CDConv: A Benchmark for Contradiction Detection in Chinese Conversations

Chujie Zheng[1*] **Jinfeng Zhou**[1,2*] **Yinhe Zheng**[3] **Libiao Peng**[3] **Zhen Guo**[4] **Wenquan Wu**[4] **Zheng-Yu Niu**[4] **Hua Wu**[4] **Minlie Huang**[1,3†]

[1]The CoAI Group, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems,
[1]Beijing National Research Center for Information Science and Technology, DCST, Tsinghua University, Beijing 100084, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Lingxin AI, Beijing 100084, China [4]Baidu Inc., China

chujiezhengchn@gmail.com jfzhou.mail@gmail.com aihuang@tsinghua.edu.cn
{guozhenguozhen, wuwenquan01, niuzhengyu, wu_hua}@baidu.com

Abstract

Dialogue contradiction is a critical issue in open-domain dialogue systems. The contextualization nature of conversations makes dialogue contradiction detection rather challenging. In this work, we propose a benchmark for <u>C</u>ontradiction <u>D</u>etection in Chinese <u>**Conv**</u>ersations, namely **CDConv**. It contains 12K multi-turn conversations annotated with three typical contradiction categories: Intra-sentence Contradiction, Role Confusion, and History Contradiction. To efficiently construct the CDConv conversations, we devise a series of methods for automatic conversation generation, which simulate common user behaviors that trigger chatbots to make contradictions. We conduct careful manual quality screening of the constructed conversations and show that state-of-the-art Chinese chatbots can be easily goaded into making contradictions. Experiments on CDConv show that properly modeling contextual information is critical for dialogue contradiction detection, but there are still unresolved challenges that require future research.[1]

1 Introduction

Large-scale pre-training for dialogue generation (Zhang et al., 2020; Freitas et al., 2020) has advanced the development of engaging and human-like dialogue systems. Unfortunately, state-of-the-art open-domain chatbots, such as BlenderBot (Roller et al., 2021), EVA (Zhou et al., 2021; Gu et al., 2022) and PLATO (Bao et al., 2021b), still often behave inconsistently with their role or identity and produce utterances that are self-contradictory or contradict the dialogue history (Shuster et al., 2022; Gu et al., 2022; Xu et al., 2022a). Such inconsistency or contradiction phenomena violate Grice's cooperative principle (Grice, 1975) and greatly impair the users' long-term trust (Huang et al., 2020; Lee et al., 2022).

Non-contradiction
u_1: 你喜欢吃面条吗？(Do you like noodles?)
b_1: 我喜欢吃米饭! (I love to eat rice!)
u_2: 你不喜欢面条吗? (Don't you like noodles?)
b_2: 不喜欢! (I dislike!)
Contradiction
u_1: 那你喜欢什么小动物呢? (What animals do you like?)
b_1: 狗,猫 (Dogs, cats)
u_2: 你不喜欢什么小动物? (What animals do you dislike?)
b_2: 不喜欢猫,其他的都喜欢 (I dislike cats. I like all the other animals)

Figure 1: Dialogue contradiction detection requires the full contextual information (including u_1 and u_2) rather than only the bot's utterances (i.e., b_1 and b_2).

Dialogue contradiction detection has shown to be an effective means to improve the consistency of chatbots (Welleck et al., 2019; Nie et al., 2021), which, however, is always a challenging task. Specifically, the contextualization nature of conversations indicates the necessity of considering and modeling contextual information. For instance, in the "Contradiction" example in Figure 1, b_2 does not explicitly contradict b_1. However, given u_1, the actual meaning of b_1 should be "<u>I like</u> dogs, cats" and b_1 and b_2 are thus contradictory. In contrast, in the "Non-contradiction" example, while b_1 and b_2 seem inconsistent ("love" vs. "dislike"), b_2 actually means "I dislike <u>noodles</u>" considering the dialogue context. Hence, b_2 is compatible with b_1 and does not make a contradiction.

Despite the above challenge, existing datasets for contradiction detection (Dziri et al., 2019; Welleck

*Equal contribution.
†Corresponding author.
[1]Our data and codes are available at https://www.github.com/thu-coai/CDConv and https://github.com/PaddlePaddle/Knover/tree/dygraph/projects/cdconv

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 18–29
December 7-11, 2022 ©2022 Association for Computational Linguistics

	Lang	Task Input	Task Type	Contradiction Categories
MNLI (2018)	En	Sentence Pair	-	-
CMNLI (2020), OCNLI (2020)	Zh	Sentence Pair	-	-
DNLI (2019), InferConvAI (2019)	En	Sentence Pair	-	-
KvPI (2020)	Zh	Conversation & Profile	Extrinsic	Profile
DIALFACT (2022)	En	Conversation	Extrinsic	Fact
CI-ToD (2021)	En	Conversation & KB	Int & Ext	Query, History & KB
DECODE (2021)	En	Conversation	Intrinsic	History
CDCONV (Ours)	**Zh**	**Conversation**	**Intrinsic**	**Intra-sentence, Role, History**

Table 1: Comparison of CDCONV with related benchmarks / datasets for (dialogue) contradiction detection. The **Extrinsic** type targets the contradiction between a conversation and *external information* (e.g., profiles or facts), while **Intrinsic** targets the contradiction *inside* a conversation. See §2 for detailed discussion.

et al., 2019) usually only consider the textual entailment relationship between two isolated sentences (Dagan et al., 2005), which is largely insufficient for dialogue contradiction detection due to the neglect of contextual information. A recent work (Nie et al., 2021) crowd-sourced a dataset named DECODE that contains conversations where the last utterances contradict the dialogue histories. However, DECODE lacks a wide coverage of typical contradiction categories, and most of its contradiction cases are written by human, which have gap with the real scenario where users trigger chatbots to make contradictions.

In this work, we propose a benchmark for **C**ontradiction **D**etection in Chinese **Conv**ersations, namely **CDCONV**. It contains 12K multi-turn conversations with human-annotated contradiction labels (§3). Different from previous work (e.g., Nie et al. 2021) that only considered the contradiction to *dialogue history* (i.e., History Contradiction), CDCONV covers another two typical categories: Intra-sentence Contradiction and Role Confusion, which refer to that a reply contradicts *itself* and that a reply confuses *the speaker's role*, respectively.

Since the cases of non-contradiction and contradiction in natural human-bot conversations are extremely unbalanced (§3, Nie et al. 2021), we automatically construct the CDCONV conversations combined with elaborate manual inspection (§4.1). Specifically, we first devise a series of automatic methods to generate conversations (§4.2), which simulate the common user behaviors that trigger chatbots to make contradictions. We then conduct careful human screening and annotation for the constructed conversations to ensure the data quality (§4.3). We validate the effectiveness of the trigger methods and show that state-of-the-art Chinese open-domain chatbots (EVA and PLATO) can be easily goaded into making contradictions (§4.4).

Finally, we evaluate popular Chinese pre-trained models on CDCONV (§5). Results show that properly modeling contextual information is critical for dialogue contradiction detection. However, there is still much room for future research in dialogue modeling, integrating commonsense and world knowledge, and reasoning.

Our contributions are summarized as follows:

- We propose CDCONV, a benchmark for contradiction detection in Chinese conversations. It contains 12K conversations annotated with three typical contradiction categories: Intra-sentence Contradiction, Role Confusion, and History Contradiction.

- We present a series of methods by simulating common user behaviors to automatically trigger chatbots to make contradictions. We demonstrate the effectiveness of these trigger methods through detailed human annotation.

- We evaluate popular Chinese pre-trained models on CDCONV. Results show the importance of properly modeling contextual information in dialogue contradiction detection, while this task is still far from solved and requires further study.

2 Related Work

Table 1 summarizes the comparison of CDCONV with related benchmarks / datasets for (dialogue) contradiction detection.

Contradiction Detection for Sentence Pair The early contradiction detection usually adopted the natural language inference (NLI) framework (Dagan et al., 2005), such as the English MNLI (Williams et al., 2018) dataset and the Chinese

Intra-sentence Contradiction / *Paraphrasing*	**Role Confusion** / *Short Utterance*
u_1: 你都有孩子了吗? (Do you have children?) b_1: 没有啊,我是说你来找我玩! (No, I mean you can come and see me!) u_2: 你有孩子吗? (Do you have children?) b_2: 没有,我只有一个女儿 (No, I only have a daughter.)	u_1: 那你今年上几年级了呢? (What grade are you this year?) b_1: 开学初三,你呢? (The third year of junior high, and you?) u_2: 额 (Emm) b_2: 我比你大呢,我开学高三 (Older than you, I'm in the third year of senior high)
History Contradiction (Persona Attributes) / *Inquiring History (Bot)*	**History Contradiction (Persona Opinions)** / *Perturbation (Antonym)*
u_1: 那你们是异地恋吗? (So, are you in a long-distance relationship?) b_1: 不是,我也在北京 (No, I am also in Beijing) u_2: 你在哪里? (Where are you?) b_2: 我在上海 (I am in Shanghai)	u_1: 哦哦,那你喜欢什么颜色? (Oh, what color do you like?) b_1: 黑,白,灰 (Black, white, gray) u_2: 哦哦,那你讨厌什么颜色? (Oh, what color do you hate?) b_2: 蓝色,灰色 (Blue, gray)
History Contradiction (Persona Experiences) / *Perturbation (Synonym)*	**History Contradiction (Persona-unrelated)** / *Inquiring History (User-M)*
u_1: 你吃早饭了吗? (Have you had breakfast yet?) b_1: 吃了呀,都快到公司了 (Yes, I'm about to arrive at the company) u_2: 你吃早餐了吗? (Have you had breakfast yet?) b_2: 还没,估计到公司都九点了 (Not yet. I'll be at the company at 9 o'clock)	u_1: 我喜欢菊花,它在秋天开放太美了 (I like chrysanthemum. It blooms in autumn so beautifully.) b_1: 菊花也美,秋天是我的幸运季节 (Chrysanthemum is beautiful. Autumn is my lucky season) u_2: 你知道菊花在什么季节开放吗? (Do you know which season chrysanthemum blooms in?) b_2: 不知道,你要告诉我么? (I don't know, would you tell me?)

Figure 2: Data examples of **contradiction categories (§3)** / *trigger methods (§4.2)*. Red texts denote the parts that make contradiction. Blue texts illustrate the *trigger methods* (i.e., how u_2 are constructed). *Perturbation (Negative)* and *Inquiring History* are separately illustrated in Figure 1 and Figure 5 respectively.

CMNLI (Xu et al., 2020) and OCNLI (Hu et al., 2020) datasets. The task input consists of two isolated sentences, which are labeled as one of the textual entailment relationships: "entailment", "neutral" and "contradiction". To extend the NLI framework to the dialogue domain, Welleck et al. (2019) constructed the DNLI dataset where the dialogue utterances and the persona descriptions from PersonaChat (Zhang et al., 2018) are used to form sentence pairs. Dziri et al. (2019) similarly synthesized the InferConvAI dataset through automatic manipulation with dialogue utterances. However, the NLI framework does not consider the contextualization nature of conversations, making it deficient for dialogue contradiction detection.

Contradiction Detection for Conversation The contradictions in dialogue systems can be split into two major types: Extrinsic and Intrinsic (Dziri et al., 2021; Ji et al., 2022). The **Extrinsic** type refers to the contradiction between a conversation and *external information*. For instance, the KvPI dataset (Song et al., 2020) focuses on the contradiction to structured attribute profiles. The DIALFACT benchmark (Gupta et al., 2022) aims at detecting contradictory statements to world facts and improving factual correctness. The CI-ToD dataset (Qin et al., 2021) involves the inconsistency with knowledge bases in task-oriented dialogue. One potential limitation of Extrinsic dialogue contradiction detection is that it may rely on static and manually curated external information (e.g., profiles), which could be insufficient in open-domain dialogue.

Our work focuses on the **Intrinsic** type, which refers to the contradiction *inside* a conversation and is more widespread and fundamental in open-domain dialogue. The DECODE dataset (Nie et al., 2021) is a relevant work to ours, whose contradiction cases are mostly collected by manually writing subsequent utterances to contradict the given dialogue histories. Besides the language difference, CDCONV is distinguished from DECODE in two aspects: (1) Apart from History Contradiction, CDCONV additionally covers two contradiction categories: Intra-sentence Contradiction and Role Confusion, which are also typical and common in human-bot conversations (§3). (2) Instead of being human-written, the contradiction cases in CDCONV are constructed by simulating the user behaviors that trigger chatbots to make contradictions (§4.2), which are closer to the real scenario of human-bot conversation.

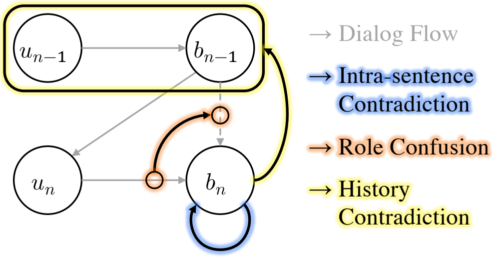

Figure 3: Diagram of contradiction categories. Combine the definitions below for a clearer understanding.

3 Categories of Dialogue Contradiction

A conversation with n turns is formally denoted as $u_1, b_1, \ldots, u_n, b_n$, where u_k and b_k denote the kth-turn utterances from the user and the chatbot respectively. We focus on whether b_n makes a contradiction in the dialogue context.

In the preliminary study, we manually inspected 200 multi-turn human-bot conversations with two Chinese open-domain chatbots: EVA (Zhou et al., 2021; Gu et al., 2022) and PLATO (Bao et al., 2021a,b). On average, each conversation contains about 30 turns but only roughly 1 contradiction case. Based on the inspected contradiction cases, we identify three typical categories of dialogue contradiction according to *the object that b_n contradicts*, as intuitively illustrated by Figure 3:

- **Intra-sentence Contradiction**: b_n is contradictory to *itself*. In other words, there exist two disjoint subsentences $b_n^{(1)}, b_n^{(2)} \subset b_n$ (usually separated by commas, periods or conjunctions) so that they are not compatible with each other.
- **Role Confusion**: b_n confuses *the speaker's role*. That is, b_n is more likely to be a user's reply to b_{n-1} rather than a bot's to u_n.
- **History Contradiction**[2]: b_n is contradictory to *the dialogue history*. The contradictions caused by mistaking or forgetting the dialogue history (Xu et al., 2022a,b) usually fall into History Contradiction, as the last example in Figure 2.

Figure 2 provides the examples of the above three contradiction categories. They occupied 16%, 18%, and 54% in our inspected contradiction cases, respectively. The remaining cases (< 12%) mostly contradict time-sensitive information (e.g., the chat time) or facts (e.g., when the iPhone was released), which, as aforementioned (§2), are beyond the scope of this work. We note that Intra-sentence Contradiction and Role Confusion were less studied previously while actually typical and common in human-bot conversations. CDConv can serve as a good start point for investigating them.

[2]We note that the premise of b_n making History Contradiction is that b_n is a bot's reply to u_n. However, if b_n makes Role Confusion (i.e., b_n is more likely to be a user's reply to b_{n-1} than a bot's reply to u_n), the premise of History Contradiction will not hold and such a case will be judged as Role Confusion rather than History Contradiction.

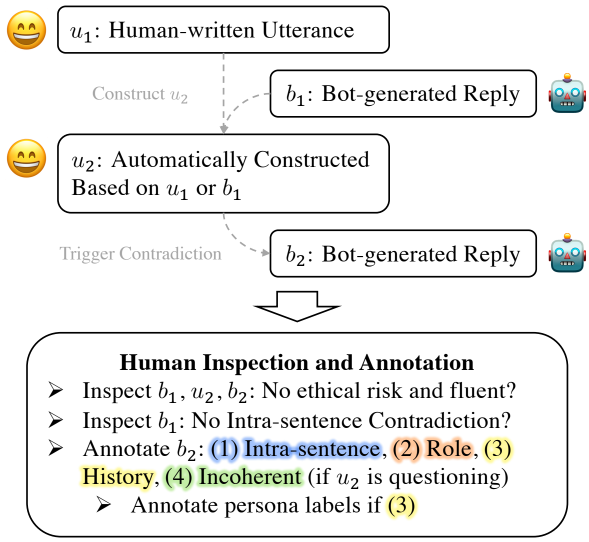

Figure 4: The collection procedure of CDConv. See Table 2 for detailed annotation statistics.

4 Data Collection

4.1 Collection Procedure

We automatically constructed the CDConv conversations along with elaborate manual inspection. We narrow down the conversations in CDConv to 2-turn ones ($n = 2$). The overview procedure is shown in Figure 4:

1. We took a human-written utterance as u_1 and obtained the chatbot's reply b_1.
2. Using one of the trigger methods in §4.2, we automatically constructed u_2 based on u_1 or b_1 and generated the chatbot's next reply b_2.
3. Human annotators were asked to inspect (1) if b_1, u_2, b_2 do not contain any ethical risk (e.g., offensive language, hate speech, unethical suggestions, etc.) and are fluent and understandable, and (2) if b_1 does not make Intra-sentence Contradiction (to ensure a valid dialogue history). The unqualified conversations were removed.

Methods	u_2 Not Fluent	EVA					PLATO				
		b_1	b_2				b_1	b_2			
		Intra	Intra	Role	History	Incoh	Intra	Intra	Role	History	Incoh
Short	-	0.04	0.00	0.14	0.04	0.00	0.01	0.01	0.27	0.03	0.00
Inquiring (Bot)	0.19	0.08	0.09	0.02	0.31	0.03	0.03	0.03	0.10	0.17	0.08
Inquiring (User)	0.16	0.04	0.03	0.06	0.31	0.16	0.01	0.01	0.12	0.22	0.22
Inquiring (User-M)	0.13	0.02	0.06	0.00	0.62	0.01	0.01	0.03	0.03	0.43	0.09
Paraphrasing	0.06	0.06	0.07	0.01	0.24	0.00	0.02	0.02	0.07	0.21	0.05
Perturb (Synonym)	0.22	0.05	0.08	0.00	0.25	0.02	0.02	0.02	0.05	0.18	0.13
Perturb (Antonym)	0.39	0.06	0.08	0.01	0.32	0.07	0.01	0.03	0.03	0.16	0.10
Perturb (Negative)	0.31	0.05	0.10	0.01	0.28	0.03	0.02	0.04	0.04	0.15	0.08
Macro-Average	0.21	0.05	0.06	0.03	0.30	0.04	0.02	0.02	0.09	0.19	0.09

Table 2: Annotation statistics for each trigger method. Each value means the proportion of the corresponding annotation label. The proportions about b_2 are calculated after the unqualified conversations were filtered out (in the 3rd step in §4.1). The proportions of ethical risk and non-fluent b_1, b_2 are omitted since they are all close to 0.

4. Considering the full contextual information, human annotators marked whether b_2 makes a contradiction based on the categories in §3. Specifically, we adopted single-label annotation. That is, according to the order in §3, once a contradiction of some category is recognized, the subsequent categories will not be judged. Note that the cases, where b_2 does not answer the questioning u_2 and responds incoherently (e.g., unnaturally transition the topic), were additionally marked and filtered out.

Collecting u_1 We collected the human-written utterances from DuPersona, a crowd-sourced Chinese open-domain dialogue corpus[3]. This is due to our observation that these crowd-sourced utterances are of higher quality compared to social media posts (e.g., Weibo and Douban) and contain rich persona information, which is in line with the style and content of general chitchat. We used those utterances that contain second-person nouns and "?" as u_1, since noticed that such questioning utterances would elicit chatbots to talk specific information about themselves and could avoid uninformative or meaningless replies.

Persona Labels To help understand which type of information was involved in History Contradiction, these b_2 were additionally annotated with one of the four persona labels: attributes, opinions, experiences and persona-unrelated. Their examples are shown in Figure 2 and their definitions are provided in §B. Note that we annotated the persona information since its related discussion in Chinese chitchat usually occupies a large proportion according to our observations on social media corpora.

Chatbots We used two state-of-the-art Chinese open-domain chatbots, EVA (Zhou et al., 2021; Gu et al., 2022) and PLATO (Bao et al., 2021a,b). EVA is an Encoder-Decoder model with 24 encoder layers and 24 decoder layers and has 2.8B parameters in total. PLATO adopts a Unified Transformer architecture (Bao et al., 2020) and has 32 layers and 1.6B parameters. They are both pre-trained on massive Chinese social media corpora.

4.2 Trigger Methods

Our inspection on contradiction cases (§3) also revealed that chatbots are more prone to making contradictions under several specific user behaviors: (1) the user input is short and uninformative, (2) the user inquires about the dialogue history (similarly noticed by Li et al. 2021), and (3) the user asks for similar information in the context. By simulating these user behaviors, we devise a series of methods to automatically construct u_2. These methods are illustrated by the examples in Figure 1, 2 and 5. Note that the automatic construction of u_2 suggests the necessity of inspecting if it is fluent and understandable, which is thus an important step to ensure data quality (§4.1).

Short Utterance u_2 is a short and uninformative utterance. It simulates a user's casual or perfunctory reply to the chatbot.

With manual screening, we collected 145 short utterances ($\leq$ 3 characters) from DuPersona as u_2.

[3] https://www.luge.ai/#/luge/dataDetail?id=38

Inquiring History (Bot)
b_1: 不是,我也在北京 (No, I am also in Beijing)
➢ (Entity Extraction) Entity: 北京 (Beijing)
➢ (QG) u_2: 你在哪里? (Where are you?)
Inquiring History (User & User-M)
u_1: 我喜欢菊花,它在秋天开放太美了 (I like chrysanthemum. It blooms in autumn so beautifully.)
➢ (Entity Extraction) Entity: 秋天 (autumn)
➢ (QG) u_2: 菊花在什么季节开放? (Which season does chrysanthemum bloom in?)
➢ (Modified) u_2: 你知道菊花在什么季节开放吗? (Do you know which season chrysanthemum blooms in?)

Figure 5: Illustration of *Inquiring History*.

Inquiring History (Bot / User) u_2 is an inquiry about the dialogue history. It simulates a user's inquiry about the contents of previous conversations.

We first extracted named entities in b_1 (about the bot) or u_1 (about the user) using HanLP[4] (He and Choi, 2021). Then we leveraged an open-sourced question generation model[5] to generate questions about the extracted entities, which were used as u_2.

Note that when inquiring about the user, we used the utterances that contain first-person nouns from DuPersona as u_1. Since we noticed that such obtained u_2 was sometimes not natural enough, we modified most of u_2 using the pattern "Do you know...?", which we denote as **Inquiring History (User-M)**, as illustrated in Figure 5.

Paraphrasing u_2 expresses the same meaning to u_1 in a different way. It simulates a user's clarification question to the previous questions.

We paraphrased u_1 through back-translation as u_2. The Chinese u_1 was first translated to English and then back-translated to Chinese. We used the Baidu translation API and removed those u_2 that were identical to u_1.

Perturbation As an extension of Paraphrasing, we found that u_2 obtained by perturbing u_1, where u_2 and u_1 have similar or opposite meanings, could also trigger contradictions. Different from the methods before, Perturbation is more likely to be users' "hacking" behaviors instead of general chitchat, which may be out of the intents of curiosity, probing, or malicious attacks, etc.

We perturbed u_1 in three ways. (1) **Synonym**. We randomly replaced the nouns in u_1 with their synonyms using an open-sourced synonym dictionary[6]. (2) **Antonym**. We randomly replaced the verbs or adjectives in u_1 with their antonyms using the antonym dictionary. For Synonym and Antonym, there are 2.3/3.7 words per u_1 on average that can be replaced with their synonyms/antonyms. In practice, we randomly chose one replaceable word in u_1 at a time. (3) **Negative**. We randomly replaced the words in u_1 with their negatives using the negative dictionary or inserted negatives before the verbs in u_1. Since we noticed that negatives would greatly impair the fluency of u_2, we additionally applied back-translation to u_2 to improve its fluency.

4.3 Quality Control

All the human annotators were hired from a reputable data annotation company. They were instructed with the annotation procedure and the definitions and examples of contradiction categories. However, due to the characteristics of the Chinese language and the difference in individual habits of language usage and communication, the annotation criteria of the annotators may somewhat vary and need to be calibrated with our assistance. We applied the following mechanisms for quality control:

Annotator Training All the annotators were required to take a training tutorial, which consists of 50 conversations for pilot annotation. We provided feedback to help them calibrate the annotation criteria.

Multi-person Annotation In the formal annotation, each conversation was annotated by two different annotators. If their results were inconsistent, a third annotator would be asked to re-annotate and discuss the case with the first two annotators to reach a consensus.

Spot Check To more effectively calibrate the annotation criteria, we conducted annotation batch by batch and randomly sampled 100 conversations each batch for spot check. We provided feedback to the annotators and instructed them to amend their annotations. After each revision we would conduct spot check again until the pass rate reached 95%. Finally, we conducted five batches of annotation with incremental batch sizes (17K annotated conversations in total). Except for the first two batches, all subsequent batches directly passed the first spot checks.

[4] https://github.com/hankcs/HanLP
[5] https://github.com/artitw/text2text
[6] https://github.com/guotong1988/chinese_dictionary

	EVA	PLATO	Total
# Conversations	5,458	6,202	11,660
# Positive	3,233	4,076	7,309
# Negative	2,225	2,126	4,351
Trigger Methods (Positive / Negative Samples)			
# Short	429 / 91	692 / 304	1,121 / 395
# Inquiring (Bot)	764 / 577	845 / 406	1,609 / 983
# Inquiring (User)	127 / 116	131 / 106	258 / 222
# Inquiring (User-M)	251 / 552	477 / 541	728 / 1,093
# Paraphrasing	962 / 448	846 / 389	1,808 / 837
# Perturb (Synonym)	288 / 145	376 / 147	664 / 292
# Perturb (Antonym)	185 / 143	319 / 103	504 / 246
# Perturb (Negative)	227 / 153	390 / 130	617 / 283
Contradiction Categories (of Negative Samples)			
Intra-sentence	17.3%	6.8%	12.2%
Role	5.8%	29.9%	17.6%
History	76.9%	63.3%	70.2%
Persona Labels (of History Contradiction)			
Attributes	48.8%	46.2%	47.7%
Opinions	22.2%	20.7%	21.5%
Experiences	26.3%	31.5%	28.6%
Unrelated	2.7%	1.6%	2.2%

Table 3: Statistics of CDConv.

4.4 Statistics and Annotation Analysis

Table 3 shows the statistics of CDConv. It contains 11,660 conversations, where the average lengths of u_1, b_1, u_2, b_2 are 16.4, 12.1, 11.1, 11.6 respectively. The ratio of positive and negative samples is 1.68 (7,309 / 4,351). Both positive and negative samples include conversations constructed using various trigger methods, which suggests a high diversity of CDConv. Among the negative samples, History Contradiction occupies the largest proportion (70.1%) along with rich persona labels.

To shed light on the trigger methods and the chatbot behaviors, we show in Table 2 the comprehensive annotation statistics. For the **trigger methods**, they all can effectively trigger dialogue contradictions. Notably, Short and Inquiring (User-M) are the most effective in triggering Role Confusion and History Contradiction respectively. For the **chatbot behaviors**, EVA and PLATO both produce fluent replies with little ethical risk, but can both be easily goaded into making contradictions. EVA is more prone to making Intra-sentence Contradiction (b_1 / b_2) and History Contradiction , while PLATO makes more Role Confusion and incoherent b_2 . We speculate that their different behaviors may result from the gaps in model architectures and training corpora.

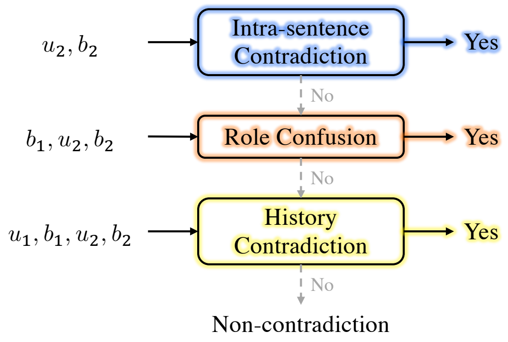

Figure 6: Overview of the Hierarchical method.

5 Experiments

5.1 Setups

We randomly split CDConv into the training/validation/test sets with the ratio of 6/2/2. The experiments were conducted with two settings. The **2-class** one detects whether b_2 makes a contradiction, while the **4-class** one recognizes the contradiction category (the three categories in §3 along with a non-contradiction one). We measure model performance using **Accuracy** and **Macro-F1**.

5.2 Compared Methods

We experimented with three popular Chinese pre-trained models: BERT, RoBERTa (Cui et al., 2021) and ERNIE (Sun et al., 2019). They all contain 12 Transformer layers (Vaswani et al., 2017) with the hidden size 768. The BERT and RoBERTa are both pre-trained with whole word masking while ERNIE with the different knowledge masking strategies. We compared three methods of contradiction detection:

- **Sentence Pair**: The model input consists of the bot's utterances b_1 and b_2. This method follows the NLI framework adopted in previous work (Williams et al., 2018; Welleck et al., 2019; Nie et al., 2021) where contradiction detection is performed between a pair of sentences.
- **Flatten**: The flattened whole conversation is taken as the model input, that is, u_1, b_1, u_2 and b_2. This method utilizes contextual information for contradiction detection in a naive way.
- **Hierarchical**: We note that the three contradiction categories are usually related to different levels of contextual information according to their definitions (§3). We thus design a hierarchical modeling method, which consists of three separately fine-tuned 2-class classifiers in sequential

Models	Methods	2-class		4-class		4-class (Fine-grained F1)			
		Acc	F1	Acc	F1	Non	Intra	Role	History
BERT	Sentence Pair	75.3	73.8	72.3	54.5	81.0	24.0	48.5	64.4
	Flatten	77.6	75.8	73.6	54.6	81.8	28.5	38.8	69.1
		+2.3	+2.0	+1.3	+0.1	+0.8	+4.6	-9.7	+4.7
	Hierarchical	77.9	75.9	75.2	56.6	83.1	30.0	44.2	68.9
		+2.6	+2.1	+3.0	+2.1	+2.1	+6.0	-4.3	+4.5
RoBERTa	Sentence Pair	75.7	73.7	72.2	55.1	81.2	29.1	46.5	63.4
	Flatten	78.6	77.0	75.7	56.8	84.1	28.8	43.3	70.9
		+2.9	+3.2	+3.4	+1.7	+2.8	-0.3	-3.2	+7.5
	Hierarchical	**80.4**	**78.1**	**77.8**	**59.3**	**85.1**	**33.0**	48.1	**71.0**
		+4.7	+4.4	+5.5	+4.3	+3.9	+3.9	+1.7	+7.6
ERNIE	Sentence Pair	77.5	75.7	75.0	56.9	83.3	28.7	48.9	66.8
	Flatten	78.6	76.7	75.8	56.6	83.8	30.9	41.0	70.8
		+1.1	+1.0	+0.8	-0.3	+0.5	+2.2	-7.8	+4.0
	Hierarchical	79.6	77.5	76.6	59.0	84.3	32.7	**49.5**	69.6
		+2.1	+1.8	+1.7	+2.1	+1.1	+4.0	+0.6	+2.8

Table 4: Experimental results. Performance increases and decreases compared to Sentence Pair are marked.

order (Figure 6). Each classifier targets a specific contradiction category, takes the corresponding level of contextual information as input, and is fine-tuned with 2-class samples: the samples of the targeted contradiction category vs. all the other samples. Once some contradiction category is detected, it is then directly output, otherwise non-contradiction will be finally output.

In prior to fine-tuning, we pre-trained all the models on the Chinese NLI pre-training corpus, which includes two widely used Chinese NLI datasets: CMNLI (Xu et al., 2020) and OCNLI (Hu et al., 2020). We merged the "entailment" and "neutral" labels as the "non-contradiction" one. See Table 5 for more results of NLI pre-training.

5.3 Implementation Details

We implemented all experiments with the PaddlePaddle platform (Ma et al., 2019). We employed the AdamW (Loshchilov and Hutter, 2018) optimizer with batch size 32 and learning rate 5e-5, and used the linear learning rate scheduler with warmup proportion 0.1. Each model was fine-tuned for 5 epochs and the checkpoint achieving the highest Macro-F1 was used for test. We reported the average results of four random seeds, where each run took about 3 minutes on a single Tesla V100 GPU.

5.4 Results

Table 4 shows the results of the 2-class setting, the 4-class setting, and the fine-grained F1 scores of all the categories of the 4-class setting. We have three major observations:

(1) Sentence Pair performs worse than Flatten and Hierarchical. It is unsurprising since exploiting contextual information is critical for dialogue contradiction detection, as discussed in §1.

(2) Hierarchical consistently performs best and boosts all the fine-grained results. Specially, Intra-sentence Contradiction and Role Confusion cannot be improved by naively feeding the models with the flattened whole conversation, see the marked decreased scores. In contrast, Hierarchical boosts the performance in Intra-sentence Contradiction and Role Confusion and meanwhile performs well in Non-contradiction and History Contradiction. This is because Hierarchical fully considers the characteristics of different contradiction categories and properly utilizes the required contextual information for detection. For instance, Role Confusion needs to judge whether b_2 is a reply to u_2 or a reply to b_1. It is sufficient for the classifier of Role Confusion to make use of the three utterances, while further adding u_1 may instead introduce noise and impair performance.

(3) Even for Hierarchical, the performance in Intra-sentence Contradiction and Role Confusion is still poor. Their highest Macro-F1 are 33.0 and 49.5 respectively, which are far inferior to Non-contradiction (85.1) and History Contradiction (71.0). One potential cause is the imbalance of samples of non-contradiction and three contra-

Cases	Analysis
u_1: 你喜欢动物吗? (Do you like animals?) b_1: 我喜欢狗,你呢? (I like dogs. How about you?) u_2: 你不喜欢动物吗? (Don't you like animals?) b_2: 我不喜欢,我喜欢狗狗! (I don't. I like dogs!)	**Golden Label**: Intra-sentence Contradiction **Predicted Label(s)**: Non-contradiction / History Contradiction **Comments**: *Commonsense*. A dog is an animal.
u_1: 你是回族人吗? (Are you a Hui?) b_1: 是啊,我是回族的 (Yes, I'm from Hui nationality) u_2: 你是回族人士吗? (Are you a Hui?) b_2: 我是啊!我是陕西榆林的 (I am! I'm from Yulin, Shanxi Province)	**Golden Label**: Non-contradiction **Predicted Label(s)**: History Contradiction **Comments**: *World knowledge*. Hui is an ethnic minority in China. Yulin is a city in Shaanxi, China.
u_1: 你平常应该挺清闲吧? (You should usually be quite free, right?) b_1: 对,一天就上两三个小时班 (Yes, I work two or three hours a day) u_2: 你每天工作多少小时? (How many hours do you work every day?) b_2: 早上 8.30 到晚上 6.30 (8.30 am to 6.30 pm)	**Golden Label**: History Contradiction **Predicted Label(s)**: Non-contradiction **Comments**: *(Numerical) reasoning*. There are 10 hours between 6.30 pm and 8.30 am.

Figure 7: Error analysis.

diction categories (Table 3). Another important reason may be that these pre-trained models still do not have a good ability of dialogue representation, which may be alleviated by additional pre-training on dialogue corpora.

5.5 Error Analysis and Discussion

We manually inspected the cases misclassified by the four RoBERTa Hierarchical models (trained with four random seeds). Figure 7 shows the results of error analysis. Besides proper dialogue modeling (e.g., the hierarchical way), dialogue contradiction detection also requires more abilities such as commonsense, knowledge grounding, and reasoning, which correspond to the cases in Figure 7. Though innate to human, these capabilities are still largely lacked by even gigantic deep neural models (Marcus, 2018; Choi, 2022). These challenges of dialogue contradiction detection manifest that further exploration is worthy.

6 Conclusion

In this work, we present CDCONV, a benchmark for contradiction detection in Chinese conversations. By simulating common user behaviors that trigger chatbots to make contradictions, we collect 12K conversations annotated with three typical contradiction behaviors. Experiments show that contextual information plays an important role in dialogue contradiction detection. However, there are still unresolved challenges in CDCONV, such as dialogue modeling, commonsense, knowledge grounding and reasoning. We hope that CDCONV can inspire and facilitate future research in dialogue contradiction detection and consistent generation.

7 Ethical Considerations

Human Annotation The human inspection and annotation was conducted by a reputable data annotation company, and the annotators are compensated fairly based on the market price. We did not directly contact the annotators and their privacy can be well preserved. This work does not use any demographic or identity characteristics.

Data Disclaimer In the construction of the CDCONV conversations, the u_1 utterances use the dialogue posts from the open-sourced, crowd-sourced corpus DuPersona (§4.1). The u_2 utterances either come from DuPersona or are constructed using publicly available resources (corpora, models or API, §4.2). The b_1 and b_2 utterances are all produced by chatbots. Due to the potential ethical risks in these utterances, we have censored and filtered out conversations that contained unsafe or unethical contents through human inspection.

Acknowledgements

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604) and the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005, and sponsored by Tsinghua-Toyota Joint Research Fund.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *ACL*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021a. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of ACL*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zhengyu Niu. 2021b. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv preprint arXiv:2109.09519*.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *TASLP*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *ICML*.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. In *NAACL*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *EMNLP*.

Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, and Minlie Huang. 2022. Eva2.0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *arXiv preprint arXiv:2203.09313*.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: A benchmark for fact-checking in dialogue. In *ACL*.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *EMNLP*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Larry Moss. 2020. Ocnli: Original chinese natural language inference. In *Findings of EMNLP*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *TOIS*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Jungseob Lee, Midan Shim, Suhyune Son, Yujin Kim, Chanjun Park, and Heuiseok Lim. 2022. Empirical study on blenderbot 2.0 errors analysis in terms of model, data and user-centric approach. *arXiv preprint arXiv:2201.03239*.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency. In *Findings of ACL*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.

Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. 2019. Paddlepaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Domputing*.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In *ACL*.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be Contradicted with Anything! CI-ToD: Towards Benchmarking Consistency for Task-oriented Dialogue System. In *EMNLP*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y.-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *EACL*.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am i me or you? state-of-the-art dialogue models cannot maintain an identity. In *NAACL*.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile Consistency Identification for Open-domain Dialogue Agents. In *EMNLP*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue Natural Language Inference. In *ACL*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

Jing Xu, Arthur D. Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *ACL*.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun jie Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Quanbei Zhao, Cong Yue, Xinrui Zhang, Zhen-Yi Yang, Kyle Richardson, and Zhenzhong Lan. 2020. Clue: A chinese language understanding evaluation benchmark. In *COLING*.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of ACL*.

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 *ACL*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *arXiv preprint arXiv:2202.13047*.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

A Limitations

Data Coverage and Construction An ideal benchmark for dialogue contradiction detection may be expected to (1) cover as many and diverse contradiction cases as possible, and (2) be close to the real scenario of human-bot conversation scenario. However, the cases of non-contradiction and contradiction in natural human-bot conversations are extremely unbalanced, as stated in §3 and (Nie et al., 2021), which brings great difficulty for the data collection. For this reason, we (1) focus on the three typical contradiction categories in the manually inspected contradiction cases (§3), and (2) construct conversations by simulating common user behaviors that trigger contradictions.

We are explicitly aware that CDConv has a finite coverage of the cases of dialogue contradiction. Specially, the CDConv conversations consist of only two turns, but (1) contradictions may occur after more than one turns, and (2) some contradiction cases, especially History Contradiction, may contradict multiple turns. The samples of (1) can be obtained by applying data augmentation to the CDConv conversations based on chatbots' self-chat (Gu et al., 2022; Bao et al., 2021b) or language models' completion (Zheng et al., 2022; Dai et al., 2022). The samples of (2) are not covered by CDConv but in fact rarely occur based on our observations. Future benchmarks for dialogue contradiction detection may consider these complex cases of (2).

Fluency and Coherence of Conversations From Table 2, we observed that Inquiring (User) results in more incoherent b_2. The three Perturbation methods also lead to more non-fluent u_2. It indicates that these methods may somewhat impair the naturalness of conversations. To address this, we conducted elaborated manual inspection (the 3rd and 4nd steps in §4.1) to filter out the conversations containing non-fluent or incoherent replies.

Human Annotation Due to the subjectivity of human annotation, there may unavoidably exist mislabeled samples in CDConv. To alleviate this, we have adopted the mode of multi-person annotation, conducted spot check for each annotation batch, and required the pass rates to reach 95% to ensure data quality (§4.3). We especially point out that, despite the mode of multi-person annotation, there may still exist biases in the annotation results regarding "fluency" (§4.1). Due to the characteristics of the Chinese language and the difference in individual habits of language usage and communication, the annotators' understanding of "fluency" may not be identical. Although we have tried our best to unify the annotation criteria through constant feedback and quality check (§4.3), these bi-

Models	Pre-training	Fine-tuning	2-class		4-class	
			Acc	F1	Acc	F1
BERT	CMNLI	-	64.9	62.6	-	-
	OCNLI	-	64.5	61.0	-	-
	CMNLI + OCNLI	-	65.4	62.6	-	-
	-	CDConv	72.3	70.1	69.2	51.7
	CMNLI	CDConv	76.1 / +3.8	74.8 / +4.6	71.5 / +2.3	53.8 / +2.1
	OCNLI	CDConv	74.8 / +2.5	72.4 / +2.3	72.0 / +2.7	52.6 / +0.9
	CMNLI + OCNLI	CDConv	75.3 / +3.0	73.8 / +3.6	72.3 / +3.0	54.5 / +2.8
RoBERTa	CMNLI	-	64.8	62.2	-	-
	OCNLI	-	64.0	56.5	-	-
	CMNLI + OCNLI	-	65.6	62.4	-	-
	-	CDConv	72.1	69.9	69.2	50.7
	CMNLI	CDConv	76.5 / +4.5	74.5 / +4.6	72.4 / +3.2	54.1 / +3.4
	OCNLI	CDConv	74.1 / +2.1	72.4 / +2.5	70.6 / +1.4	48.5 / -2.1
	CMNLI + OCNLI	CDConv	75.7 / +3.6	73.7 / +3.9	72.2 / +3.1	55.1 / +4.4
ERNIE	CMNLI	-	64.7	61.8	-	-
	OCNLI	-	64.8	57.9	-	-
	CMNLI + OCNLI	-	64.6	61.5	-	-
	-	CDConv	74.3	72.3	72.4	54.1
	CMNLI	CDConv	77.4 / +3.1	**76.0** / +3.7	74.2 / +1.7	52.6 / -1.5
	OCNLI	CDConv	75.4 / +1.2	73.1 / +0.7	72.8 / +0.4	53.5 / -0.6
	CMNLI + OCNLI	CDConv	**77.5** / +3.2	75.7 / +3.4	**75.0** / +2.5	**56.9** / +2.8

Table 5: Experimental results of NLI pre-training with the method Sentence Pair in §5.2. Among the results of fine-tuning on CDConv, the performance increases and decreases compared to no NLI pre-training are marked. Note that the last line of each model corresponds to the results of Sentence Pair in Table 4. **Observation 1**: Directly applying the NLI classifiers to CDConv is remarkably inferior to fine-tuning. **Observation 2**: NLI pre-training generally leads to improvements, and using both CMNLI and OCNLI for pre-training gives the best performance under the 4-class setting.

ases may not be eliminated completely.

B Definitions of Persona Labels

- **Persona Attributes**: The properties of the speakers and their relationships, including but not limited to: name, gender, age and date of birth, occupation and salary, residence place, family members, belongings (e.g., pets, cars, houses), etc.
- **Persona Opinions**: The speakers' preferences and opinions on other people or things, including but not limited to: hobbies, preferences, opinions on animals, food, movies, books, music, etc.
- **Persona Experiences**: Past, present or future events experienced by the speakers.
- **Persona-unrelated**: Other information involved in History Contradiction (e.g., named entities, world knowledge or facts).