DIALOGUE NATURAL LANGUAGE INFERENCE

Sean Welleck   JasonWeston   Arthur Szlam   Kyunghyun Cho

Motivation - 1

- Consistency of dialogue systems

Human: *what is your job ?*
Machine: *i 'm a lawyer .*
Human: *what do you do ?*
Machine: *i 'm a doctor .*

Semantic plausibility is not enough to root them out, preventing them is challenging.

Previous work

- Personalizing Dialogue Agents: I have a dog, do you have pets too?
  - The dialogue agent was given a set of personal facts describing its character (a persona)
  - Intended outcome: utterances that consistent with its given persona.
  - However, consistency issue still exists.

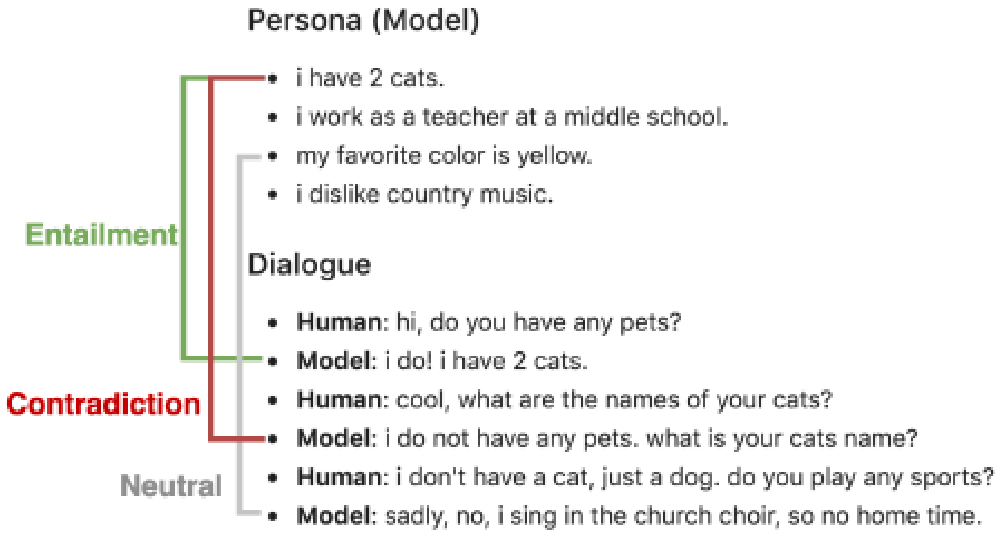


Figure 1: Persona-based dialogue with a Key-Value Memory Network trained on Persona-Chat [21].

Motivation - 2

- **Natural Language Inference task**
  - Learning a mapping between a sentence pair and a category *{Entail, Neutral, Contradict}*
  - Expected to be useful in downstream tasks.

Contributions

- Leveraging an NLI model to reduce the problem pf consistency in dialogue.
- Create a dataset, Dialogue NLI: contains sentence pairs labeled as entailment, neutral, or contradiction.
- Improve consistency of dialogue models.

Problem Formulation

- **Dialogue Generation**
  - An alternating two-agent dialogue with agent A and ends with agent B is written as: $u_1^A, u_2^B, u_3^A, u_4^B, \ldots, u_T^B$
- **Persona-based dialogue**
  - Each agent is associated with a persona, P_A and P_B
  - Persona is represented by a set of utterances $P = \{p_1, \ldots, p_m\}$
- **Consistency**
- **Natural Language inference**

$$\mathcal{D} = \{(s_1, s_2)_i, y_i\}_{i=1}^{N}$$

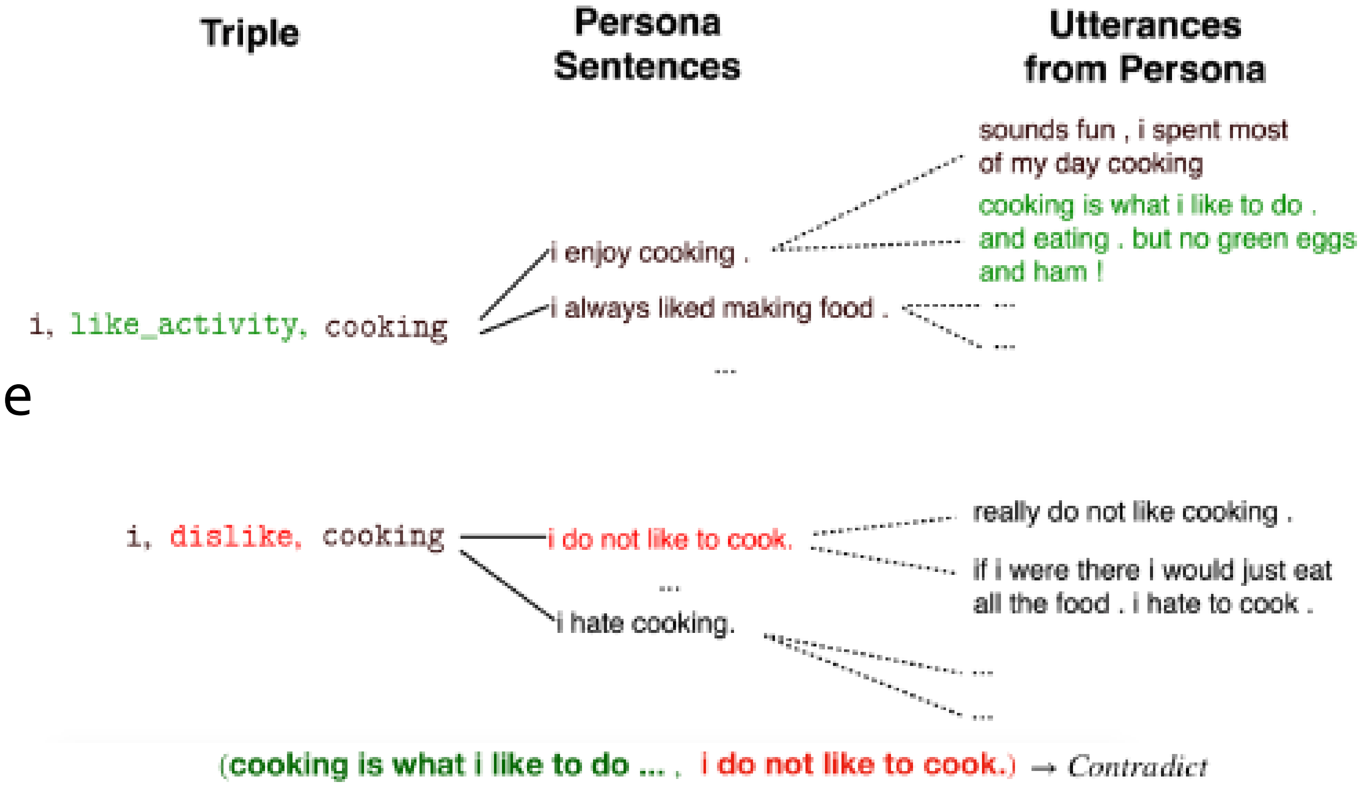
$$f_{\text{NLI}}(s_1, s_2) \rightarrow \{E, N, C\}$$

Problem Formulation

- **Reducing dialogue consistency to NLI**
  - Given a persona $P_A = \{p_1^A, \dots, p_m^A\}$ for agent A and a length T dialogue $u_1^A, u_2^B, \dots, u_{T-1}^A, u_T^B$.
  - Dialogue contradiction is contained in pair (u_i^A, u_j^A)
  - Persona contradiction is contained in a pair (u_i^A, p_k^A)

Dialogue NLI Dataset

- Sentences
  - Consists of $(\mathrm{u_i}, p_j)$ and (p_i, p_j) pairs
- Labels
  - Associate a human-labeled triple (e_1, r, e_2) with each persona sentence and a subset of utterances.



Triples Annotation

- Each **persona sentence** is annotated through a Mechanical Turk task.
  - 10832 persona sentences are annotated
- **Utterances**
  - Let p be a persona sentence with triple (e_1, r, e_2)
  - If e_2 is a sub-string of u, or word similarity $\mathrm{sim}(u, p) \geq \tau$,
    - u is associated with triple (e_1, r, e_2) and persona p

Dialogue NLI Dataset

- Pairs $(\mathrm{u_i}, p_j)$ and (p_i, p_j) are defined as entailment , neutral or contradiction based on their triple.
- **Entailment**: share the same triple
- **Neutral**
  - Miscellaneous utterance (u, p)
  - Persona pairing (p, p')
  - Relation swap (r, r'), e.g. *have_vehicle* and *have_pet*
- **Contradiction**
  - Relation swap, e.g. *like_activity* and *dislike.*
  - Entity swap, e.g. *physical_attribute, short -> tall*
  - Numeric contradiction

Dialogue NLI Dataset

Triple	Premise	Hypothesis	Triple	Label
(i, like_activity, chess)	i listen to a bit of everything . it helps me focus for my chess tournaments .	i like to play chess .	(i, like_activity, chess)	E
-	how are you today?	i drink espresso .	(i, like_drink, espresso)	N
(i, like_goto, spain)	i love spain so much , i been there 6 times .	i think i will retire in a few years .	(i, want_do, retire)	N
(i, have_vehicle, car)	my vehicle is older model car .	i have pets .	(i, have_pet, pets)	N
(i, dislike, cooking)	i really do not enjoy preparing food for myself .	i like to cook with food i grow in my garden .	(i, like_activity, cooking)	C
(i, physical_attribute, short)	height is missing from my stature .	i am 7 foot tall .	(i, physical_attribute, tall)	C
(i, have_family, 3 sister)	i have a brother and 3 sisters .	i have a brother and four sisters .	(i, have_family, 4 sister)	C

Dialogue NLI Dataset

Data Type	Label	Train (u,p)	Train (p,p)	Valid (u,p)	Valid (p,p)	Test (u,p)	Test (p,p)
Matching Triple	E	43,000	57,000	5,000	500	4,500	900
Misc. Utterance	N	50,000	-	3,350	-	3,000	-
Persona Pairing	N	20,000	10,000	2,000	-	2,000	-
Relation Swap	N	20,000	-	150	-	400	-
Relation Swap	C	19,116	2,600	85	14	422	50
Entity Swap	C	47,194	31,200	4,069	832	3,400	828
Numerics	C	10,000	-	500	-	1,000	-
Dialogue NLI Overall		**310,110**		**16,500**		**16,500**	

Consistent Dialogue Agent via NLI

- Assume a dialogue model and a Dialogue NLI model

$$f^{\text{dialogue}}(P, u_{<t}, U) \quad \rightarrow \quad (s_1, s_2, ..., s_{|U|})$$

$$f^{\text{NLI}}(u, p) \rightarrow \{E, N, C\}$$

- NLI model run on each $(\mathrm{u_i}, p_j)$ pair, predicting a label $y_{i,j} \in \{E, N, C\}$, with confidence $c_{i,j}$

$$s_i^{\text{contradict}} = \begin{cases} 0 & \text{if } y_{i,j} \neq C \; \forall \; p_j \in P \\ \max\limits_{j:y_{i,j}=C} c_{i,j} & \text{otherwise.} \end{cases}$$

- New candidate scores

$$s_i^{\text{re-rank}} = s_i - \lambda(s_1 - s_k) s_i^{\text{contradict}}$$

Experiments - NLI

- NLI models that have achieved competitive performance on existing NLI benchmark datasets

Model	Valid	Test
ESIM	**86.31**	**88.20**
InferSent	85.82	85.68
InferSent SNLI	47.86	46.36
InferSent Hypothesis-Only	55.98	57.19
Most Common Class	33.33	34.54
ESIM Ground-Truth Triples	99.53	99.49

Table 3: Dialogue NLI Results

Experiments – Consistency in Dialogue

- Key-Value Memory Network
  - Trained on the persona-chat dataset
- Evaluation sets

Persona (Model)

- i work in retail .
- i enjoy singers like jason aldea .
- i love country music .
- i have an economical suv .

Dialogue

- **Model:** hello ! do you like the new song by taylor swift ?
- **Human:** even though i have lived on earth for 100 years , i have not heard anything better .

Next-Utterance Candidates:

KVMemnn Score	Original	Re-ranked
0.261	**yes . i do not like country though .**	do you like country music ?
0.203	**i hate country music . you ?**	i really like country . do you have any pets ?
0.185	do you like country music ?	cool , what is your favorite type of music ? mine is country .
0.149	i really like country . do you have any pets ?	my favorite type of music is country .
0.142	cool , what is your favorite type of music ? mine is country .	cool i love country music sone songs are in spanish .

NLI Model Output:

Candidate	Persona Sentence Labeled as Contradiction	Confidence
yes . i do not like country though .	i love country music .	1.000
i hate country music . you ?	i enjoy singers like jason aldea .	0.986
	i love country music .	1.000

Experiments – Consistency in Dialogue

- Metrics
  - Hits@k
  - Contradict@k
    - Measures the proportion of top-k candidates which are contradicting
  - Entail@k
    - Measures the proportion of top-k candidates which are entailment.
- Results

	Haves		Likes		Attributes	
	Orig.	**Rerank**	**Orig.**	**Rerank**	**Orig.**	**Rerank**
Hits@1 ↑	30.2	**37.3**	16.9	**18.7**	35.2	**36.4**
Contradict@1 ↓	32.5	**8.96**	17.6	**4.1**	8.0	**5.7**
Entail@1 ↑	55.2	**74.6**	77.9	**90.6**	87.5	**88.6**

Human evaluation

- Scoring
  - **Overall score:** how well the model represented its persona {1, 2, 3, 4, 5}
  - **Consistent:** a marking of each model utterance consistent with the models persona {0, 1}
  - **Contradiction:** a marking of each model utterance that contradicted a previous utterance or model's persona {0,1}

	Overall Score ↑		% Consistent ↑		% Contradiction ↓	
	Raw	Calibrated	Raw	Calibrated	Raw	Calibrated
KV-Mem	2.11± 1.12	2.21± 0.26	0.24	0.27± 0.07	0.23	0.25± 0.08
KV-Mem + NLI	**2.34± 1.21**	**2.38± 0.26**	**0.28**	**0.35± 0.08**	**0.19**	**0.16± 0.06**