대화 시스템에서의 자연스러운 대화를 위한 Memory Attention기반 Breakdown Detection

이설화○, 박기남, 임희석
고려대학교 정보대학 컴퓨터학과
whiteldark@korea.ac.kr, spknn@korea.ac.kr, limhseok@korea.ac.kr

Memory Attention-based Breakdown Detection for Natural Conversation in Dialogue System

Seolhwa Lee○, Kinam Park, Heuiseok Lim
Korea University

요 약

대화 시스템에서 사람과 기계와의 모든 발화에서 발생하는 상황들을 모두 규칙화할 수 없기 때문에 자연스러운 대화가 단절되는 breakdown 현상이 빈번하게 일어날 수 있다. 이런 현상이 발생하는 이유는 다음과 같다. 첫째, 대화에서는 다양한 도메인이 등장하기 때문에 시스템이 커버할 수 있는 리소스가 부족하며, 둘째, 대화 데이터에서 학습을 위한 annotation되어 있는 많은 양의 코퍼스를 보유하기에는 한계가 있으며, 모델에 모든 대화 흐름의 히스토리를 반영하기 어렵다. 이런 한계점이 존재함에도 breakdown detection은 자연스러운 대화 시스템을 위해서는 필수적인 기능이다. 본 논문은 이런 이슈들을 해소하기 위해서 memory attention기반의 새로운 모델을 제안하였다. 제안한 모델은 대화내에 발화에 대해 memory attention을 이용하여 과거 히스토리가 반영되기 때문에 자연스러운 대화흐름을 잘 detection할 수 있으며, 기존 모델과의 성능비교에서 state-of-the art 결과를 도출하였다.

주제어: 대화시스템, memory attention, breakdown detection

1. 서론

대화 시스템에서 중요하게 다루는 이슈(Issue) 중 하나는 사람과 시스템간의 대화가 사람간의 대화처럼 자연스럽게 지속되지 않는 경우가 많다는 것이다. Martinovsky et al.(2006)[1]은 사람과 시스템 사이의 대화에서 문맥의 흐름이 끊겨 사람이 더 이상 대화 흐름에 참여할 수 없게 되는 상황일 때, 시스템의 발화가 breakdown을 유발하는지를 탐색하는 것을 breakdown detection이라고 정의하였다. 즉, 대화 시스템에서의 breakdown은 시스템이 대화의 문맥 흐름에 맞지 않는 부적절한 발화를 생성 하였을 때 발생하며, 이런 대화의 흐름이 끊기는 현상을 탐색해내는 것이 breakdown detection이라고 할 수 있다. 표 1은 대화 시스템내에서 발생하는 breakdown에 대한 예시이다. 'User' 는 비가 싫어서 저녁에 집에 있겠다고 하였으나 시스템은 반대되는 의견을 제시하여 'Turn 3' 에서 시스템의 발화가 breakdown을 유발한 것을 확인할 수 있다.

대화 시스템내에 breakdown이 발생하는 이유는 크게 두 가지로 나뉠 수 있다. 첫째, 사람과 시스템이 발화를 여러번 주고 받는 멀티 턴(Multi-turn)으로 인해 시스템이 사용자의 의도를 잘못 이해하여 발생할 수 있다. 둘째, 대화에서 많은 주제들을 다루며, 그에 알맞은 반응을 해야하는데 대화의 흐름을 모두 파악하기 어렵기때문에 시스템이 적절한 답변을 하는 것에 어려움이 있을 수 있다. 다음과 같은 이유들로 인해 시스템이 잘못된 발화를 생성하는 breakdown이 유도될 수 있다.

대화 시스템은 크게 목적지향(Goal/Task-oriented) 대화 시스템[2]과 대화지향(Chat-oriented) 대화 시스템[3] 두 가지로 분류할 수 있다. 목적지향 대화 시스템은 식당 예약, 항공권 예약 등의 명령 전달의 특정 목적을 수행하기 위한 대화 시스템이다. 반면, 대화 지향 대화 시스템은 챗봇(Chat-bot)이 이에 해당되며, 사람과 시스템간의 자연스러운 대화를 하는데 목적이 있다.

Breakdown detection 기술은 대화지향 대화 시스템과 같이 자연스러운 대화 유지가 중요한 경우에 유용할 수 있다. 또한 대화 시스템에서 시스템이 잘못 발화한 경우에 대한 error recovery에도 활용될 수 있다.

과거 대부분의 대화지향 대화 시스템의 경우 모델을 학습시키기 위해서 사람이 방대한 학습 코퍼스에 직접 annotation하여 hand-crafted feature를 선별하는 과정을 거쳤었지만, 이런 과정은 도메인 영역이 작을 수 있고 모델을 개발하기까지 많은 시간이 소비될 수 있다. 최근에는 이런 문제점들을 해결하기 위해서, end-to-end[4]방식을 사용하는 방법들이 제안되고 있다.

End-to-end 방식은 hand-crafted feature에 의존적이지 않으며 긴 대화를 유지하는데 유용하다[5]. 하지만, 이와 같은 연구가 있음에도 불구하고 멀티 턴이 존재하는 연속성이 커다란 대화 시스템에서 일일이 문맥의 흐름을 파악하여 breakdown detection을 하는 것에는 어려움이 있다.

표 1 대화 시스템에서의 breakdown 예시

Turn	Speaker	발화
1	Sys1	오늘은 축구하기에 좋은 날씨네요.
	User1	맞아, 그런데 곧 비가 온대.
2	Sys2	일기예보에서 저녁에 비가 온다고 했어요.
	User2	나는 비가 싫어서 저녁에 집에 있을 거야.
3	Sys3	그러면 저녁에 나가서 축구하면 되겠네요. ***(Breakdown)***

따라서, 본 논문은 이런 한계점을 해소하기 위해서 memory attention기반의 temporal utterance encoding 방법을 제안한다. 제안하는 방법은 end-to-end 기반의 breakdown detection모델이며, LSTM(Long short-term memory)을 이용하여 대화내에 사용자와 시스템의 발화를 인코딩하고 시스템 발화에 대해 memory network[6]기반의 attention 기법을 이용하여 breakdown detection을 수행하는 구조를 가지고 있다. 본 논문의 구성은 다음과 같다. 2장에서는 대화 시스템에서의 breakdown detection에 대한 관련 연구, 3장에서는 제안한 모델 설명, 4장에서는 실험 결과, 5장은 결론으로 구성된다.

2. 관련 연구

대화 시스템에서 breakdown detection은 여전히 발전의 여지가 많은 유망한 태스크 중 하나이다. 해당 태스크는 Higashinaka et al.(2016)[7]에 의해서 처음으로 제안되었다. Breakdown detection 태스크에서 전통적인 방법과 딥 러닝 기반의 방법에 대한 초기 연구는 사람과 시스템 발화에 대한 키워드 추출을 통해 breakdown이 되기까지 휴리스틱(Heuristic)규칙을 이용한 rule-based 방법[8]과 LSTM-RNN[9]을 이용해 빈번히 나오는 단어에 대한 벡터를 feature로 구성하는 딥러닝 기반의 방법이 있다. 규칙기반 방법은 규칙을 휴리스틱하게 일일이 비교해보며 찾아야 하기 때문에 시간이 많이 소비되며, LSTM-RNN 기반의 모델 또한 멀티 턴을 모두 반영하기에는 부족함이 있다. Iki et al.(2017)[11]의 연구에서는 external memory network를 사용하여 character-level convolutional attention을 이용하였다. 해당 모델 또한 기존 한계점을 해소하기에는 한계가 있다.

기존 연구들은 멀티 턴에 대한 문맥의 흐름을 파악하는데 한계가 있기 때문에 시스템 발화에 대한 히스토리 정보를 이용하면 해소할 수 있다.

3. Temporal Utterance Encoding 모델

그림 1은 본 연구에서 제안한 memory attention기반 temporal utterance encoding모델에 대한 개요이다. 그림에서 $\oplus, \otimes, \sum, w$는 각각 concatenation, 매트릭스 multiplication, summation, 문장을 구성하는 단어를 의미한다. 본 모델의 학습 과정은 다음과 같다. (1) 사용자 발화 및 시스템 발화에 대한 sentence representation을 수행한다.

$$\frac{1}{n}\sum_{i=1}^{n} w^{t_n} \quad (1)$$

(2) 현재 시점 t에서 시스템 $s^t = \{w_1^t, w^{t_2}, ..., w_n^t\}$ 및 사용자 $u^t = \{w_1^t, w^{t_2}, ..., w_n^t\}$의 발화를 인코딩(Encoding)하기 위해 LSTM을 이용하여 h_t를 도출한다.

$$\begin{aligned} x_t &= u^t \oplus s^t \\ h_t &= LSTM(x_t) \end{aligned} \quad (2)$$

(3) LSTM으로부터 획득한 인코딩 벡터와 현재 시점에서 모든 이전 시스템 발화에 대한 memory를 저장하여 attention을 이용한 attention weight값을 도출한다. $e_{ti}, m_i^T, W_{\text{attn}}$는 각각 attention score, memory 벡터, 학습을 통해 얻는 weight matrice이다. $\alpha_i, \overline{m_t}$는 각각 attention weight, memory에 대한 context 벡터이다.

$$\begin{aligned} e_{ti} &= m_i^T W_{\text{attn}} h_t \\ \alpha_i &= \frac{\exp(e_{ti})}{\sum_{i=1}^{t} \exp(e_{ti})} \\ \overline{m_t} &= \sum_{i=1}^{t} \alpha_i m_i \end{aligned} \quad (3)$$

(4) 마지막으로 대화 시스템내의 발화에서 breakdown을 예측한다. W_o, b_o는 각각 weight metrice와 bias 벡터이다.

$$\hat{y} = softmax(W_o(h_t \oplus \overline{m_t}) + b_o) \quad (4)$$

3. 실험

3.1 데이터셋

개발 데이터는 DBDC3[7]에서 제공한 데이터를 사용하였으며, 해당 데이터에는 20 또는 21개의 발화로 구성된 대화 데이터 415개로 구성된다. 모든 대화 데이터에는 크게 3종류의 레이블링이 되어 있는데, breakdown이 아닐때는 NB(Not breakdown), breakdown일 가능성이 높을 때는 PB(Possibly breakdown), breakdown일 때는 B(Breakdown)으로 구성된다. 레이블링은 사람이 voting 방법을 통하여 태깅을 하였다[7]. 평가 데이터 역시 [7]의 연구에서 오픈한 데이터셋을 이용하였다.

3.2 평가 방법 및 모델 성능 비교

평가 방법은 Accuracy와 F1 score를 측정하였으며,

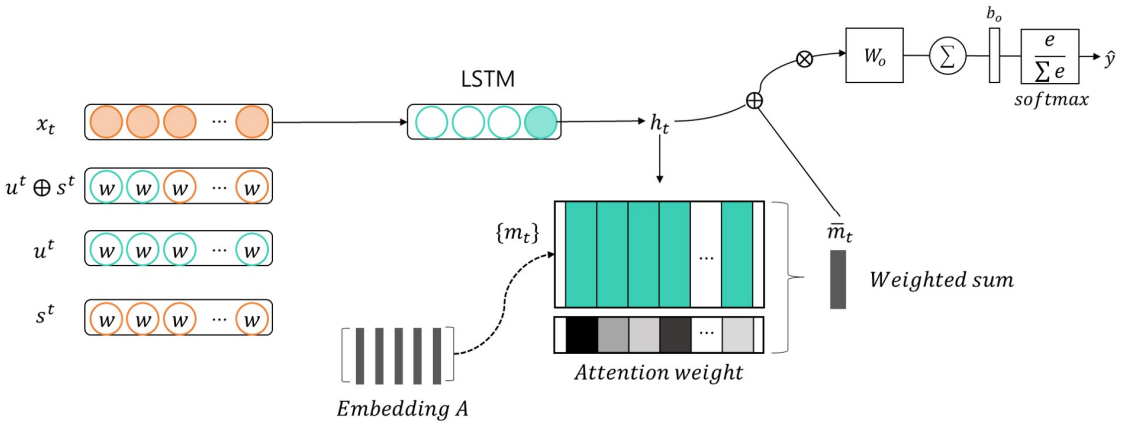

그림 1 Memory attention기반 temporal utterances encoding 모델 개요

[7]에서 보고한 기존 연구들과 성능 비교를 수행하였다.

모델 성능 비교는 Baseline으로 CRF(Conditional random field)와 Majority(레이블 중 가장 빈번하게 나온 확률분포의 평균 값)와 DBDC3 태스크에서 accuracy와 F1 score에서 제일 성능이 잘 나왔던 KTH[11], PLECO[4] 모델과 하였다.

3.3 학습 파라미터

sentence representation은 위키피디아 영어 데이터를 이용하여 word2vec 300차원으로 학습시킨 모델을 이용했으며, LSTM은 128차원으로 학습하였다. 학습 loss optimizer로는 Adadelta[10]를 사용하였으며, learning rate는 0.1, batch 사이즈는 10으로 설정하였다. 최종 학습 epoch는 30으로 설정하였다.

3.4 결과

표 2는 제안한 모델과 다른 연구들과의 성능 비교를 나타낸 것이다. 제안한 모델은 Accuracy에서 제일 좋은 성능을 획득하였으며, breakdown 레이블링에 대한 F1 score에서도 제일 좋은 성능을 보였다. 하지만 PB와 B 레이블링에 대한 F1 score값은 Majority baseline이 가장 높게 나왔다. 이러한 결과가 나온 이유는 데이터셋의 레이블링 구성 비율이 PB와 B에서 제일 낮았기 때문에 학습에 반영되기에 충분한 양이 아니어서 발생할 수 있는 결과라고 유추할 수 있다.

표 2 모델 성능 평가 결과

	Accuracy	F1(B)	F1(PB+B)
CRF baseline	0.4285	0.3543	0.7622
Majority baseline	0.3720	0.3343	**0.8927**
KTH run2	0.4415	0.2949	0.7440
PLECO run1	0.2950	0.3636	0.8744
제안 모델	**0.4515**	**0.5130**	0.6426

5. 결론

대화 시스템에서 breakdown detection은 자연스러운 대화를 위해서는 필수적인 태스크이지만 아직까지 많은 연구가 이루어지고 있지는 않다. 따라서, 본 연구에서 제안한 방법을 시작으로 더 많은 연구가 이루어져 챗봇 시스템의 성능을 더 높일 수 있기를 기대해본다.

Acknowledgement

이 논문은 2018년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.NRF-2018R1D1A1B07047618)

참고문헌

[1] Martinovsky, Bilyana, and David Traum. The error is the clue: Breakdown in human-machine interaction. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES, 2006.

[2] Williams, J.D. and Young, S.. Partially observable Markov decision processes for spoken dialog systems. Computer Speech & Language, 21(2), pp.393-422, 2007.

[3] Zhou, Xiangyang, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view response selection for human-computer conversation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 372-381. 2016.
[4] Iki, Taichi, and Atsushi Saito. End-to-end Character-Level Dialogue Breakdown Detection with External Memory Models. Dialog System Technology Challenges, 2017.
[5] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models. In AAAI, volume 16, pp,3776-3784, 2016.
[6] Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. End-to-end memory networks. In Advances in neural information processing systems, pp. 2440-2448. 2015.
[7] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In LREC, 2016.
[8] Unno Y. Kobayashi, S. and M Fukuda. Multitask learning of recurrent neural network for detecting breakdowns of dialog and language modeling. JSAI Technical Report (SIG-SLUD-75-B502), pp. 41-46, 2015.
[9] R. Taniguchi and Y Kano. Construction of automatic detector for dialogue breakdowns based on rules with keywords extraction. JSAI Technical Report (SIG-SLUD-75-B502), pp. 37-40, 2015.
[10] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. CoRR, abs/1212.5701, 2012.
[11] Jos´e Lopes. How generic can dialogue breakdown detection be? the kth entry to dbdc3, 2017.