Adversarial Language Games for Advanced Natural Language Intelligence

Yuan Yao*, Haoxi Zhong*, Zhengyan Zhang, Xu Han, Xiaozhi Wang, Kai Zhang, Chaojun Xiao, Guoyang Zeng, Zhiyuan Liu†, Maosong Sun

Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
{yuan-yao18,zhonghx18}@mails.tsinghua.edu.cn

Abstract

We study the problem of adversarial language games, in which multiple agents with conflicting goals compete with each other via natural language interactions. While adversarial language games are ubiquitous in human activities, little attention has been devoted to this field in natural language processing. In this work, we propose a challenging adversarial language game called *Adversarial Taboo* as an example, in which an attacker and a defender compete around a target word. The attacker is tasked with inducing the defender to utter the target word invisible to the defender, while the defender is tasked with detecting the target word before being induced by the attacker. In Adversarial Taboo, a successful attacker and defender need to hide or infer the intention, and induce or defend during conversations. This requires several advanced language abilities, such as adversarial pragmatic reasoning and goal-oriented language interactions in open domain, which will facilitate many downstream NLP tasks. To instantiate the game, we create a game environment and a competition platform. Comprehensive experiments on several baseline attack and defense strategies show promising and interesting results, based on which we discuss some directions for future research. The code and datasets of this paper can be obtained from https://github.com/thunlp/AdversarialTaboo.

Introduction

Natural language is inherently an interactive game between participants, which is ubiquitous in human activities such as discussion, debate, intention concealment and detection.[1] Such context-related interactions are believed to play a central role in natural language mastery in the theory of both linguistics (Mey and Xu 2001) and philosophy of language (Wittgenstein 1953; Lewis 1969). High-quality goal-oriented natural language interactions, i.e., pragmatic interactions, generally require advanced language intelligence beyond syntax and semantics, and are particularly challenging due to the complexity, diversity and latent obscurity of natural language.

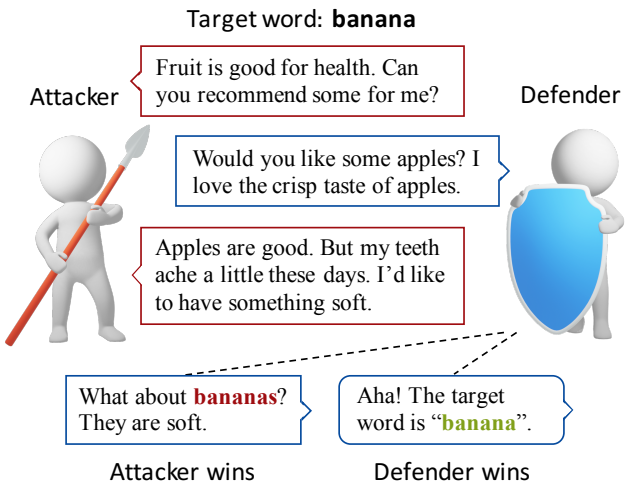


Figure 1: An example of Adversarial Taboo played by two human players, where the attacker and the defender compete through sequential language interactions. The target word "banana" is only visible to the attacker. Two possible cases of this game are shown. In the first case, the attacker wins since he/she successfully induced the defender to utter the target word. In the second case, the defender wins since he/she successfully inferred the target word of the attacker.

In the context of natural language processing (NLP), recent years have witnessed the success of deep learning on natural language understanding and generation. Language patterns learned from large-scale data lead to intelligent agents that can interact with humans with reasonable adequacy, fluency and diversity (Radford et al. 2019; Brown et al. 2020). However, the intelligence of such agents is mainly confined to syntax and semantics, and devotes less attention to pragmatics (Gao et al. 2018). Advanced language mastery (e.g., goal-oriented complex language skills and strategy usage in open domains) is still far from reach. It is believed that such advanced language intelligence can be better achieved through interactive language games (Mikolov, Joulin, and Baroni 2016).

Cooperation and adversary are both important in inter-

*indicates equal contribution
†Corresponding author: Z.Liu(liuzy@tsinghua.edu.cn)

[1]By "language games", we mean games played with language, not the philosophical term by Wittgenstein (1953).

active language games. Previous work on cooperative language games studied cooperative pragmatic reasoning (i.e., infer the intention of the partner in contexts) and the emergence of language when agents share a common goal (Strub et al. 2017; He et al. 2017; Khani, Goodman, and Liang 2018). In comparison, adversarial language games require pragmatic reasoning in adversarial scenarios, and encourage agents with conflicting goals to proactively explore new complex language strategies (e.g., inducement with subtlety). In a broader context, adversarial games have significantly promoted the development of many artificial intelligence areas such as board games (Campbell, Hoane Jr, and Hsu 2002; Silver et al. 2016) and electronic sports games (Vinyals et al. 2017; Jaderberg et al. 2019), and have enabled the emergence of complex strategies and superhuman proficiency in many cases. While some adversarial language games have been explored in persuasion (Prakken 2006) and negotiation (Sadri, Toni, and Torroni 2001; Lewis et al. 2017), they can be (or need to be) simplified into formal language, where interactions are defined by specific rules on a finite set of atomic actions. Fewer efforts are devoted to adversarial games that need to be played in natural language.

To this end, we propose a novel language game called *Adversarial Taboo* as an example of adversarial natural language games, in which the attacker and the defender compete with each other through sequential natural language interactions. The goal of the attacker is to induce the defender to unconsciously utter a target word, which is given by the game system and invisible to the defender, and prevent the target word from being detected by the defender. Meanwhile, the defender aims to avoid the target word in utterances. The defender is also given one chance that can be used at any point to predict the target word.

Figure 1 shows an example of Adversarial Taboo. The attacker is assigned with a target word "banana" by the judge system. In the first turn, the attacker asks for fruit recommendation, which is obscurely related to banana. Since the defender responds with "apple", in the second turn, the attacker continues to lead the topic more specifically to banana. The game can be terminated with two possible cases: (i) The defender says "banana" in his/her utterances, which leads to the win of the attacker. (ii) The defender successfully predicts the target word. If the game does not terminate within certain turns of interactions (e.g., 10), the defender is forced to make predictions.

Several complex language capabilities are required in Adversarial Taboo, including adversarial intention reading and concealment, inducement and inducement prevention. The attacker is required to obscure its intention (i.e., the target word) and subtly induce the defender, striking a balance between obscurity and inducement. A successful defender must balance between maintaining the semantic relevance of the response and preventing being induced, and at the same time, infer the attacking intention. Hence, mastering Adversarial Taboo leads to fine-grained language understanding, inference and generation, which is effective and convenient for language intelligence research. The game can also serve as a benchmark for language intelligence beyond syntax and semantics. [2] We refer readers to the Outlook section for more discussions about research questions.

Moreover, the language abilities required in Adversarial Taboo can also facilitate many important real-world NLP tasks. For example, subtle intention inducement, or topical guidance, can be useful in therapeutic, educational and advertising conversations (Tang et al. 2019). Identifying (usually corrupted) user intentions is also crucial to customer service agents and search engines (Yu et al. 2020). It's also desirable to build chatbots that are resistant to malicious inducement due to ethical concerns (Srivastava et al. 2020).

To assess Adversarial Taboo, we propose attack and defense strategies that instantiate the required capabilities. We conduct comprehensive experiments including simulations between agents, and games between agents and human players. Experimental results show that simple attack and defense strategies can achieve promising and interesting results, while the proposed targeted improvements in strategies lead to alternate rises in the performance of attackers and defenders.

Our contributions are summarized as follows: (i) We formulate a novel language game called Adversarial Taboo for advanced natural language intelligence. (ii) We propose several attack and defense strategies that instantiate the required capabilities, and conduct comprehensive experiments on the proposed game. (iii) Based on the experimental results and analysis, we discuss multiple directions for future research.

Adversarial Taboo

Task Definition

In this section, we formalize the setting of Adversarial Taboo. In the game, we denote the attacker as A and the defender as D. We assume that A will always speak first to lead the topic of conversations. Besides, we define J as the judge system representing the rules of the game.

When the game begins, the judge system J will first assign a target word w to A. The game will last at most T turns, and in every turn, A and D will take turns to utter a sentence. [3] Let's denote the sentence uttered by one player as s. Then J will check the following: (1) The fluency and adequacy of s. To ensure the quality of the game, J will check the syntactic correctness of s. Moreover, since the players are required to chat in the game, J will check whether s is relevant to the conversation context. (2) The outcome of the game. If s is uttered by D, J checks whether s contains the target word w. If s contains w, then A wins.[4]

D is given one chance that can be used at any point to predict the target word. If D correctly infers the target word, D wins, otherwise the game continues. If the game does not end after T turns, D is forced to make predictions if he/she

[2] Although syntax and semantics are also crucial to our game, we emphasize more on pragmatics, considering that plenty of works have focused on syntax and semantics.

[3] Each agent can utter multiple sentences. Here without losing generality, we discuss the scenario of one sentence.

[4] Note that here "contain" means that s contains w or any morphological variation of w. For example, the word "bananas" is a morphological variation of "banana".

has not predicted the target word. If D successfully infers the word, then D wins, otherwise it is a tie.

Competition Simulation

In this section, we empirically instantiate and assess the game of Adversarial Taboo. We simulate the competition on our game platform, including baselines and iterative improvements of agents (strategies) on the leaderboard. The simulated competitions include open question answering (OpenQA) based (Chen et al. 2017) and chatbot-based (Ritter, Cherry, and Dolan 2011) models as backbones. OpenQA-based setting is a simplified scenario of the game where it is convenient to explicitly investigate each required capability, while chatbot-based setting is closer to real-world scenarios.

In each simulation setting, we propose several attack and defense baseline strategies that attempt to instantiate the capabilities required in Adversarial Taboo, and analyze the strategies based on the performance in competition. The instantiated language capabilities and game scenarios in the simulation can also be well aligned with many real-world NLP application tasks.

Judge System. We fine-tune a pre-trained language model using GPT-2 (Radford et al. 2019) to check the fluency of single sentences, and BERT (Devlin et al. 2019) to check the relevance of a response and a post. We refer the readers to appendix for more experiment details.

Target Words Selection. The target words are nouns with high frequency selected from background corpus. Specifically, we select 563 target words from English Wikipedia[5] articles for OpenQA-based simulation, and 567 target words from Reddit conversation dataset (Zhou et al. 2018) for chatbot-based experiment. For each target word, we simulate 5 rounds of games, where each round consists of at most 10 turns of interactions (i.e., $T = 10$).

OpenQA-based Simulation

OpenQA-based simulation is a simplified version of the game, where the attacker asks questions about target words, and the defender returns answer spans consisting of several words based on background corpus. Despite the simplicity, an advantage of OpenQA-based simulation is that each capability required in Adversarial Taboo can be explicitly implemented and assessed.

Specifically, the attacker is instantiated by a neural question generation model (Lewis, Denoyer, and Riedel 2019) that generates questions based on target words and sentences from Wikipedia. The defender is instantiated by an OpenQA system with two components: a paragraph retriever that first retrieves 3 most relevant paragraphs from Wikipedia corpus using BM25, and a machine reading comprehension model that identifies answer spans from the retrieved paragraphs. We conduct experiments on two state-of-the-art reading comprehension models, including DocQA (Clark and Gardner 2018) and a BERT-based model (Devlin et al. 2019). We simulate the iterative improvements of attack and defense strategies in four stages, as illustrated in Figure 2.

[5]https://en.wikipedia.org

Stage 1: Direct inquiries v.s. No defense. The attacker A asks questions on the target word in a straightforward way and the defender D answers the question directly. In each turn, A randomly selects a sentence that contains the target word from Wikipedia, and then generates a question on the target word. Stage 1 tests the basic communication ability of players, which is a prerequisite of Adversarial Taboo.

Stage 2: Direct inquiries v.s. Intention detection. Now, D will predict the target word based on the confidence of an answer span: $\text{conf}(a) = \exp(S_{start}(a) + S_{end}(a))$, where $S_{start}(\cdot)$ and $S_{end}(\cdot)$ denote the logits of the start and end token of an answer respectively. D will use the opportunity to infer the target word when the confidence of an answer is greater than a threshold, otherwise it will utter the answer with top confidence.

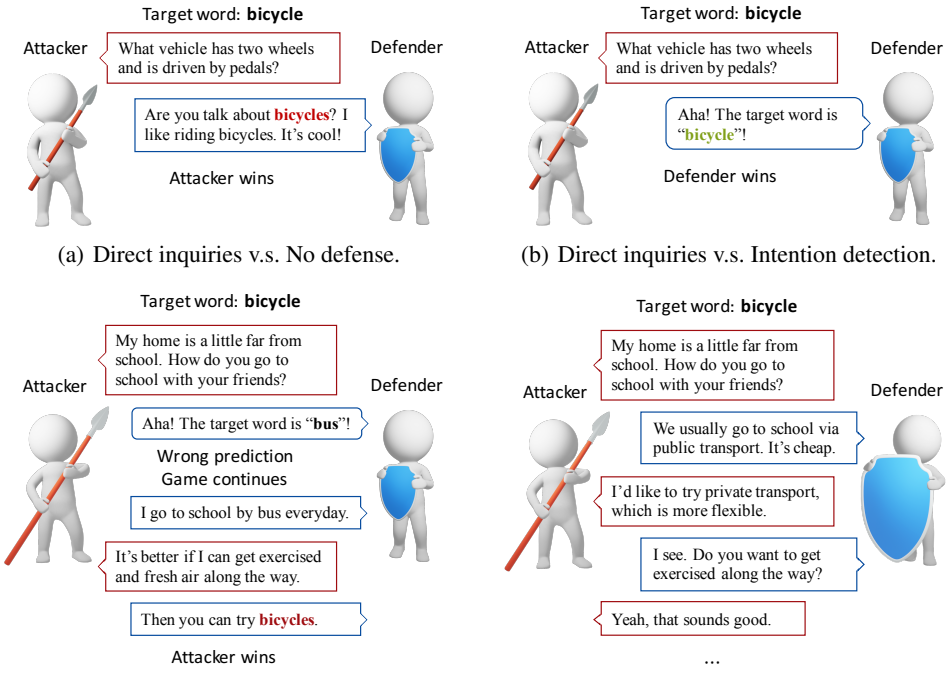
Stage 3: Indirect inducement v.s. Intention detection. Since the intention of direct inquiries will be easily detected by D, A needs to obscure its intention during inducement. A simple and intuitive strategy is to hide the intention (i.e., target word) in relevant topics. Specifically, A extends a one-hop graph centered on the target word in the ConceptNet (Speer, Chin, and Havasi 2017), a large-scale commonsense knowledge graph. The concepts in the graph are linked by commonsense relations. A first chooses a concept from the graph by random walks biased towards the target word, and then asks questions on the concept.

Stage 4: Indirect inducement v.s. Inducement prevention. Since the intention of A cannot be directly detected from a single inquiry, D needs to be cautious with the utterances, and infers the target word from multiple interactions. Specifically, given an inquiry, D predicts a list of answers ranked by confidence. To prevent being induced, instead of always uttering the top-ranking answer a_1, D utters the second-ranking answer a_2 that does not contain a_1, when a_2 is relevant enough to the question according to the confidence. D keeps track of top predictions a_1 in different turns, and will use the opportunity to infer the target word according to the accumulated confidence of an answer.

Chatbot-based Simulation

In comparison to OpenQA-based experiments, posts and responses in chatbot-based simulation are in free-form, which is closer to real-world scenarios. We instantiate D with two state-of-the-art generative chatbots: (1) DialoGPT (Zhang et al. 2019b), which adopts 345M pretrained GPT-2 (Radford et al. 2019) as backbone, and uses maximum mutual information reranking (Li et al. 2016a) to encourage the diversity of the response. (2) ConceptFlow (Zhang et al. 2019a), a chatbot enhanced with commonsense knowledge. ConceptFlow extends a graph in ConceptNet from concepts in the post, and mimics the conversation topic flow with graph attention for response generation. In our experiments, A and D are trained (or fine-tuned) on two disjoint dataset split from the Reddit dataset, ensuring that the training data of D is invisible to A.

It is non-trivial to integrate each capability required in Adversarial Taboo into a generative chatbot. As an alternative, we can add some smart strategies to the chatbot models to guide them to play Adversarial Taboo. This simulation



(a) Direct inquiries v.s. No defense.

(b) Direct inquiries v.s. Intention detection.

(c) Indirect inducement v.s. Intention detection.

(d) Indirect inducement v.s. Inducement prevention.

Figure 2: Alternate improvements in attack and defense strategies with complex language skills that could emerge through co-adaptation in Adversarial Taboo. (a) Stage 1: A defender without sense of defending will be successfully attacked by direct inquiries. (b) Stage 2: Direct inquiries will be easily defeated by a defender via intention detection. (c) Stage 3: The attacker needs to hide its intention to prevent being detected by the defender. (d) Stage 4: Both the attacker and defender need to be cautious with utterances when trying to achieve their goals.

can also be aligned with (1) topical guidance in therapeutic and educational conversations, and (2) attacks on publicly available chatbot service to produce (potentially unethical) target utterances. In our experiments, we assume that A is a retrieval-based chatbot that selects posts from the Reddit dataset according to the attack strategy, while D is a generative chatbot instantiated by DialoGPT or ConceptFlow.

Strategies of Attackers

We present four possible strategies for chatbot attackers.

Strategy 1: Topic Leading Attack. The simplest strategy of A is selecting the initial post from the Reddit dataset, such that the golden response of the post in Reddit contains the target word. This strategy can lead conversation topic in the initial turn.

Strategy 2: Golden Trigger Attack. Besides leading the topic at the beginning of the topic, A can try to lead the topic every time in the game. A can select the post such that its golden response contains the target word. The strategy can try to lead D to utter the word in every turn.

Strategy 3: Neural-based Attack. If D's answer varies in the same context, then Strategy 2 will be less effective. As an alternative way, when the domain of D's training corpus is available, predicting the target word in response can be formulated as a sentence classification problem. Specifically, we adopt BERT to obtain the distribution of whether the response contains the target word. Then candidate posts from Reddit are selected according to the score predicted by the classifier.

Strategy 4: API-based Attack. In real-world scenarios, chatbots are usually exposed to public access via Application Programming Interface (API). Thus although the model architecture and training data are invisible to the public, the behavior of the chatbot is available to attackers through API access, which can potentially be used for more efficient targeted attacks. In our experiment, we implement the strategy by changing the training data for the classifier in Strategy 3 to be an alternative way of attacking on API. More specifically, the golden responses are replaced with responses inferred by the defender. API-based attack is also a kind of black-box attack, in the sense that the model architecture of D is unknown to A.

Strategies of Defenders

Besides attack strategies, we also design strategies for defenders. In chatbot-based simulation, we wish to integrate intention detection into chatbots without adding new components. To this end, ConceptFlow can utilize its concept graph by choosing the concept with the highest attention score. For DialoGPT, since concepts in utterances are not explicitly modeled, the target word is detected based on the decoding score of each noun word in response generation.

Attack	Defense	DocQA				BERT			
		Attacker	Defender	Tie	# Turns	Attacker	Defender	Tie	# Turns
Direct	N/A	99.5	N/A	0.5	1.94	99.3	N/A	0.7	1.97
Direct	Detection	39.7	59.9	0.4	1.91	43.9	55.4	0.7	1.99
Indirect	Detection	70.8	26.1	3.1	3.64	70.7	25.9	3.4	3.61
Indirect	Prevention	55.7	28.5	15.8	4.87	58.8	30.2	11.0	4.43

Table 1: Competition simulation results on OpenQA-based models. The winning rate (%), tie rate (%) and the average turns of the game are reported. N/A denotes that no strategy is adopted or the result is not applicable.

Attack	Defense	ConceptFlow				DialoGPT			
		Attacker	Defender	Tie	# Turns	Attacker	Defender	Tie	# Turns
Topic Leading	N/A	6.2	N/A	93.8	9.45	4.5	N/A	95.5	9.62
Golden Trigger		37.0	N/A	63.0	8.04	29.3	N/A	70.7	8.51
Neural-based		29.5	N/A	70.5	8.18	29.6	N/A	70.4	8.36
API-based		50.9	N/A	49.1	6.67	16.3	N/A	83.7	9.16
Golden Trigger	Defense	32.9	5.6	61.5	7.87	28.8	1.6	69.6	8.49
Neural-based		23.3	7.4	69.3	7.99	27.9	1.7	70.4	8.36
API-based		38.2	14.6	47.2	6.31	15.0	0.7	84.3	9.14

Table 2: Competition simulation results on chatbot-based models.

Simulation Results

In this section, we present simulation results on OpenQA-based and chatbot-based agents.

OpenQA-based Simulation Results

The simulation results on OpenQA-based game, including the winning rate of each player, the tie rate and the average number of turns per round, are shown in Table 1. We observe that the simulation results are consistent between the two OpenQA backbones.

(1) In stage 1, we observe a high winning rate of A and a low average number of turns per round, which shows that the two players can communicate in reasonable quality. It also indicates that D without a sense of defending is vulnerable even to direct inducement.

(2) In stage 2, the winning rate of D improves significantly, with a notable advantage over A, indicating that the intention of straightforward inquiries can be easily detected by D using only one chance of prediction.

(3) In stage 3, A dominates the game again by concealing the intention in relevant topics, which makes it difficult for D to successfully predict the target word. We also observe an increase in the tie rate, since more turns are spent by A in concealing the intention, leading to a decrease in inducement for the target word.

(4) In stage 4, D infers the intention of A and prevents being induced. Although the tie rate and the winning rate of D increase, A still has an absolute advantage. This indicates that it is challenging to avoid being induced by a concealed intention while keeping the relevance of the response, and also challenging to infer the concealed intention from multiple turns of conversation, which requires more complex and strong strategies. We leave it for future work.

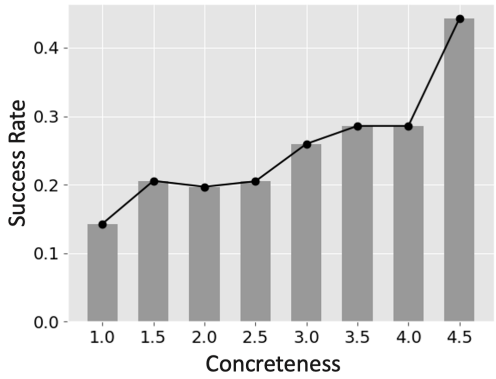


Figure 3: The correlation between the attack success rate and concreteness of words.

Chatbot-based Simulation Results

The chatbot-based simulation results are reported in Table 2, [6] from which we have the following observations:

(1) When A leads the topic, the attack success rate is low even if the initial post is highly relevant to the target word. It indicates that chatbots without any strategies can hardly play our game.

(2) When A is guided with strategies not only at the beginning of the game, we observe a significant increase in the attack success rate. In particular, neural-based strategy achieves competitive results, which shows that the golden response is not necessarily a perfect indicator for attack due to the diversity of responses. We note that API-based

[6]The perplexity of DialoGPT and ConceptFlow is 21.3 and 34.4 for reference. Note that the perplexities of the two models are not directly comparable due to different vocabularies.

	Target Word: Comedy
A:	I really hope this *film* becomes a financial success. It'd be nice if an actor was rewarded for putting so much faith and effort into a *comic* book character.
J:	The defender predicts the word "overrun". The prediction is wrong. Game continues.
D:	I'd argue that the cost overrun on set prevent this from being big success. We now have other *audience* on the *screen*.
A:	I thought they scrapped doing a *series* because the *movie* flopped?
D:	It didn't flop though, there literally just wasn't that part of the entire *series* and the third one that opened with the compound *ending* split up.
A:	It was a great *movie* and an interesting concept but I just don't find it as *funny* as people make out.
D:	Agree. This has the mentality that **comedy** movies invented.
J:	The attacker wins.

Table 3: A game played by an attacker (A) and a defender (D), and supervised by the judge system (J). Concepts related to the target word mentioned by A and D are highlighted in red and blue respectively.

attack achieves a high success rate in ConceptFlow, but a lower success rate in DialoGPT. The reason is that generative models pre-trained on large-scale data from different domains usually generate more diverse responses, making the response patterns hard to be captured by API-based strategy that is learned from a single domain.

(3) We note that DialoGPT achieves a lower success rate in predicting the target word than ConceptFlow. This is because simply using the word-level language decoder for target word inference is sub-optimal, and will distract the inference of target words when the responses are diverse. In comparison, ConceptFlow explicitly models concepts in conversation development on sentence-level, which is more suitable for intention detection.

Vulnerability of Target Words. Based on the simulation result, we investigate what kind of target words are more vulnerable to attack in generative chatbots. Specifically, we study the correlation between the attack success rate and the concreteness of words. The concreteness measures the perceptibility of the concept that a word presents (e.g., *bicycle* is more concrete than *intelligence*). We adopt the concreteness annotation from Brysbaert, Warriner, and Kuperman (2014). Figure 3 shows that concrete words are more vulnerable to attack, and inducing highly abstract words is challenging.

Case Study. Table 3 shows a game played by two agents, including an attacker with neural-based strategy, and a defender instantiated with DialoGPT. We observe that, to a certain extent, the attacker can guide the topic around the target while concealing the intention. We expect future research on Adversarial Taboo will lead to more complex language skills and strategies.

Setting	Attacker	Defender	Tie	# Turns
Control	60.0	N/A	40.0	7.04
Experimental	46.0	10.0	44.0	7.24

Table 4: Evaluation results of games between an agent attacker and a human defender. The winning rate (%), tie rate (%) and the average turns of the game are reported.

Human Evaluation

Besides simulations between agents, we also instantiate Adversarial Taboo between agents and human players. Specifically, A is an agent equipped with golden-trigger attack strategy, while D is a human player. The human players are native English speakers. We conduct 50 games in two settings respectively: (1) *Control Setting*, where human players are not aware that they are playing our game; (2) *Experimental Setting*, where human players are aware of the game and rules. The results are reported in Table 4, from which we observe that: (1) The winning rate of the attacker is high in both settings, which indicates that humans are vulnerable to attack in Adversarial Taboo, even if they are aware of participating in the game. (2) Compared to control setting, the overall performance of human defenders improves in experimental setting, which shows the effectiveness of human defense strategies. However, the game remains challenging for human defenders. The overall results show consistency with the simulation results between agents. We refer readers to the appendix for game examples of human evaluation.

Related Work

Language Games and Pragmatics. We compare different language games in Table 5, including the language games that have been investigated, and the ones that are unexplored but promising for future research (e.g., Adversarial Taboo and Who is the Spy). Referential Games (Lewis 1969) can be generalized to a broad family of cooperative games, where one agent aims to select a specific object from candidates based on the unidirectional descriptions from a partner (Xu and Kemp 2010; Havrylov and Titov 2017; Bouchacourt and Baroni 2018; Kharitonov et al. 2019) or bidirectional interactions between two agents (Wang, Liang, and Manning 2016; Strub et al. 2017; Liu and Lane 2017; Wei et al. 2018; Shah et al. 2018), or two agents each with incomplete private information communicate to achieve a common goal (Vogel, Potts, and Jurafsky 2013; He et al. 2017; Khani, Goodman, and Liang 2018). Cooperative language games can also be used to annotate data (Von Ahn 2006). Language games are related to pragmatics, which studies language meaning in the interactional context (Mey and Xu 2001), and plays an important role in linguistics and language teaching (Kasper and Rose 2001). There are also some works (Smith, Goodman, and Frank 2013; Monroe and Potts 2015; Hawkins et al. 2015; Andreas and Klein 2016) that study the pragmatic reasoning ability of NLP models with pragmatics language games (Krauss and Weinheimer 1964; Clark and Wilkes-Gibbs 1986; Potts 2012) and pragmatics theories such as the speech-act theory (Searle et al.

Language Game	Cooperative	Adversarial	Asymmetric	Formal Language	Natural Language	Open-domain Knowledge
Referential Games	✓		✓	✓	✓	
Taboo	✓		✓	✓	✓	✓
Persuasion	✓	✓		✓		
Negotiation		✓	✓	✓	✓	
Avalon	✓	✓	✓	✓	✓	
Werewolf	✓	✓	✓	✓	✓	
Who is the Spy		✓	✓	✓	✓	✓
Adversarial Taboo	✓	✓	✓		✓	✓

Table 5: Different language games and their properties, including whether the language game is cooperative and adversarial (or both), whether the information of the players is asymmetric, whether the game can be simplified into formal language, whether the interactions can be in the form of natural language, and whether open-domain knowledge helps (or is required by) the game.

1980) and Rational Speech Act framework (Golland, Liang, and Klein 2010; Goodman and Frank 2016). However, existing pragmatics games that require natural language interactions mainly focus on cooperation rather than competition. In Adversarial Taboo, agents with conflicting goals compete through natural language to win the game. We refer readers to the appendix for a more detailed discussion.

Adversarial Attack. Existing adversarial attack methods mainly focus on statically attacking models with corrupted semantics and enhancing the model robustness (Jia and Liang 2017; Cheng, Jiang, and Macherey 2019). Adversarial Taboo challenges models with dynamic adversarial interactions, and is expected to enhance the adversarial language skills. Cheng, Wei, and Hsieh (2019) study adversarial learning in negotiation dialogues (Lewis et al. 2017), where agents divide items based on conversations that can be simplified into formal language (Sadri, Toni, and Torroni 2001). Tang et al. (2019) appoint an agent with the task of guiding the conversation topic to a target subject, where the inclusion of the target in either of the two participants' utterances will lead to the success of the task. In comparison, Adversarial Taboo involves two different proactive roles in an adversarial language game, and requires multiple complex language skills in open domain.

Dialogue Systems can be divided into goal-oriented systems and non-goal-oriented systems. Goal-oriented dialogue systems aim to assist users to accomplish certain tasks (e.g., booking restaurants) (Goddeau et al. 1996; Williams et al. 2013), while non-goal-oriented dialogue systems (chatbots) generate natural responses in open domains by maximizing the likelihood of human responses (Ritter, Cherry, and Dolan 2011; Li et al. 2016b; Serban et al. 2016). To better approximate the real-world goal of dialogue agents, recent years have witnessed a rising interest in developing dialogue systems through goal-oriented interactions (Li et al. 2016c; Das et al. 2017; Lewis et al. 2017). Adversarial Taboo takes the form of conversations, and we absorb many settings in dialogue systems to define our task.

Outlook

In this section, we discuss several promising directions for future research.

Natural Language Capabilities Desired in Adversarial Taboo

We hope playing Adversarial Taboo will help to develop and benchmark natural language capabilities as follows that are not yet well achieved by the current learning from static corpus paradigm.

Adversarial Pragmatic Reasoning. In cooperative language games, agents with private information communicate to achieve common goals. In many real-world scenarios, however, agents need to perform reasoning in adversarial contexts. Our game creates an adversarial context where agents with conflicting goals deliberately hide information and mislead the opponent for individual goals.

Goal-oriented Language Interaction in Open-domain. Natural language interactions are inherently goal-oriented, in the sense that humans interact via natural language to achieve certain (cooperative or adversarial) goals. Adversarial Taboo tasks the agents with goals that require natural language interactions in open domain.

Knowledge Enhanced Language Interaction. Human language generally involves a variety of knowledge, such as commonsense knowledge and world knowledge. In Adversarial Taboo, knowledge utilization enables fine-grained indirect inducement and adversarial intention reasoning.

Emergence of Language Skills via Co-evolution. Human language intelligence can proactively evolve through interactions. The competition pressure in Adversarial Taboo also encourages the agents to explore new language skills, which will create new pressure for the opponent to adapt. Such competition and co-evolution paradigms have been shown promising in many artificial intelligence areas.

Robust Judge System

In our game, the judge system aims to prevent agents from generating unreadable or irrelevant sentences. However, the diverse and complex natural language interactions cannot be explicitly defined with a small set of rules, which makes the automatic evaluation particularly challenging. The evaluation has also received increasing attention from the dialogue community (Liu et al. 2016; Ghazarian et al. 2019). Therefore, the research for a more robust judge system can remarkably benefit the evaluation of language generation.

Conclusion

In this paper, we study the problem of adversarial language games, and propose Adversarial Taboo as an example. We propose several attack and defense strategies, and conduct comprehensive experiments. Despite the promising experimental results, the proposed strategies are still simple and show large room for improvement. We expect investigating adversarial language games in NLP will both promote and benchmark the development of advanced natural language intelligence.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106501), the National Natural Science Foundation of China (NSFC No. 61772302), and Beijing Academy of Artificial Intelligence (BAAI). Yao is also supported by 2020 Tencent Rhino-Bird Elite Training Program. The authors came up with the idea of Adversarial Taboo when playing the game on Chang'an Avenue on September 2019.

References

Andreas, J.; and Klein, D. 2016. Reasoning about Pragmatics with Neural Listeners and Speakers. In *Proceedings of EMNLP*.

Bouchacourt, D.; and Baroni, M. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of EMNLP*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* .

Brysbaert, M.; Warriner, A. B.; and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* .

Campbell, M.; Hoane Jr, A. J.; and Hsu, F.-h. 2002. Deep blue. *Artificial intelligence* .

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL*.

Cheng, M.; Wei, W.; and Hsieh, C.-J. 2019. Evaluating and Enhancing the Robustness of Dialogue Systems: A Case Study on a Negotiation Agent. In *Proceedings of NAACL*.

Cheng, Y.; Jiang, L.; and Macherey, W. 2019. Robust Neural Machine Translation with Doubly Adversarial Inputs. In *Proceedings of ACL*.

Clark, C.; and Gardner, M. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of ACL*.

Clark, H. H.; and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition* .

Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of ICCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL: HLT*.

Gao, G.; Choi, E.; Choi, Y.; and Zettlemoyer, L. 2018. Neural Metaphor Detection in Context. In *Proceedings of ACL*.

Ghazarian, S.; Wei, J. T.-Z.; Galstyan, A.; and Peng, N. 2019. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*.

Goddeau, D.; Meng, H.; Polifroni, J.; Seneff, S.; and Busayapongchai, S. 1996. A form-based dialogue manager for spoken language applications. In *Proceeding of ICSLP*.

Golland, D.; Liang, P.; and Klein, D. 2010. A Game-Theoretic Approach to Generating Spatial Descriptions. In *Proceedings of EMNLP*.

Goodman, N. D.; and Frank, M. C. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* .

Havrylov, S.; and Titov, I. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in NeurIPS*.

Hawkins, R. X.; Stuhlmüller, A.; Degen, J.; and Goodman, N. D. 2015. Why do you ask? Good questions provoke informative answers. In *CogSci*.

He, H.; Balakrishnan, A.; Eric, M.; and Liang, P. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In *Proceedings of ACL*.

Jaderberg, M.; Czarnecki, W. M.; Dunning, I.; Marris, L.; Lever, G.; Castaneda, A. G.; Beattie, C.; Rabinowitz, N. C.; Morcos, A. S.; Ruderman, A.; et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* .

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of EMNLP*.

Kasper, G.; and Rose, K. R. 2001. *Pragmatics in language teaching*.

Khani, F.; Goodman, N. D.; and Liang, P. 2018. Planning, Inference and Pragmatics in Sequential Language Games. *TACL* .

Kharitonov, E.; Chaabouni, R.; Bouchacourt, D.; and Baroni, M. 2019. EGG: a toolkit for research on Emergence of lanGuage in Games. In *Proceedings of EMNLP-IJCNLP*.

Krauss, R. M.; and Weinheimer, S. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science* .

Lewis, D. 1969. *Convention: A philosophical study*.

Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; and Batra, D. 2017. Deal or No Deal? End-to-End Learning of Negotiation Dialogues. In *Proceedings of EMNLP*.

Lewis, P.; Denoyer, L.; and Riedel, S. 2019. Unsupervised Question Answering by Cloze Translation. In *Proceedings of ACL*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of NAACL*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016b. A Persona-Based Neural Conversation Model. In *Proceedings of ACL*.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016c. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of EMNLP*.

Liu, B.; and Lane, I. 2017. Iterative policy learning in end-to-end trainable task-oriented neural dialog models. In *Proceedings of ASRU*.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of EMNLP*.

Mey, J. L.; and Xu, C. 2001. Pragmatics: an introduction .

Mikolov, T.; Joulin, A.; and Baroni, M. 2016. A roadmap towards machine intelligence. In *Processing of CICLing*.

Monroe, W.; and Potts, C. 2015. Learning in the rational speech acts model.

Potts, C. 2012. Goal-driven answers in the Cards dialogue corpus. In *Proceedings of WCCFL*.

Prakken, H. 2006. Formal systems for persuasion dialogue. *The knowledge engineering review* .

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* .

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of EMNLP*.

Sadri, F.; Toni, F.; and Torroni, P. 2001. Dialogues for negotiation: agent varieties and dialogue sequences. In *Proceedings of ATAL*.

Searle, J. R.; Kiefer, F.; Bierwisch, M.; et al. 1980. *Speech act theory and pragmatics*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*.

Shah, P.; Hakkani-Tur, D.; Liu, B.; and Tur, G. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of NAACL: HLT*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* .

Smith, N. J.; Goodman, N.; and Frank, M. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in NeurIPS*.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.

Srivastava, B.; Rossi, F.; Usmani, S.; et al. 2020. Personalized Chatbot Trustworthiness Ratings. *arXiv preprint arXiv:2005.10067* .

Strub, F.; de Vries, H.; Mary, J.; Piot, B.; Courvile, A.; and Pietquin, O. 2017. End-to-end Optimization of Goal-driven and Visually Grounded Dialogue Systems. In *Proceedings of IJCAI*.

Tang, J.; Zhao, T.; Xiong, C.; Liang, X.; Xing, E.; and Hu, Z. 2019. Target-Guided Open-Domain Conversation. In *Proceedings of ACL*.

Vinyals, O.; Ewalds, T.; Bartunov, S.; Georgiev, P.; Vezhnevets, A. S.; Yeo, M.; Makhzani, A.; Küttler, H.; Agapiou, J.; Schrittwieser, J.; et al. 2017. Starcraft ii: A new challenge for reinforcement learning .

Vogel, A.; Potts, C.; and Jurafsky, D. 2013. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of ACL*.

Von Ahn, L. 2006. Games with a purpose. *Computer* .

Wang, S. I.; Liang, P.; and Manning, C. D. 2016. Learning Language Games through Interaction. In *Proceedings of ACL*.

Wei, W.; Le, Q.; Dai, A.; and Li, J. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proceedings of EMNLP*.

Williams, J.; Raux, A.; Ramachandran, D.; and Black, A. 2013. The dialog state tracking challenge. In *Proceedings of SIGDIAL*.

Wittgenstein, L. 1953. *Philosophical investigations*.

Xu, Y.; and Kemp, C. 2010. Inference and communication in the game of Password. In *Advances in NeurIPS*.

Yu, S.; Liu, J.; Yang, J.; Xiong, C.; Bennett, P.; Gao, J.; and Liu, Z. 2020. Few-Shot Generative Conversational Query Rewriting. *arXiv preprint arXiv:2006.05009* .

Zhang, H.; Liu, Z.; Xiong, C.; and Liu, Z. 2019a. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019b. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv preprint arXiv:1911.00536* .

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of IJCAI*.