I like fish especially dolphins : Addressing Contradictions in Dialogue Modeling

01. Problem

Figure 1: Two dialogue examples demonstrating a state-of-the-art chatbot (B) (Roller et al., 2020) contradicting itself when talking to a human (A).

- 문맥상 일관적이지 않은 대화는 유저의 사용감을 현저히 떨어뜨린다.

02. Solution

1. Dialogue Contradiction Detection Task 소개
2. 데이터 수집 : 화자 중 한 명이 대화 중 특정 지점에서 이전에 말한 것과 의도적으로 모순되는 대화 수집.
3. 모델 비교 : Unstructured와 Structured 방법을 비교.

02. Solution

DECODE

- Dialogue Contradiction Detection를 supervised classification task로 설정.
- 입력 x = {문장1, 문장2, 문장3, … , 문장n}
- 출력 y = 마지막 발화 문장n이 이전에 대화와 모순되는지 여부를 0과 1로 판단.

02. Solution

Data Collection

1. 다양한 토픽을 담고 있는 Wizard of Wikipedia (Dinan et al., 2018), EMPATHETICDIALOGUES (Rashkin et al., 2019), Blended Skill Talk (Smith et al., 2020a), and ConvAI2(Dinan et al., 2020) 대화 데이터를 미리 수집한다.
2. 해당 대화 문맥을 기반으로 사람이 직접 DECODE Task 데이터를 생성한다.
3. 생성한 데이터의 품질을 세가지 방법(Onboarding Test, Maximum Annotation Count Limit, Verification)으로 체크한다.

02. Solution

Conversation Data

	Train	Dev	Test
Wizard of Wikipedia	6,234	1,208	1,160
EMPATHETICDIALOGUES	6,182	1,046	1,050
Blended Skill Talk	8,554	1,200	1,310
ConvAI2	6,214	572	696
Total	**27,184**	**4,026**	**4,216**

Table 1: Our DECODE Main Dataset source statistics. The labels in each split are balanced. There are a total of 2,013+2,108 contradicting examples in the dev and test sets which are the collected 4,121 verified examples. The first column indicates the source of the dialogue.

User: I'm looking for a cheaper restaurant
`inform(price=cheap)`
System: Sure. What kind - and where?
User: Thai food, somewhere downtown
`inform(price=cheap, food=Thai, area=centre)`
System: The House serves cheap Thai food
User: Where is it?
`inform(price=cheap, food=Thai, area=centre); request(address)`
System: The House is at 106 Regent Street

Figure 1: Annotated dialogue states in a sample dialogue. Underlined words show rephrasings which are typically handled using semantic dictionaries.

- 해당 대화 데이터를 기반으로 Annotation Tool을 사용해 데이터를 변형한다.

02. Solution

Annotation Interface

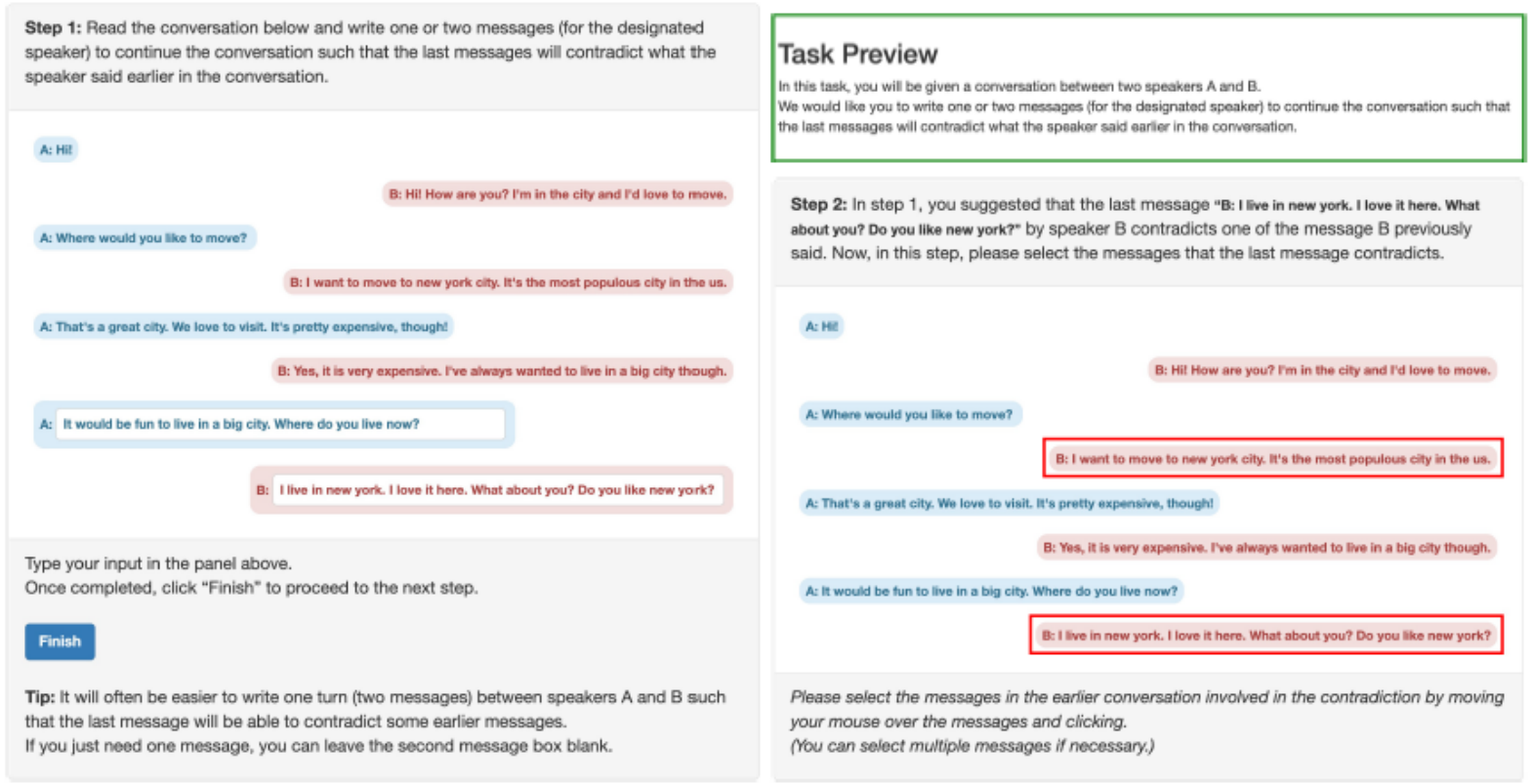

Figure 2: The collection interface. The task preview box (top right) gives a short description of the task before the annotator will work on the writing. The collection consists of two steps. In Step 1 (on the left), the annotators are asked to write one or two utterances such that the last utterance will contradict some previous utterances in the conversation. In Step 2 (on the right), the annotators are asked to pick the utterances in the conversation that are involved in the contradiction. We use a casual term "message" instead of "utterance" in the instructions.

02. Solution

Quality Control

1. Onboarding Test : 모든 Annotator는 Onboarding Test를 통과해야한다. 테스트는 실제 DECODE Task 데이터이며, 5개의 대화 중 3개의 대화가 대화 흐름과 모순되는 대화를 포함한다. Annotator는 테스트를 통과하기 위해 5개의 모든 대화에 대해 올바른 레이블을 선택해야 한다. 이 방법은 Annotator가 작업을 이해하는지 여부를 테스트한다.
2. Maximum Annotation Count Limit : 한 Annotator가 생성할 수 있는 최대 예제 수는 20개로 제한한다. 이 방법은 유사한 패턴을 줄여 대화 예제를 더욱 다양화하는 데 도움된다.
3. Verification : 3명의 추가 Annotator에게 수집된 예제 중 일부를 확인하고 3명의 검증자가 모두 모순 레이블에 동의한 것을 선택하여 결과 검증 및 테스트 세트에 사용하도록 요청한다. 이 방법은 대화에 명확하고 합의된 모순이 있는지 확인하여 일부 NLU 작업에서 주관성과 모호성 문제를 방지한다.

03. Model

Unstructured Approach

$$\hat{y}_{pred} = f_\theta([u_0, u_1, u_2, ..., u_{n-1}], u_n) \qquad (1)$$

- Unstructured Approach에서는 single context를 형성하기 위해 대화0 부터 대화 n-1까지를 한 문장으로 연결한다. 그후, contradiction 발언의 확률을 추론하기 위해 문맥과 마지막 발언에 f를 적용한다.
- 모델에 대화 구조의 신호를 제공하기 위해, 발화를 연결할 때 각 발화 앞에 특별한 토큰을 삽입하여 해당 발화의 화자를 나타낸다. 해당 방법은 모델이 학습 중 대화의 기본 구조를 암묵적으로 학습 할 것 이라 가정한다.

03. Model

Structured Utterance-based Approach

$$\hat{y}_{pred} = \max \left\{ f_{\theta}^{UB}(u_i, u_n) : u_i \in \mathbf{S} \right\} \quad (2)$$

- 추론은 결정적으로 마지막 발화에 의존하기 때문에, 먼저 집합 S를 형성하기 위해 마지막 화자의 모든 발화를 선택한다. 그런 다음 세트의 모든 발화와 마지막 발화를 짝을 지어 하나씩 fUB로 입력한다. 최종 모순 확률은 모든 출력에 대한 최대값이다.
- 해당 접근법의 장점은 높은 모순 확률을 가진 쌍을 선택함으로써 결정에 대한 뒷받침 증거를 제공할 수 있다. 이것은 예측에 대한 설명을 제공할 뿐만 아니라 모델 자체를 진단하는 데도 도움이 될 수 있다.
- 이 모델링 접근법의 한 가지 단점은 화자 간의 추론을 포착할 수 없다는 것이다. 대화에는 대명사가 사용되는데 이러한 것들을 추론할 수 없다는 단점이 있다.

03. Model

Thresholding

- 모든 방법의 임계값은 0.5이며 크면 모순 작은면 합리적 대화로 분류된다.

04. Result

Pre-trained Model	Training Data	Main (Test)	Main (Test-Strict)	Human-Bot	SE (Precision / Recall / F1)
Unstructured Approach					
RoBERTa	All	**97.46**	-	77.09	-
	All - DNLI	97.44	-	73.17	-
	All - ANLI-R3	98.04	-	73.56	-
	All - DECODE	84.42	-	61.91	-
	DNLI	57.19	-	60.34	-
	ANLI-R3	82.21	-	59.69	-
	DECODE	96.85	-	70.03	-
Utterance-based Approach					
RoBERTa	SNLI + MNLI	77.40	47.70	73.17	63.3 / 84.6 / 72.4
	All	94.19	80.08	83.64	85.9 / **91.2** / **88.5**
	All - DNLI	94.38	**80.93**	81.68	86.7 / 90.1 / 88.4
	All - ANLI-R3	94.07	79.32	82.85	85.2 / 91.8 / 88.4
	All - DECODE	86.67	66.95	77.36	78.0 / 83.4 / 80.6
	DNLI	76.54	63.09	75.26	85.1 / 61.2 / 71.2
	ANLI-R3	81.59	69.11	70.52	**88.2** / 64.3 / 74.3
	DECODE	93.19	80.86	**84.69**	87.9 / 87.2 / 87.5
BERT	DECODE	88.88	74.14	75.52	84.9 / 83.7 / 84.3
Electra	DECODE	93.17	81.19	80.76	87.9 / 87.1 / 87.5
BART	DECODE	94.47	80.10	79.19	85.8 / 90.7 / 88.2
Majority					
-	-	50.00	50.00	50.00	50.4 / 47.1 / 48.7

04. Result

DECODE를 사용한 모델이 NLI 데이터를 사용한 모델보다 성능이 좋다.

Pre-trained Model	Training Data	Main (Test)	Main (Test-Strict)	Human-Bot	SE (Precision / Recall / F1)
Unstructured Approach					
RoBERTa	All	**97.46**	-	77.09	-
	All - DNLI	97.44	-	73.17	-
	All - ANLI-R3	98.04	-	73.56	-
	All - DECODE	84.42	-	61.91	-
	DNLI	57.19	-	60.34	-
	ANLI-R3	82.21	-	59.69	-
	DECODE	96.85	-	70.03	-
Utterance-based Approach					
RoBERTa	SNLI + MNLI	77.40	47.70	73.17	63.3 / 84.6 / 72.4
	All	94.19	80.08	83.64	85.9 / **91.2** / **88.5**
	All - DNLI	94.38	**80.93**	81.68	86.7 / 90.1 / 88.4
	All - ANLI-R3	94.07	79.32	82.85	85.2 / 91.8 / 88.4
	All - DECODE	86.67	66.95	77.36	78.0 / 83.4 / 80.6
	DNLI	76.54	63.09	75.26	85.1 / 61.2 / 71.2
	ANLI-R3	81.59	69.11	70.52	**88.2** / 64.3 / 74.3
	DECODE	93.19	80.86	**84.69**	87.9 / 87.2 / 87.5
BERT	DECODE	88.88	74.14	75.52	84.9 / 83.7 / 84.3
Electra	DECODE	93.17	81.19	80.76	87.9 / 87.1 / 87.5
BART	DECODE	94.47	80.10	79.19	85.8 / 90.7 / 88.2
Majority					
-	-	50.00	50.00	50.00	50.4 / 47.1 / 48.7

04. Result

다른 테스트 데이터와 다르게 Human-Bot 데이터에서 비교적 더 낮은 성능을 보인다.

Pre-trained Model	Training Data	Main (Test)	Main (Test-Strict)	Human-Bot	SE (Precision / Recall / F1)
Unstructured Approach					
RoBERTa	All	**97.46**	-	77.09	-
	All - DNLI	97.44	-	73.17	-
	All - ANLI-R3	98.04	-	73.56	-
	All - DECODE	84.42	-	61.91	-
	DNLI	57.19	-	60.34	-
	ANLI-R3	82.21	-	59.69	-
	DECODE	96.85	-	70.03	-
Utterance-based Approach					
RoBERTa	SNLI + MNLI	77.40		73.17	63.3 / 84.6 / 72.4
	All	94.19		83.64	85.9 / **91.2** / **88.5**
	All - DNLI	94.38		81.68	86.7 / 90.1 / 88.4
	All - ANLI-R3	94.07		82.85	85.2 / 91.8 / 88.4
	All - DECODE	86.67		77.36	78.0 / 83.4 / 80.6
	DNLI	76.54		75.26	85.1 / 61.2 / 71.2
	ANLI-R3	81.59		70.52	**88.2** / 64.3 / 74.3
	DECODE	93.19		**84.69**	87.9 / 87.2 / 87.5
BERT	DECODE	88.88		75.52	84.9 / 83.7 / 84.3
Electra	DECODE	93.17		80.76	87.9 / 87.1 / 87.5
BART	DECODE	94.47		79.19	85.8 / 90.7 / 88.2
Majority					
-	-	50.00	50.00	50.00	50.4 / 47.1 / 48.7

- RoBERTa, Electra, BART는 DECODE에서 약 93%-94%의 정확도를 얻었다.
- RoBERTa는 Humna-Bot에서 84.69%로 가장 높은 점수(다른 모델 75.52, 79.19, 80.76)를 얻었다.
- 이는 RoBERTa 사전 훈련 데이터가 Electra와 BART보다 범위가 넓기 때문으로 추측된다.

04. Result

Unstructured approach가 in-domain에서 더 좋은 성능을 보인다.

Pre-trained Model	Training Data	Main (Test)	Main (Test-Strict)	Human-Bot	SE (Precision / Recall / F1)
Unstructured Approach					
RoBERTa	All	**97.46**	-	77.09	-
	All - DNLI	97.44	-	73.17	-
	All - ANLI-R3	98.04	-	73.56	-
	All - DECODE	84.42	-	61.91	-
	DNLI	57.19	-	60.34	-
	ANLI-R3	82.21	-	59.69	-
	DECODE	96.85	-	70.03	-
Utterance-based Approach					
RoBERTa	SNLI + MNLI	77.40	47.70	73.17	63.3 / 84.6 / 72.4
	All	94.19	80.08	83.64	85.9 / **91.2** / **88.5**
	All - DNLI	94.38	**80.93**	81.68	86.7 / 90.1 / 88.4
	All - ANLI-R3	94.07	79.32	82.85	85.2 / 91.8 / 88.4
	All - DECODE	86.67	66.95	77.36	78.0 / 83.4 / 80.6
	DNLI	76.54	63.09	75.26	85.1 / 61.2 / 71.2
	ANLI-R3	81.59	69.11	70.52	**88.2** / 64.3 / 74.3
	DECODE	93.19	80.86	**84.69**	87.9 / 87.2 / 87.5
BERT	DECODE	88.88	74.14	75.52	84.9 / 83.7 / 84.3
Electra	DECODE	93.17	81.19	80.76	87.9 / 87.1 / 87.5
BART	DECODE	94.47	80.10	79.19	85.8 / 90.7 / 88.2
Majority					
-	-	50.00	50.00	50.00	50.4 / 47.1 / 48.7

- Unstructured 방법은 학습할수록 대화의 일관성을 잘 표현할 수 있다. 그러나 Human-Bot 성능 결과를 보면 이러한 높은 정확도는 모델의 실제 이해 능력을 과도하게 증폭시키는 것을 보여준다.

04. Result

Structured approach가 더 robust하고 transferable하다.

Pre-trained Model	Training Data	Main (Test)	Main (Test-Strict)	Human-Bot	SE (Precision / Recall / F1)
Unstructured Approach					
RoBERTa	All	**97.46**	-	77.09	-
	All - DNLI	97.44	-	73.17	-
	All - ANLI-R3	98.04	-	73.56	-
	All - DECODE	84.42	-	61.91	-
	DNLI	57.19	-	60.34	-
	ANLI-R3	82.21	-	59.69	-
	DECODE	96.85	-	70.03	-
Utterance-based Approach					
RoBERTa	SNLI + MNLI	77.40	47.70	73.17	63.3 / 84.6 / 72.4
	All	94.19	80.08	83.64	85.9 / **91.2** / **88.5**
	All - DNLI	94.38	**80.93**	81.68	86.7 / 90.1 / 88.4
	All - ANLI-R3	94.07	79.32	82.85	85.2 / 91.8 / 88.4
	All - DECODE	86.67	66.95	77.36	78.0 / 83.4 / 80.6
	DNLI	76.54	63.09	75.26	85.1 / 61.2 / 71.2
	ANLI-R3	81.59	69.11	70.52	**88.2** / 64.3 / 74.3
	DECODE	93.19	80.86	**84.69**	87.9 / 87.2 / 87.5
BERT	DECODE	88.88	74.14	75.52	84.9 / 83.7 / 84.3
Electra	DECODE	93.17	81.19	80.76	87.9 / 87.1 / 87.5
BART	DECODE	94.47	80.10	79.19	85.8 / 90.7 / 88.2
Majority					
-	-	50.00	50.00	50.00	50.4 / 47.1 / 48.7

- 현재 일부 실무자들은 표준 트랜스포머를 사용하면 모든 구조를 스스로 올바르게 학습할 수 있다고 믿는다. 하지만 해당 실험을 통해 그렇지 않다는 것을 확인했다.
- 해당 논문에서는 Sturctured 방법을 통해 해결방법은 제시했다.

05. Conclusion

- DECODE와 Human-Bot 같은 새로운 대화 데이터 세트를 소개한다. DECODE로 학습한 모델은 기존의 다른 NLI 데이터보다 높은 성능을 발휘한다.
- 모델에 입력되기 전에 각각의 대화가 다른 대화와 짝을 이루는 Structured 방법을 제안한다. Unstructured 방법과 비교하며 학습에 사용되지 않은 대화 데이터를 사용할 때 성능이 좋은 것을 확인했다. NLU 모듈을 NLG 시스템에 통합할 때 도메인 데이터가 부족한 경우가 많기 때문에 이는 중요한 특성이다.
- 추후 NLU와 NLG의 모델링과 두 가지 통합에 대한 상호 보완적인 진전을 예상한다.