70 others B: Nice airport terminal. Have a safe flight.
3	A: Obsessed with [*jewelry company*] :-* B: oh my gosh obsessed with their bracelets and the meaning behind them !
4	A: Live right now on [*FM radio*] Tune in!!!!! B: Listen to Lisa Paige
5	A: Is that [...] in your photos? It's on my list of places to visit in NYC. B: Don't forget to check out the 5th floor, while you are here, it's a great view.
6	A: Are you [*coworking community*] today or tomorrow? B: I think so. Try the cookies and coffee.
7	A: Cant wait for my [*pizzeria*] tonight B: Order bacon n pepperoni thin crust pizza.
8	A: So [...] is down to one copy of Pound Foolish. I'm curious to see if they are re-ordering it. B: Check out the video feed on 6 and take a picture of the Simpsons on the 3rd floor.
9	A: I wish [...] would introduce another vegetarian option besides the shroomburger. It's delicious but kind of ridicu B: This is the best j.crew in the world. Try the lemonade!
10	A: Just had an awesome dinner at [*restaurant*] Great recommendation [*person*] B: One of my favorite places I've ever been to in NYC. The food is great and the service is lackluster.

Figure 6: Responses produced by our neural response generation system MTASK-RF. Entities marked [...] are anc

b.
   i. 2 번은 엉뚱한 대답 하는 경우가 있고, 나머지는 그럭저럭이다

7. 결론

a. 일단 Vocabulary 이 없이 response 한다는 점은 좋은 것 같다
b. Conversation data preprocessing 하는 부분이 설명되어 있어서 좋았음 (Non-conversation and Conversation data)
c. 아마 데이터가 general 한게 위에처럼 대답이 되는 것 같다. 생각보다 성능이 좋은지는 잘 모르겠고, 역시 General 하다는 것 자체의 문제가 있다고 판다
d. 다른 Goal oriented Chatbot 이 특정 도메인에서 더 자연스러운 대화를 이끌어낸다.
e. 하지만 역시 General 한 대화는 모든 Chatbot 의 궁극적인 목표이니 연구할 충분한 이유가 있다고 생각된다.