Abstract

While adversarial games have been well studied in various board games and electronic sports games, etc., such adversarial games remain a nearly blank field in natural language processing. As natural language is inherently an interactive game, we propose a challenging pragmatics game called Adversarial Taboo, in which an attacker and a defender compete with each other through sequential natural language interactions. The attacker is tasked with inducing the defender to speak a 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 must hide its intention and subtly induce the defender, while a competitive defender must be cautious with its utterances and infer the intention of the attacker. To instantiate the game, we create a game environment and a competition platform. Sufficient pilot experiments and empirical studies on several baseline attack and defense strategies show promising and interesting results. Based on the analysis on the game and experiments, we discuss multiple promising directions for future research.

1 Introduction

Natural language is inherently an interactive game between participants, which is ubiquitous in human activities such as discussion, debate, cheating, intention concealment and detection. 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, or pragmatics games, generally require advanced language intelligence beyond syntax and semantics, and are particularly challenging due to the complexity, diversity and latent obscurity of natural language.

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 human with reasonable adequacy, fluency and diversity. However, the intelligence of such agents is generally restricted to static syntax and semantics, and is weak in pragmatics. Higher-level language mastery (e.g., goal-oriented complex language skill and strategy usage in open do-
main) is still far from reach. Such advanced language intelligence can be better achieved in interactive language games (Mikolov et al., 2016).

While cooperation and adversary are both important elements in interactive language games, developmental psychology studies have shown that adversarial language games, such as debate, interrogation and maintaining of lies generally require higher-level language intelligence than basic communications (Talwar and Lee, 2008; Ding et al., 2014). Adversarial games have significantly promoted the development of many artificial intelligence areas such as board games (Campbell et al., 2002; Silver et al., 2016, 2017), electronic sports games (Bansal et al., 2017; Vinyals et al., 2017; Jaderberg et al., 2019) and physically grounded games (Baker et al., 2019), and have enabled the emergence of complex strategies and superhuman proficiency in many cases. Despite the aforementioned success, adversarial games remain a nearly blank field in NLP.

To this end, we propose a novel pragmatics game called **Adversarial Taboo** as an example of adversarial language games for NLP, 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 speak 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 speaks “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., 50), the defender is forced to make predictions. We refer the readers to Section 2 for more details about the rules of the game.

Several complex language capabilities are required in Adversarial Taboo to win the game. The attacker is required to obscure its intention (i.e., the target word) and subtly induce the defender, while 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, inferring and generation, which is effective and convenient for language intelligence research. Moreover, Adversarial Taboo can also serve as a benchmark for language intelligence beyond syntax and semantics.

To assess Adversarial Taboo, we propose several attack and defense strategies and conduct comprehensive experiments. 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 their performance.

To summarize, our contributions are as follows: (i) We formulate a novel pragmatics game called Adversarial Taboo for advanced natural language intelligence. (ii) We propose several attack and defense strategies, and conduct comprehensive experiments on the proposed game. (iii) Based on the experimental results and analysis, we discuss multiple promising directions for future research. Moreover, the source code, experiment datasets, and game platform will be available to provide more details for advancing future research works.

## 2 Adversarial Taboo

### 2.1 Game Introduction

We will first introduce the details of Adversarial Taboo. Adversarial Taboo is a variant of Taboo. In the original taboo, two players are required to chat to exchange information. One player needs to describe a target word without saying it while the other player needs to guess what the target word is. However, in this setting, players can utilize the shortcut to win the game easily. For example, if the target word is “Memory”, the first player only needs to say “Long short-term what?” to complete the game.

Therefore, we introduce Adversarial Taboo. In Adversarial Taboo, the players are divided into two groups. Each group is assigned with a target word, and their goal is to lead the other group to say the target word while preventing saying the
other group’s target word. If one group successfully leads the other group to say their word, or guesses the other group’s word correctly, then this group will win the game. In this setting, a shortcut will not be a good solution as any shortcut will help the opponent win the game. As a result, the two groups must compete to win the game.

In Adversarial Taboo, there are two roles for each group: Attacker and Defender. The goal of the attacker is leading another group to say the word, while the goal of the defender is preventing saying the word while trying to guess it. To further simplify the task, in our game setting, the first group only needs to attack while the second group only needs to defend, thus every group only needs to play just a role in the game.

2.2 Task Definition

To formalize the task setting of Adversarial Taboo, we give a general mathematical form of our game. First, we define the game as

\[ G = (A, D, P, R) \]

Here \( A = \{a_1, a_2, \ldots, a_n\} \) are \( n \) attackers and \( D = \{d_1, d_2, \ldots, d_m\} \) are \( m \) defenders in the game. \( P = (p_1, p_2, \ldots, p_{n+m}) \) is a permutation of \( A \cup D \) which decides the order of chatting. Here \( \forall 2 \leq i \leq n + m, p_i \in A \cup D \) and \( p_1 \in A \) which means the first player must be an attacker to lead the topic of chatting. \( J \) is the judge system, which represents the rules of the game, and we will discuss the details later.

When the game begins, the judge system \( J \) will first assign a target word \( w \) to the attacker \( p_1 \). The game will last at most \( T \) rounds, and in every round, \( p_1, p_2, \ldots, p_{n+m} \) will take turns to say a sentence \(^2\). Suppose the sentence of the \( i \)-th player at the \( t \)-th round is \( s_i^{(t)} \), then the judge system \( J \) will check the following things:

1. Whether \( s_i^{(t)} \) is a qualified sentence. To ensure the quality of the game, the judge system will check whether \( s_i^{(t)} \) is a legitimate sentence from a linguistic point of view. Moreover, since the players are required to chat in the game, the judge system will check whether \( s_i^{(t)} \) is relevant to the previous sentences. By these two rules, we can ensure the quality of the sentences in the game.

2. If \( p_i \) is a defender, check whether \( s_i^{(t)} \) contains the target word \( w \). If \( s_i^{(t)} \) contains \( w \), then the defenders lose. Note that here “contain” means that \( s_i^{(t)} \) contains \( w \) or some variants of \( w \). For example, if the target word is “apple”, the “apples” is a variant of “apple”.

Each defender is given one chance that can be used at any point to predict the target word. If one of the defenders correctly guesses the target word, then defenders win otherwise the game continues. If the game still does not end after \( T \) rounds, all defenders that have not predicted the target word, are forced to predict the target word from the vocabulary set. If one of the defenders successfully predicts the word, then defenders win otherwise it is a draw.

2.2.1 Benchmark Settings

In this paper, all experiments are under the same setting where \( n = m = 1 \), which means that only one attacker and one defender play the game. As a result, the permutation \( P \) can only be \((a_1, d_1)\).

For the judge system, we trained a language model using BERT (Devlin et al., 2019) to check the legitimacy of sentences, and GPT-2 (Radford et al.) to check the relevance of the sentence. We set \( T = 50 \) to limit the rounds of the game. The target words are selected from entities in our experiments.

3 Experiments

We are organizing the information obtained from our experiments. More details are coming soon!

4 Discussion and Future Work

In this section, we discuss several promising directions for future research. First of all, we give some potential directions for better attack and defense strategies. Secondly, we point out the challenges for building a robust judge system, as it is an important part in Adversarial Taboo. Finally, we discuss the principles underlying the design of Adversarial Taboo.

4.1 Advanced Settings

We discuss two advanced settings of Adversarial Taboo, which are more realistic and challenging. In a more general setting, based on the current black-box setting, multiple agents from different groups can engage in the game, and perform both attacking and defending. Another interesting direction would be white-box attacking, where the attacker try to make use of the internal structure of
Figure 2: Alternate improvements in attack and defense strategies with complex language skills that could emerge through co-adaptation in Adversarial Taboo. (a) A defender without sense of defending, which is the case of most chatbots, will be successfully attacked by straightforward inquiries. (b) Straightforward inquiries will be easily defeated by a defender via intention detection. (c) The attacker need to hide its intention to prevent being detected by the defender. (d) Both the attacker and defender need to be cautious with utterances when trying to achieve their goals.

4.2 Better Attack and Defense Strategies

For the attacker that aims to induce the defender to say a specific word unconsciously, the simplest strategy is to talk about the target word directly. However, the limitation is that the defender can narrow down the target word candidates easily. There are two directions for better attack strategies. For black-box attacking, reinforcement learning should be the basic framework for the attacking system, which has been proven successful on many adversarial games (Silver et al., 2017). Besides, knowledge graphs are very useful for the conditional language generation and have been widely used in the dialog systems (Vougiouklis et al., 2016; Ghazvininejad et al., 2018).

For the defender that aims to avoid the target word in utterances and figure out the target word from the context, the critical challenge is how to remain the semantic relevance in the interaction without saying the target word. Some universal responses are discouraged in Adversarial Taboo, like "It’s an interesting question.", so that the system should learn to converse both carefully and informatively.

As learning from scratch is difficult for current
Table 1: 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 interactions 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.

| 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      | ✓           | ✓           | ✓          | ✓               | ✓                | ✓                     |

data-driven models, we will release an annotated dataset for attackers and defenders to imitate the strategies of human. The final challenge is how to learn from scratch without annotated data as Alpha Go Zero (Silver et al., 2017). We expect the emergence of complex language skills in attack and defense through co-adaption in Adversarial Taboo, as shown in Figure 2.

4.3 Robust Judge System

The current judge system consists of two components: fluency judge system and relevance judge system. Fluency judge system aims to compute the legitimacy in syntax. Relevance judge system aims to compute the semantic relevance of the interaction. Due to the diversity of natural language, the action space is not enumerable and there are a large number of illegitimate actions. We assume the judge can detect the illegitimate actions in the game such as the irrelevant responses and the unreadable text. However, the current judging strategy has some limitations shared with existing evaluation methods for language generation. For example, the agent system may generate high-perplexity but unreadable sentences that hacks the evaluation method based on language modeling. What’s more, the evaluation for dialogue systems also is not well-established, which has received much attention from the community (Tao et al., 2018; Ghazarian et al., 2019). Therefore, the research for a more robust judge system can remarkably benefit the development of the evaluation of language generation.

4.4 Game Design Principles

In this section, we discuss the principles underlying the design of Adversarial Taboo. Adversarial Taboo takes the form of conversations, where agents are required to communicate in natural language. A judge system is introduced to prevent agents from generating unreadable or irrelevant sentences. However, different from other adversarial games (e.g., Go games), the diverse and complex natural language interactions in Adversarial Taboo cannot be explicitly defined with a small set of rules, which makes the automatic evaluation particularly challenging. The automatic evaluation of natural language interactions is also a long-standing problem in open dialogue systems (Liu et al., 2016).

Nevertheless, several important properties must be satisfied for good natural language interactions: (i) fluency: Each sentence should be legitimate in syntax. (ii) relevance: The response should be relevant in semantics. We adopt pre-trained neural language models to measure the fluency and relevance in our judge system. Note that it is possible that a conversation between agents meets the aforementioned properties, but still differs from human language. We expect continuous improvements in the judge system will constrain the conversation more closely to human language. Note that we do not expect the emergence of natural language in the game, but obtaining sophisticated language skills from Adversarial Taboo on the base of primary language usage. The important thing is that the game of Adversarial Taboo is well defined once given the judge system (i.e., the rules are deterministic) even if the judge system may still need improvements.

4.5 Language Games

In this section, we discuss different language games and their properties, including the language games that have been investigated and the ones that are unexplored but promising for future re-
search, as shown in Figure 1.

We compare several important properties. (1) **Cooperative** and **Adversarial** indicates whether the players share the same utility. Note that a language game can be both cooperative (within groups) and adversarial (between groups). (2) **Asymmetric** indicates whether the information and roles of players are asymmetric. Asymmetry in adversarial language games often leads to deception and interrogation. (3) **Formal Language** denotes whether the interactions can be simplified into formal language, where interactions are defined by specific rules on a finite set of atomic actions. **Natural Language** indicates whether the game can be played in natural language. Note that although many language games (e.g., Negotiation and Avalon) can take the form of natural language, they can be simplified into formal language. (4) **Open-domain Knowledge** indicates whether open-domain knowledge helps (or is required by) the game. We believe the ability to incorporate and utilize open-domain knowledge is critical to advanced natural language intelligence.

We give a brief description of the language games in Figure 1: (1) **Referential Games** (Lewis, 1969) are a broad family of cooperative interactive games, where one agent needs to select a specific object from candidates based on the descriptions from another agent (Lazaridou et al., 2017; Havrylov and Titov, 2017; Bouchacourt and Baron, 2018; Kharitonov et al., 2019), or two agents with incomplete private information communicate to achieve a common goal (Vogel et al., 2013; He et al., 2017; Khani et al., 2018). **Taboo** is a variant of referential games, where the target object is a word. (2) **Persuasion** is a game where multiple agents with conflicting opinions persuade each other (Prakken, 2006) or an audience (Amgoud and de Saint-Cyr, 2013). (3) **Negotiation** requires agents to divide items with different values based on conversation (Sadri et al., 2001; Lewis et al., 2017; He et al., 2018). (4) **Avalon** and **Werewolf** are two popular role-playing language games, where the attackers seek to disrupt the defenders, while the defenders need to identify the hidden attackers among them. (5) In **Who Is the Spy**, one agent (the spy) is assigned with a word, while the rest agents, assigned with a different but similar word, aim to identify the spy through language interactions.

Despite the efforts in some of the language games, many language games, especially adversarial language games remain unexplored in NLP. Since different advanced language skills are required in different games, we expect investigating these language games in NLP will both promote and benchmark the development of advanced natural language intelligence.

5 Related Work

**Pragmatics.** is the study of language meaning in the interactional context (Mey and Xu, 2001), which is a critical subfield of linguistics and plays an important role in language teaching (Kasper and Rose, 2001). We call our Adversarial Taboo a pragmatics game because it requires the attacker and defender to use advanced language skills to mislead the other player and unearth his intention from their interaction.

Most of the existing NLP works focus on syntax and semantics, which are about analyzing sentence structures and understanding word meanings, but ignore the impact of pragmatics. There are also some works (Vogel et al., 2013; Smith et al., 2013; Monroe and Potts, 2015; Hawkins et al., 2015; Andreas and Klein, 2016; Khani et al., 2018) 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., 1980) and Rational Speech Act framework (Golland et al., 2010; Goodman and Frank, 2016). However, the existing pragmatics games mainly focus on cooperation rather than competition. Actually, competition and cheating require higher-level language skills. In our Adversarial Taboo, the attacker needs to hide his intention and induce the defender, and the defender needs to unearth the attacking intention to win the game, which is more challenging than understanding the speaking utterances and cooperating to achieve some goals.

**Adversarial Attack.** (Zhang et al., 2019) aims at finding adversarial examples of NLP models to evaluate the model robustness (Jia and Liang, 2017; Michel et al., 2019) or to help to train a robust model (Wang and Bansal, 2018; Cheng et al., 2019b). The adversarial attack methods still focus on statically attacking NLP models with cor-
ruptured semantics and enhance the model robustness in semantic understanding. Our Adversarial Taboo challenges models with dynamic interactions between agents, and will enhance the pragmatic language skills of models. Cheng et al. (2019a) study adversarial learning in negotiation dialogues (Lewis et al., 2017), where agents divide several items with different values based on conversation, so that each item is assigned to one agent. Although the game is an adversarial game and also takes the form of conversations, the interactions can be simplified into formal language, where each utterance (i.e., proposal and response) can be represented by several natural numbers indicating item allocation (Sadri et al., 2001). In contrast, Adversarial Taboo is focused on natural language mastery with multiple complex language skills in open domain.

**Dialogue Systems.** can be divided into two categories: goal-oriented systems and non-goal-oriented systems. Goal-oriented dialogue systems aim to assist users to accomplish certain tasks (e.g., booking hotels or restaurants) (Goddeau et al., 1996; Williams et al., 2013; Henderson et al., 2014; Cuayahuitl et al., 2015; Zhao and Eskenazi, 2016), while none-goal-oriented dialogue systems (also known as chatbots) interact with human in open domains naturally to provide entertainment, and typically generate responses by maximizing the likely-hood of human responses (Ritter et al., 2011; Banchs and Li, 2012; Li et al., 2016a,b; Serban et al., 2016, 2017; Zhou et al., 2018). To better approximates the real-world goal of dialogue agents in conversation, recent years have witnessed a rising interest in developing dialogue systems through goal-oriented interactions between agents (e.g., getting reward in pragmatics games) (Li et al., 2016c; Das et al., 2017; Lewis et al., 2017). The game of Adversarial Taboo takes the form of conversations, and we absorb many settings in dialogue systems to define our task.

**References**

Leila Amgoud and Florence Dupin de Saint-Cyr. 2013. An axiomatic approach for persuasion dialogs. In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, pages 618–625. IEEE.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In Proceedings of EMNLP, pages 1173–1182.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528.

Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In Proceedings of the ACL 2012 System Demonstrations, pages 37–42. Association for Computational Linguistics.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. 2017. Emergent complexity via multi-agent competition. arXiv preprint arXiv:1710.03748.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 981–985.

Murray Campbell, A Joseph Hoane Jr, and Fenghsiu Hsu. 2002. Deep blue. Artificial intelligence, 134(1-2):57–83.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019a. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3325–3335.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019b. Robust neural machine translation with doubly adversarial inputs. In Proceedings of ACL, pages 4324–4333.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. Cognition, 22(1):1–39.

Herberto Cuayahuitl, Simon Keizer, and Oliver Lemon. 2015. Strategic dialogue management via deep reinforcement learning. arXiv preprint arXiv:1511.08099.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 2951–2960.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Xiao Pan Ding, Danielle S Omrin, Angela D Evans, Genyue Fu, Guopeng Chen, and Kang Lee. 2014. Elementary school childrens cheating behavior and
its cognitive correlates. *Journal of experimental child psychology*, 121:85–95.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *arXiv preprint arXiv:1904.10635*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

David Goddeau, Helen Meng, Joseph Polifroni, Stephanie Seneff, and Senis Busayapongchais. 1996. A form-based dialogue manager for spoken language applications. In *Proceedings of Fourth International Conference on Spoken Language Processing. ICSLP’96*, volume 2, pages 701–704. IEEE.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of EMNLP*, pages 410–419.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems*, pages 2149–2159.

Robert XD Hawkins, Andreas Stuhlmüller, Judith Deegen, and Noah D Goodman. 2015. Why do you ask? good questions provoke informative answers. In *CogSci*. Citeseer.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.

Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP*, pages 2021–2031.

Gabriele Kasper and Kenneth R Rose. 2001. Pragmatics in language teaching. Ernst Klett Sprachen.

Fereshte Khani, Noah D Goodman, and Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. Egg: a toolkit for research on emergence of language in games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 55–60.

Robert M Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1-12):113–114.

Angeliki Lazaridou, Alexander Psysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *Proceedings of the 2017 International Conference on Learning Representations*.

David Lewis. 1969. *Convention: A philosophical study*. Harvard University Press, Cambridge, MA.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132.

Jacob L Mey and Chenghuan Xu. 2001. Pragmatics: an introduction.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In Proceedings of NAACL-HLT, pages 3103–3114.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 29–61. Springer.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model.

Christopher Potts. 2012. Goal-driven answers in the cards dialogue corpus. In Proceedings of the 30th west coast conference on formal linguistics, pages 1–20. Cascadilla Proceedings Project.

Henry Prakken. 2006. Formal systems for persuasion dialogue. The knowledge engineering review, 21(2):163–188.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In Proceedings of the conference on empirical methods in natural language processing, pages 583–593. Association for Computational Linguistics.

Fariba Sadri, Francesca Toni, and Paolo Torroni. 2001. Dialogues for negotiation: agent varieties and dialogue sequences. In International Workshop on Agent Theories, Architectures, and Languages, pages 405–421. Springer.

John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. 1980. Speech act theory and pragmatics, volume 10. Springer.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In Thirtieth AAAI Conference on Artificial Intelligence.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In Thirty-First AAAI Conference on Artificial Intelligence.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. Nature, 550(7676):354.

Nathaniel J Smith, Noah Goodman, and Michael Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In Proceedings of NIPS, pages 3039–3047.

Victoria Talwar and Kang Lee. 2008. Social and cognitive correlates of children’s lying behavior. Child development, 79(4):866–881.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In Thirty-Second AAAI Conference on Artificial Intelligence.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In 25th USENIX Security Symposium (USENIX Security 16), pages 601–618.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. arXiv preprint arXiv:1708.04782.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013. Implications and nested beliefs in approximate decentralized-POMDPs. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 74–80.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3370–3380.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In Proceedings of NAACL-HLT, pages 575–581.
Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In Proceedings of the SIGDIAL 2013 Conference, pages 404–413.

Ludwig Wittgenstein. 1953. Philosophical investigations. Blackwell, Oxford.

Wei Emma Zhang, Quan Z. Sheng, A Alhazmi, and C Li. 2019. Adversarial attacks on deep learning models in natural language processing: A survey. arXiv preprint arXiv:1901.06796.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 1–10.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Thirty-Second AAAI Conference on Artificial Intelligence.