Dialogue Natural Language Inference

Sean Welleck
New York University
wellecks@nyu.edu

Jason Weston
Facebook AI Research
New York University

Arthur Szlam
Facebook AI Research

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar

Abstract

Consistency is a long standing issue faced by dialogue models. In this paper, we frame the consistency of dialogue agents as natural language inference (NLI) and create a new natural language inference dataset called Dialogue NLI. We propose a method which demonstrates that a model trained on Dialogue NLI can be used to improve the consistency of a dialogue model, and evaluate the method with human evaluation and with automatic metrics on a suite of evaluation sets designed to measure a dialogue model's consistency.

1 Introduction

A long standing issue faced by dialogue models is *consistency* (Li et al., 2016; Vinyals et al., 2015; Zhang et al., 2018). An example from (Vinyals et al., 2015) shows a two-round dialogue in which their neural sequence model first responds to *what is your job?* with *i'm a lawyer*, then responds to *what do you do?* with *i'm a doctor*. Even when inconsistencies are relatively rare and semantically plausible, they are jarring, and because semantic plausibility is not enough to root them out, preventing them is challenging.

One approach to increasing the consistency of a chit-chat dialogue model was proposed in (Zhang et al., 2018), where the dialogue agent was given a set of personal facts describing its character (a *persona*) and produces utterances that reflect the persona. The intended outcome is that the agent produces utterances consistent with its given persona. However, these models still face the consistency issue, as shown in Figure 1.

Separately, the framework of Natural Language Inference (NLI) (Bowman et al., 2015; Dagan et al., 2006; Maccartney and Manning, 2009) involves learning a mapping between a sentence pair and an entailment category. It is hypothesized that the NLI task is a proxy for general goals in natural language processing, such as language understanding (Bowman et al., 2015; Williams et al., 2018). Thus, the NLI task has been used for learning general sentence representations (Conneau et al., 2017) and for evaluating NLP models (Poliak et al., 2018a; Wang et al., 2018), with the expectation that such models will be useful in downstream tasks.

Despite this expectation, leveraging an NLI model for a downstream task remains an under-explored research direction. An NLI model may improve downstream task performance if properly used, while downstream tasks may yield new datasets or identify issues with existing NLI models, thus expanding the NLI research domain.

In this paper, we reduce the problem of consistency in dialogue to natural language inference. We first create a dataset, Dialogue NLI,[1] which contains sentence pairs labeled as entailment, neutral, or contradiction.

Then, we demonstrate that NLI can be used to improve the consistency of dialogue models using a simple method where utterances are re-ranked using a NLI model trained on Dialogue NLI. The method results in fewer persona contradictions on three evaluation sets. The evaluation sets can be used independently to automatically evaluate a dialogue model's persona consistency, reducing the need for human evaluation. We discuss several future research directions involving this approach.

2 Dialogue Consistency and Natural Language Inference

First, we review the dialogue generation and natural language inference problems as well as the notions of consistency used throughout.

[1] The dataset is available at wellecks.github.io/dialogue_nli.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3731–3741
Florence, Italy, July 28 - August 2, 2019.

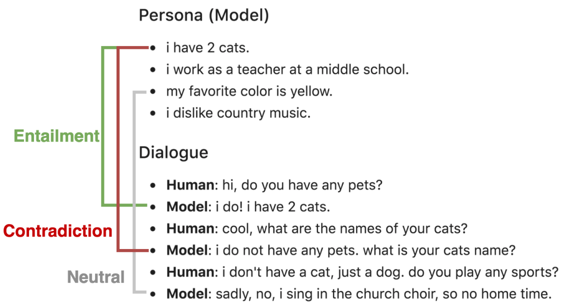

Figure 1: Persona-based dialogue with a Key-Value Memory Network trained on Persona-Chat (Zhang et al., 2018).

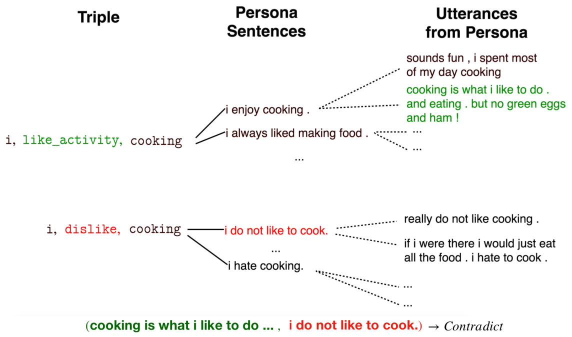

Figure 2: Relating triples, persona sentences, and utterances to derive annotated sentence pairs. Shown here is a "relation swap" contradiction.

Dialogue Generation Dialogue generation can be framed as *next utterance prediction*, in which an utterance (a sequence of tokens representing a sentence) u_{t+1} is predicted given a conversation prefix $u_{\leq t}$. A sequence of utterances is interpreted as a *dialogue* between *agents*. For instance, an alternating two-agent dialogue which starts with agent A and ends with agent B is written as $u_1^A, u_2^B, u_3^A, u_4^B, ..., u_T^B$.

Persona-Based Dialogue In *persona-based dialogue*, each agent is associated with a persona, P_A and P_B. An utterance is now predicted using the conversation prefix $u_{\leq t}$ *and* the agents own persona, e.g. P_A for agent A. It is assumed that an agent's utterances are conditionally dependent on its persona, which can be interpreted as the utterances being representative of, or reflecting, the persona.

A typical approach for representing the persona is to use a set of sentences $P = \{p_1, ..., p_m\}$.

Consistency A *consistency error*, or contradiction, occurs when an agent produces an utterance that contradicts one of their previous utterances. Similarly, a *persona consistency error*, or persona contradiction, occurs when an agent produces an utterance that contradicts a subset of its persona.

A contradiction may be a clear logical contradiction, e.g. *I have a dog* vs. *I do not have a dog*, but in general is less clearly defined. As a result, in addition to logical contradictions, we interpret a consistency error as being two utterances not likely to be said by the same persona. For instance, "i'm looking forward to going to the basketball game this weekend!" vs. "i don't like attending sporting events", as well as "i'm a lawyer" vs. "i'm a doctor" would be viewed here as contradictions, although they are not strict logical inconsistencies.

Similarly, a persona consistency error is interpreted here as an utterance which is not likely to be said given a persona described by a given set of persona sentences, in addition to logical contradictions.

Natural Language Inference Natural Language Inference (NLI) assumes a dataset $\mathcal{D} = \{(s_1, s_2)_i, y_i\}_{i=1}^N$ which associates an input pair (s_1, s_2) to one of three classes $y \in \{\text{entailment}, \text{neutral}, \text{contradiction}\}$. Each input item s_j comes from an input space $\mathcal{S}_j$, which in typical NLI tasks is the space of natural language sentences, i.e. s_j is a sequence of words $(w_1, ..., w_K)$ where each word w_k is from a vocabulary $\mathcal{V}$.

The input (s_1, s_2) are referred to as the *premise* and *hypothesis*, respectively, and each label is interpreted as meaning the premise *entails* the hypothesis, the premise is *neutral* with respect to the hypothesis, or the premise *contradicts* the hypothesis. The problem is to learn a function $f_{\text{NLI}}(s_1, s_2) \rightarrow \{E, N, C\}$ which generalizes to new input pairs.

Reducing Dialogue Consistency to NLI Identifying utterances which contradict previous utterances or an agent's persona can be reduced to natural language inference by assuming that contradictions are contained in a sentence pair. That is, given a persona $P_A = \{p_1^A, ..., p_m^A\}$ for agent A and a length-T dialogue $u_1^A, u_2^B, ...u_{T-1}^A, u_T^B$, it is assumed that a dialogue contradiction for agent A is contained in an utterance pair (u_i^A, u_j^A), and a persona contradiction is contained in a pair (u_i^A, p_k^A). Similarly, we assume that entailments and neutral interactions, defined in Section 3, are contained in sentence pairs. We do not consider relationships which require more than two sentences to express.

Under this assumption, we can use a natural language inference model f_{NLI} to identify entailing, neutral, or contradicting utterances.

Section 3 proposes a dialogue-derived dataset for training f_{NLI}, and Section 4 proposes a method which incorporates f_{NLI} with a dialogue model for next utterance prediction.

3 Dialogue NLI Dataset

The Dialogue NLI dataset consists of sentence pairs labeled as entailment (E), neutral (N), or contradiction (C).

Sentences Sentences originate from a two-agent persona-based dialogue dataset. A dialogue between agents A and B consists of a sequence of utterances $u_1^A, u_2^B, u_3^A, u_4^B, ..., u_T^B$, and each agent has a persona represented by a set of persona sentences $\{p_1^A, ..., p_{m_A}^A\}$ and $\{p_1^B, ..., p_{m_B}^B\}$. The Dialogue NLI dataset consists of (u_i, p_j) and (p_i, p_j) pairs[2] from the Persona-Chat dataset (Zhang et al., 2018)[3].

3.1 Triple Generation

In order to determine labels for our dataset, we require human annotation of the utterances and persona sentences in PersonaChat, as the original dataset does not contain this information. We perform such annotation by first associating a human-labeled *triple* (e_1, r, e_2) with each persona sentence, and a subset of all the utterances, detailed in 3.2. Each triple contains the main fact conveyed by a persona sentence, such as `(i, have_pet, dog)` for the persona sentence *I have a pet dog*, or a fact mentioned in an utterance, such as *No, but my dog sometimes does*.

Persona sentences and utterances are grouped by their triple (e.g. see Figure 2), and pairs (u, p) and (p, p) are defined as entailment, neutral, or contradiction based on their triple according to the criteria below. For examples and summary, we refer readers to Tables 1–2.

[2] We also release additional (u_i, u_j) pairs, but experiments in this paper are not based on them.

[3] The dataset collection process is applicable to other persona-based dialogue datasets such as (Mazaré et al., 2018).

Entailment Each unique pair of sentences that share the same triple are labeled as entailment.

Neutral Neutral pairs are obtained with three different methods.

First, a *miscellaneous utterance* is a (u, p) pair of which u is not associated with any triple. This includes greetings (*how are you today?*) and sentences unrelated to a persona sentence (*the weather is ok today*), so such utterances are assumed to be neutral with respect to persona sentences.

The second method, *persona pairing*, takes advantage of the fact that each ground-truth persona is typically neither redundant nor contradictory. A persona sentence pair (p, p') is first selected from a persona if p and p' do not share the same triple. Then each sentence associated with the same triple as p is paired with each sentence associated with the same triple as p'.

Lastly, we specify *relation swaps* (r, r') for certain relations (see Appendix A.2) whose triples are assumed to represent independent facts, such as `have_vehicle` and `have_pet`. A sentence pair, whose first sentence is associated with a triple $(\cdot, r, \cdot)$ and whose second sentence has triple $(\cdot, r', \cdot)$, is labeled as neutral. See Table 1 for an example.

Contradiction We obtain contradictions using three methods. See Figure 2 for an example.

First, the *relation swap* method is used by specifying contradicting relation pairs (r, r') (see Appendix A.2), such as `(like_activity,dislike)`, then pairing each sentence associated with the triple (e_1, r, e_2) with each sentence associated with (e_1, r', e_2).

Similarly, an *entity swap* consists of specifying relations, e.g., `physical_attribute`, that would yield a contradiction when the value of e_2 is changed to a different value e_2', e.g., `short` $\rightarrow$ `tall` (see Appendix A.3). Sentences associated with (e_1, r, e_2) are then paired with sentences associated with (e_1, r, e_2').

Finally, a *numeric* contradiction is obtained by first selecting a sentence which contains a number that appears in the associated triple (see Table 1). A contradicting sentence is generated by replacing the sentence's numeric surface form with a different randomly sampled integer in the number or text form.

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

Table 1: Examples from the validation set.

		Train		**Valid**		**Test**		**Test-Gold**	
Data Type	**Label**	(u,p)	(p,p)	(u,p)	(p,p)	(u,p)	(p,p)	(u,p)	(p,p)
Matching Triple	E	43,000	57,000	5,000	500	4,500	900	3,712	615
Misc. Utterance	N	50,000	-	3,350	-	3,000	-	2,282	-
Persona Pairing	N	20,000	10,000	2,000	-	2,000	-	1,466	-
Relation Swap	N	20,000	-	150	-	400	-	260	-
Relation Swap	C	19,116	2,600	85	14	422	50	279	44
Entity Swap	C	47,194	31,200	4,069	832	3,400	828	2,246	591
Numerics	C	10,000	-	500	-	1,000	-	881	-
Dialogue NLI Overall		**310,110**		**16,500**		**16,500**		**12,376**	

Table 2: Dialogue NLI Dataset Properties. (u, p) and (p, p) refer to (utterance, persona sentence) and (persona sentence, persona sentence) pairs, respectively. Numerics consist of (u, u) (u, p) and (p, p) pairs.

3.2 Triple Annotation

Each persona sentence is annotated with a triple (e_1, r, e_2) using Amazon Mechanical Turk task. We first define a schema consisting of $\langle category\rangle\langle relation\rangle\langle category\rangle$ rules, such as $\langle person\rangle have\_pet\langle animal\rangle$, where the relation comes from a fixed set of relation types $\mathcal{R}$, listed in Appendix A.1. Given a sentence, the annotator selects a relation r from a drop-down populated with the values in $\mathcal{R}$. The annotator then selects the categories and values of the entities e_1 and e_2 using drop-downs that are populated based on the schema rules. An optional drop-down contains numeric values for annotating entity quantities (e.g., 3 brothers). If selected, the numeric value is concatenated to the front of the entity value. The annotator can alternatively input an out-of-schema entity value in a text-box. Using this method, each of the 10,832 persona sentences is annotated with a triple (e_1, r, e_2), where $r \in \mathcal{R}$, $e_1 \in \mathcal{E}_1$, and $e_2 \in \mathcal{E}_2$. Here $\mathcal{E}_1$ is the set of all annotated e_1 from the drop-downs or the text-box, and $\mathcal{E}_2$ is similarly defined.

Finally, *utterances* are associated with a triple as follows. Let p be a persona sentence with triple (e_1, r, e_2). We start with all utterances, U, from agents that have p in their persona. An utterance $u \in U$ is then associated with the triple (e_1, r, e_2) and persona sentence p when e_2 is a sub-string of u, or word similarity[4] $\text{sim}(u, p) \geq \tau$ is suitably large.

[4] We use cosine similarity between the mean of TF-IDF weighted GloVe (Pennington et al., 2014) word vectors and set $\tau = 0.9$.

3.3 Statistics

Table 2 summarizes the dataset and its underlying data types. The label, triple, and data type are supplied as annotations for each sentence pair. We additionally create a gold-standard test set (*Test Gold*) by crowdsourcing three label annotations for each example in the test set. We keep each test example for which two or more annotators agreed with its dataset label. All sentences in Dialogue NLI were generated by humans during the crowdsourced dialogue collection process of the Persona-Chat dataset (Zhang et al., 2018). The resulting sentence pairs are thus drawn from a natural dialogue domain that differs from existing NLI datasets, which are either drawn from different domains such as image captions or created using synthetic templates (Bowman et al., 2015; Demszky et al., 2018; Khot et al., 2018; Marelli et al., 2014; Poliak et al., 2018b; Wang et al., 2018; Williams et al., 2018).

4 Consistent Dialogue Agents via Natural Language Inference

We now present a method which demonstrates that natural language inference can be used to improve the consistency of dialogue agents. Candidate utterances are re-ranked based on whether the candidate is predicted to contradict a persona sentence. If the NLI model predicts that a candidate contradicts a persona sentence, the candidate's score is penalized, with the penalty weighted by the NLI model's confidence[5] scaled by a constant.

Specifically, assume a dialogue model $f^{\text{dialogue}}(P, u_{\leq t}, U) \rightarrow (s_1, s_2, ..., s_{|U|})$ and a Dialogue NLI model $f^{\text{NLI}}(u, p) \rightarrow \{E, N, C\}$. Given a persona $P = \{p_1, ..., p_m\}$, previous utterances $u_{\leq t}$, and a set of candidate next-utterances U, the dialogue model outputs a ranked list of scores $s_1, s_2, ..., s_{|U|}$ corresponding to next-utterance candidates $u_1, u_2, ..., u_{|U|}$.

The NLI model is then run on each (u_i, p_j) pair, predicting a label $y_{i,j} \in \{E, N, C\}$ with confidence $c_{i,j}$. A contradiction score is computed for each candidate as:

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

That is, if the candidate u_i does not contradict any persona sentence p_j according to the NLI model, $s_i^{\text{contradict}}$ is zero. If u_i contradicts one or more persona sentences, $s_i^{\text{contradict}}$ is the highest confidence, $c_{i,j}$, out of the contradicting (u_i, p_j).[6]

New candidate scores are then computed as

$$s_i^{\text{re-rank}} = s_i - \lambda(s_1 - s_k)s_i^{\text{contradict}} \quad (1)$$

and the candidates are sorted according to $s^{\text{re-rank}}$. Hyper-parameters λ and k control the NLI model's influence in re-ranking. For example, if the top candidate has a contradiction score of 1.0, then with $\lambda = 1$, it will be moved to the k'th position in the ranking. $\lambda = 0$ corresponds to no re-ranking.

Model	Valid	Test	Test Gold
ESIM	**86.31**	**88.20**	**92.45**
InferSent	85.82	85.68	89.96
InferSent SNLI	47.86	46.36	47.03
InferSent Hyp. Only	55.98	57.19	51.52
Most Common Class	33.33	34.54	34.96
ESIM Gold Triples	99.52	99.46	99.69

Table 3: Dialogue NLI Results

5 Experiments

5.1 Experiment 1: NLI

Models Many recently proposed NLI models can be categorized into sentence encoding based methods of the form $f_{\text{MLP}}(g_{\text{enc}}(s_1), g_{\text{enc}}(s_2))$, and attention-based methods of the form $f_{\text{MLP}}(g_{\text{attn}}(s_1, s_2))$ (Lan and Xu, 2018). We thus choose and train representative models of each type which have achieved competitive performance on existing NLI benchmark datasets. For the sentence encoding method, we use InferSent (Conneau et al., 2017), which encodes a sentence using a bidirectional LSTM followed by max-pooling over the output states. As the representative attention-based method we use the enhanced sequential inference model (ESIM, (Chen et al., 2017)), which computes an attention score for each word pair.

We also report results from a model trained and evaluated using the hypothesis sentence only (InferSent Hyp. Only) (Gururangan et al., 2018; Poliak et al., 2018c), a model trained on the existing SNLI dataset (Bowman et al., 2015) but evaluated

[5] In our experiments, the softmax output corresponding to the contradiction class from Dialogue NLI.

[6] Future work could consider filtering previous-utterance contradictions (u_i, u_j) as well.

Data Type	Example	Pred.	Actual
Matching Triple (p, p)	i am a hopeless bookworm. when i have some spare time i read.	Neutral	Entail
Matching Triple (u, p)	i am from italy. i love the early mornings. i like getting up bright and early.	Neutral	Entail
Misc. Utterance	i do not understand football or baseball. i am employed as an engineer.	Contradict	Neutral
Persona Pairing	i lift weights every chance i get. i work in a warehouse driving a forklift.	Entail	Neutral
Relation Swap (p, p)	canines make me shake with fear. i love dogs but hate cats.	Entail	Contradict
Relation Swap (u, p)	i am heavy into fitness although i am rather large. i do not like exercise or physical activity.	Entail	Contradict
Entity Swap (p, p)	hawaii is where i reside. i do not drive because i live in new york.	Neutral	Contradict
Entity Swap (u, p)	tell me it was vegan food please , that is all i eat. i eat ham.	Neutral	Contradict
Numerics	i have two part time jobs. i have 7 part time jobs.	Neutral	Contradict

Table 4: Example ESIM mispredictions by data type on Test Gold.

Data Type	N	Accuracy
Matching Triple (p, p)	615	83.58
Matching Triple (u, p)	3,712	91.25
Misc. Utterance	2,282	96.85
Persona Pairing	1,466	94.48
Relation Swap (p, p)	44	79.55
Relation Swap (u, p)	539	80.71
Entity Swap (p, p)	591	93.40
Entity Swap (u, p)	2,246	92.43
Numerics	881	96.25

Table 5: ESIM Accuracy by data type on Test Gold.

on Dialogue NLI (InferSent SNLI), and a model which returns the most common class from the Dialogue NLI training set (Most Common Class).

Results Table 3 shows the performance of the two NLI models and three baselines on the Dialogue NLI validation and test sets. The test performance of ESIM (88.2%) and InferSent (85.68%) is similar to the performance reported on the existing SNLI dataset (88.0% (Chen et al., 2017) and 85.5% (Conneau et al., 2017), respectively), while the results on the Dialogue NLI gold test set (92.45%, 89.96%) are higher. As in Table 3, however, an InferSent model trained on SNLI performs poorly when evaluated on the proposed Dialogue NLI (47.03%). This is likely due to a mismatch in sentence distributions between SNLI, which is derived from image captions, and Dialogue NLI, whose sentences more closely resemble downstream dialogue applications. The hypothesis-only performance (51.52%) is lower than the hypothesis-only baseline for SNLI (69.00% (Poliak et al., 2018c)), and shows that using information from both the utterance and persona sentence is necessary to achieve good performance on Dialogue NLI.

ESIM's reasonably strong performance on Dialogue NLI suggests that the model may be useful in a downstream task - a claim which we verify in Experiment 5.1. However, there is also room for improvement. In particular, we report the performance of a model which takes the ground-truth triples as input instead of sentences. As shown in the last row of Table 3, each sentence's underlying triple contains sufficient information to achieve near-perfect accuracy (99.69%). We also show ESIM's accuracy by data type on Test Gold in Table 5, along with example mispredictions in Table 4. The accuracies and examples suggest that the NLI model could be improved further.

5.2 Experiment 2: Consistency in Dialogue

This experiment evaluates the effect of the re-ranking method from Section 4 on the dialogue model's persona consistency.

Experiment Setup The re-ranking method of Section 4 uses a dialogue next utterance prediction

	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**

Table 6: Effect of NLI re-ranking on persona consistency in dialogue. The reported metrics are percentages computed over each validation set.

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

Figure 3: Example from the Likes Evaluation Set, showing dialogue model candidates, NLI model predictions, and reranked candidates using the method proposed in Section 4.

model and the Dialogue NLI model.

For the dialogue model we train a key-value memory network (Zhang et al., 2018) on the Persona-Chat dataset, which uses persona sentences and the conversation prefix as context. This model achieved the best performance on Persona-Chat in (Zhang et al., 2018). We train the model using ParlAI (Miller et al., 2017) on the `personachat:self_original` task, using the hyper-parameters given for the `KVMemnnAgent` in the ConvAI2 competition. For the NLI model we use the ESIM model trained on Dialogue NLI, based on the results of Experiment 5.

To study the effect of re-ranking on persona consistency, we form evaluation sets which contain next-utterances which are likely to yield persona contradiction or entailment, as follows.

Evaluation Sets Each example is formed by first finding a next-utterance u_{t+1} in the Persona-Chat validation set which has an associated triple (e_1, r, e_2) of interest, e.g. $(\texttt{i}, \texttt{like\_music}, \texttt{country})$. If a sentence in the agent's profile P has triple (e_1, r, e_2), we form the validation example $(P, u_{\leq t}, u_{t+1})$. Figure 3 shows an example.

Each example is associated with candidates U, consisting of the ground-truth utterance u_{t+1}, 10 entailment candidates with the same triple as u_{t+1}, 10 contradicting candidates with a different triple than that of u_{t+1}, and 10 random candidates. The dialogue model must avoid ranking a contradicting candidate highly.

Specifically, suppose the ground-truth next-utterance u_{t+1} is associated with triple (e_1, r, e_2), e.g., $(\texttt{i}, \texttt{have\_pet}, \texttt{dog})$. Entailment candidates are utterances u from the validation or training sets such that u is associated with triple (e_1, r, e_2). Since by construction a sentence in the profile also has triple (e_1, r, e_2), these candidates entail a profile sentence. A contradicting candidate is an utterance associated with a specified contradicting triple (e'_1, r', e'_2), e.g., $(\texttt{i}, \texttt{not\_have}, \texttt{dog})$.

We construct three evaluation sets, **Haves**, **Likes**, and **Attributes** using this process.

Metrics We introduce variants of the ranking metric Hits@k, called Contradict@k and Entail@k. **Contradict@k** measures the proportion of top-k candidates returned by the model which contradict candidates, averaged over examples. This measures the propensity of a model to highly rank contradictions. Contradiction@1 is the proportion of consistency errors made by the model. For this metric lower values are better, in contrast to Hits@k.

Entail@k measures the proportion of top-k candidates returned by the model which are entailment candidates, averaged over examples. Entail-

	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**

Table 7: Human evaluation results (mean± standard deviation).

ment candidates share the same underlying triple as the ground-truth next utterance, so this metric rewards highly ranked candidates that convey similar meaning and logic to the ground-truth utterance. Thus it can be interpreted as a more permissive version of Hits@k.

Results Table 6 shows re-ranking results on the three evaluation sets ($\lambda = 1.0, k = 10$). The NLI re-ranking improves all three metrics on all the evaluation sets. Overall dialogue performance improves, as measured by Hits@1. The NLI re-ranking substantially reduces the number of contradicting utterances predicted by the model, and increases the number of utterances which entail a profile sentence, as seen in the Contradict@1 and Entail@1 scores.

Figure 3 shows an example dialogue with candidates, contradictions predicted by the NLI model, and the corresponding re-ranked candidates.

5.3 Experiment 3: Human Evaluation

This experiment evaluates the effect of the proposed NLI re-ranking method on a dialogue model's consistency, where consistency is judged by human annotators in an interactive persona-based dialogue setting.

Experiment Setup We use ParlAI (Miller et al., 2017) which integrates with Amazon Mechanical Turk for human evaluation. A human annotator is paired with a model, and each is randomly assigned a persona from 1,155 persona sets. The human and model are then asked to make a conversation of at least either five or six turns (randomly decided). After the conversation, the annotator assigns three scores to the conversation, described below. Each annotator is allowed to participate in at most ten conversations per model, and we collect 100 conversations per model. Two models are evaluated: the same key-value memory network used in Experiment 5.1 without re-ranking (**KV-Mem**), and with re-ranking (**KV-Mem + NLI**).

Scoring and Calibration Following a conversation, an annotator is shown the conversation and the model's persona, and assigns three scores: an overall score of how well the model represented its persona ({1,2,3,4,5}), a marking of each model utterance that was consistent with the model's persona ({0,1}), and a marking of each model utterance that contradicted a previous utterance or the model's persona ({0,1}).

We use Bayesian calibration to adjust for annotator bias, following (Kulikov et al., 2018). We assume a model with observed scores S_{ij} and latent variables M_i for the unobserved score of model i and B_j for the bias of annotator j. We then estimate the posterior mean and variance for the unobserved scores given the observed scores. We use Pyro (Bingham et al., 2018) and the no-u-turn sampler (Hoffman and Gelman, 2014) for posterior inference. See Appendix C for details.

Results Table 7 shows the human evaluation results. The natural language inference re-ranking improves all the metrics, notably the fine-grained consistency score (0.27 vs. 0.35) and contradiction score (0.25 vs. 0.16). The results are consistent with the conclusions from the automatic evaluation in Experiment 5.1.

6 Conclusion

In this paper, we demonstrated that natural language inference can be used to improve performance on a downstream dialogue task. To do so, we created a new dialogue-derived dataset called Dialogue NLI, a re-ranking method for incorporating a Dialogue NLI model into a dialogue task, and an evaluation set which measures a model's persona consistency. The dataset offers a new domain for natural language inference models, and suggests avenues such as devising alternative methods for using natural language inference components in downstream tasks. Future work may also incorporate contradiction information into the dialogue model itself, and extend to generic contradictions.

References

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. 2018. Pyro: Deep Universal Probabilistic Programming. *arXiv preprint arXiv:1810.09538*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, Christopher D Manning, and Stanford Linguistics. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loc Loc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. pages 177–190. Springer, Berlin, Heidelberg.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. *arXiv preprint arXiv:1809.02922*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew D. Hoffman and Andrew Gelman. 2014. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SCITAIL: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Diederik P Kingma and Jimmy Lei Ba. 2014. Adam: A Method For Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. Importance of a Search Strategy in Neural Dialogue Modelling. *arXiv preprint:1811.00907*.

Wuwei Lan and Wei Xu. 2018. Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Bill Maccartney and Christopher D Manning. 2009. An extended model of natural logic. Technical report.

M Marelli, S Menini, M Baroni, L Bentivogli, R Bernardi, and R Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. *arXiv preprint:1705.06476*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018a. On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 513–523, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018b. Collecting Diverse Natural Language Inference Problems for Sentence

Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018c. Hypothesis Only Baselines in Natural Language Inference. In *The Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*.

Oriol Vinyals, Google Quoc, and V Le. 2015. A Neural Conversational Model. In *ICML Deep Learning Workshop*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Dataset Details

A.1 Schema

Relation Types : place_origin, live_in_citystatecountry, live_in_general, nationality, employed_by_company, employed_by_general, has_profession, previous_profession, job_status, teach, school_status, has_degree, attend_school, like_general, like_food, like_drink, like_animal, like_movie, like_music, like_read, like_sports, like_watching, like_activity, like_goto, dislike, has_hobby, has_ability, member_of, want_do, want_job, want, favorite_food, favorite_color, favorite_book, favorite_movie, favorite_music, favorite_music_artist, favorite_activity, favorite_drink, favorite_show, favorite_place, favorite_hobby, favorite_season, favorite_animal, favorite_sport, favorite, own, have, have_pet, have_sibling, have_children, have_family, have_vehicle, physical_attribute, misc_attribute, has_age, marital_status, gender, other.

Additional triples with a not_have relation were extracted using a dependency tree pattern.

Entity Categories : ability, activity, animal, color, citystate, country, company, cuisine, degree_type, drink, family, food, gender, general_location, job_status, language, marital, media_genres, media_other, movie_title, music_artist, music_genre, music_instrument, noun, number, organization, person, person_attribute, person_label, personality_trait, profession, read_author, read_genre, read_title, read_other, school_name, school_status, school_type, season, sport_type, subject, time, vehicle, location, other.

A.2 Relation Swaps

Relation swaps for contradictions include `(have_*, not_have)`, `(own, not_have)`, `(has_hobby, not_have)`, `(like_*, dislike)`, `(favorite_*, dislike)`.

Neutral relation swaps include `(have_x, have_y)`, e.g. `have_pet`, `have_sibling`. Additional `(have_* A, not_have B)` swaps were defined for entities A which are a super-type of B, namely (A,B) pairs ({pet, animal}, {dog, cat}), ({sibling}, {brother, sister}), ({child, kid}, {son, daughter}), ({vehicle}, {car, truck}); this includes sentence pairs such as “i have a sibling”, “i do not have a sister”. Similarly, `(not_have B, have_* A)` swaps were defined using the (A, B) pairs above.

A.3 Entity Swaps

For contradictions, swapping entities for the following relation types was assumed to yield a contradiction:

attend_school, employed_by_company, employed_by_general, favorite_animal, favorite_book, favorite_color, favorite_drink, favorite_food, favorite_hobby, favorite_movie, favorite_music, favorite_music_artist, favorite_place, favorite_season, favorite_show, favorite_sport, gender, has_profession, job_status, live_in_citystatecountry, marital_status, nationality, place_origin, previous_profession, school_status, want_job.

Additionally, for `physical_attribute`, `misc_attribute`, or `other` relations, an entity swap was done using all WordNet antonym pairs in the personality_trait and person_attribute entity categories, as well as the swaps ({blonde}, {brunette}), ({large}, {tiny}), ({carnivore, omnivore}, {vegan, vegetarian}), ({depressed}, {happy, cheerful}), ({clean}, {dirty}) where each entity in the left set is swapped with each entity in the right set.

B Experiment Details

Experiment 1 The InferSent model used the Adam (Kingma and Lei Ba, 2014) optimizer with learning rate 0.001, and otherwise used the hyper-parameters from the open source implementation[7]. The ESIM model used a 1-layer bidirectional LSTM with hidden dimension 1024 and Adam optimizer with learning rate 0.0001, with the remaining hyper-parameters set to those used by the InferSent model.

C Score Calibration

1-5 star rating Let $M_i \sim \mathcal{N}(\mu_i, 1^2)$ be the unobserved, underlying quality of the i-th approach, where $\mu_i \sim \mathcal{U}(1, 5)$. Let $A_j \sim \mathcal{N}(0, 1^2)$ be the unobserved annotator bias, indicating whether the j-th annotator is more or less generous. We observe a score given by the j-th annotator to the i-th approach, and this score follows a normal distribution with its mean given by the sum of the underlying model score and annoator bias, i.e., $S_{ij} \sim \mathcal{N}(M_i + A_j, 1^2)$. We observe some of these scores, and given these scores, the goal is to infer $\mathbb{E}[M_i]$ and $\mathbb{V}[M_i]$ for all i.

Utterance-pair selection Each annotator is asked to label each utterance-pair as consistent and/or contradictory with respect to the personas. In this case, the unobserved, underlying model score is modelled as a pre-sigmoid normal variable, i.e., $M_i \sim \mathcal{N}(0, 1^2)$, and the annotator bias as a usual normal variable, i.e., $A_j \sim \mathcal{N}(0, 1^2)$, similarly to the 1-5 star rating case above. We however also introduce a turn bias $T_k \sim \mathcal{N}(0, 1^2)$ to incorporate the potential degradation of a neural dialogue model as the conversation lengthens. An observed score for each utterance pair then follows a Bernoulli distribution with its mean given as the sigmoid of the sum of these three latent variables, i.e., $S_{ijk} \sim \mathcal{B}(\text{sigmoid}(M_i + A_j + T_k))$. The goal of inference is to compute $\mathbb{E}[\text{sigmoid}(M_i)]$ and $\mathbb{V}[\text{sigmoid}(M_i)]$.

[7] https://github.com/facebookresearch/InferSent