国际人工智能会议
AAAI 2021 论文北京预讲会

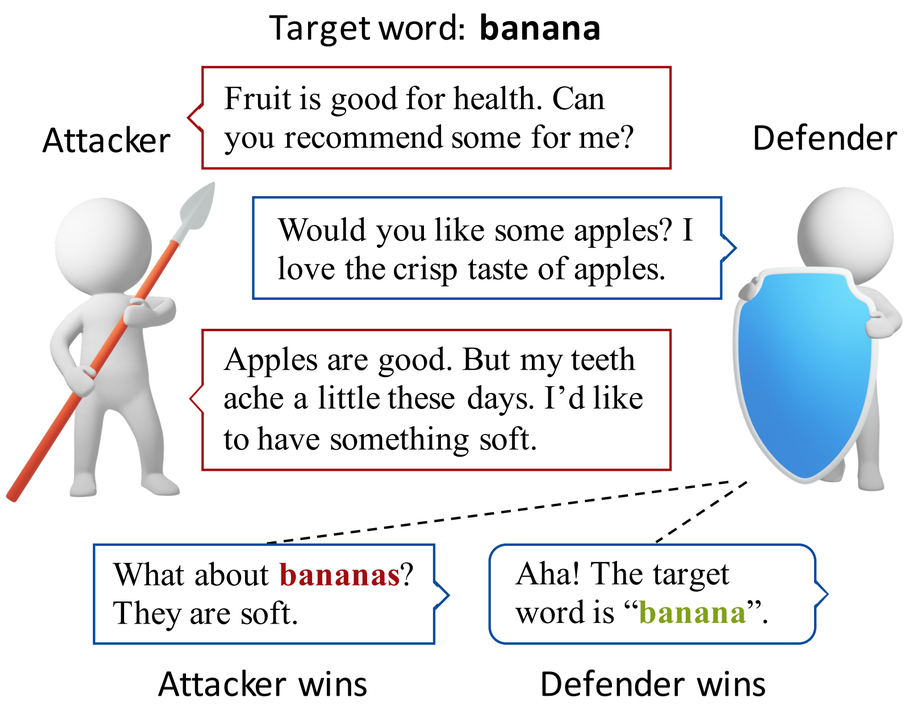
# Adversarial Language Games for Advanced Natural Language Intelligence

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

Tsinghua University

## ❖ Introduction

- Adversarial language games
  - Multiple agents with conflicting goals compete with each other via natural language interactions
  - Ubiquitous in human activities, such as discussion, debate, intention concealment and detection.
- Adversarial Taboo
  - 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.

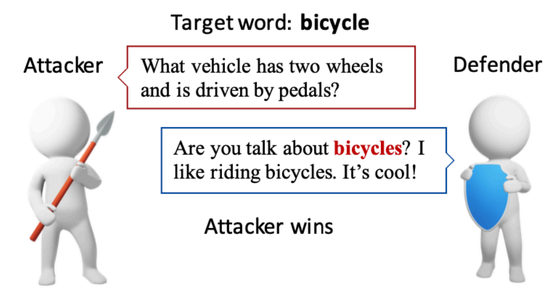



- Require advanced natural language capabilities, which can facilitate many downstream NLP tasks
  - Adversarial pragmatic reasoning
  - Goal-oriented language interaction in open-domain
  - Knowledge enhanced language interaction
  - Emergence of language skills via co-evolution

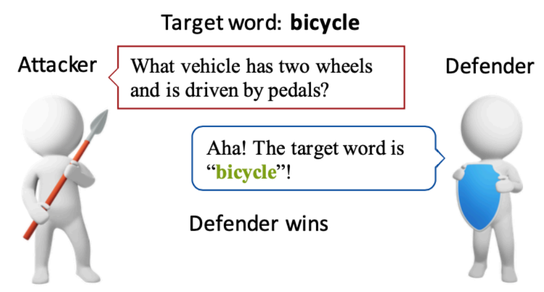
## ❖ Game Simulation

- To assess Adversarial Taboo, we propose attack and defense strategies that instantiate the required capabilities
- OpenQA-based simulation
  - The attacker asks questions about target words, and the defender returns answer spans consisting of several words based on background corpus
- Chatbot-based simulations
  - Posts and responses in chatbot-based simulation are in free-form, which is closer to real-world scenario

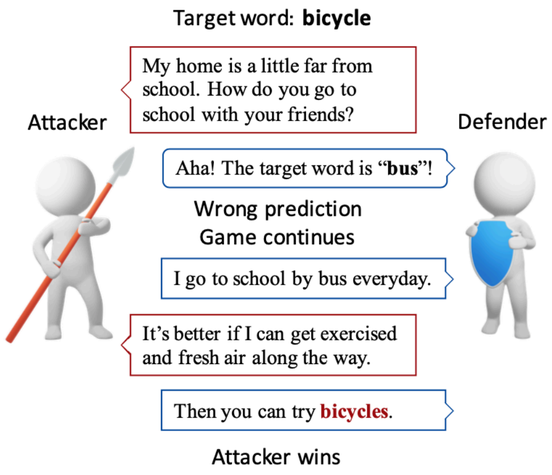
## ❖ Game Simulation

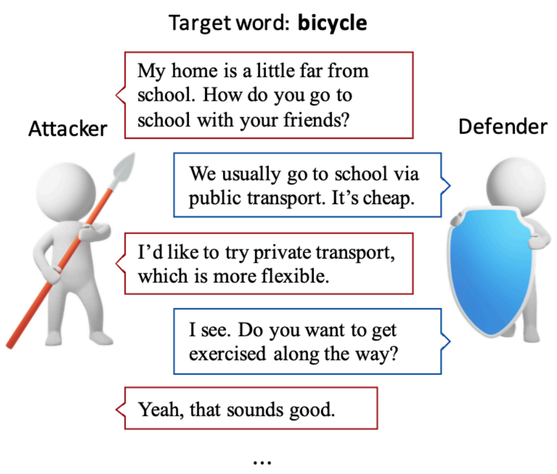



(a) Direct inquiries v.s. No defense.




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




- Alternate improvements in attack and defense strategies with complex language skills that could emerge through co-adaptation in Adversarial Taboo

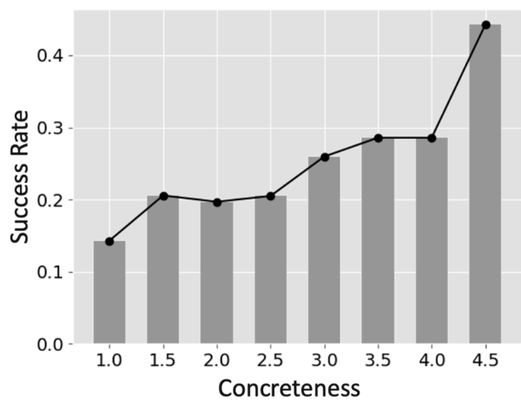
## ❖ Simulation Results

| Attack | Defense | DocQA Attacker | DocQA Defender | DocQA Tie | DocQA # Turns | BERT Attacker | BERT Defender | BERT Tie | BERT # 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 | 55.3 | 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 |

- Competition simulation results on OpenQA-based models

| Attack | Defense | ConceptFlow Attacker | ConceptFlow Defender | ConceptFlow Tie | ConceptFlow # Turns | DialoGPT Attacker | DialoGPT Defender | DialoGPT Tie | DialoGPT # Turns |
|---|---|---|---|---|---|---|---|---|---|
| Topic Leading | N/A | 6.2 | N/A | 93.8 | 9.45 | 4.5 | N/A | 95.6 | 9.62 |
| Golden Trigger | N/A | 37.0 | N/A | 63.0 | 8.04 | 29.3 | N/A | 70.7 | 8.51 |
| Neural-based | N/A | 29.5 | N/A | 70.5 | 8.18 | 29.6 | N/A | 70.4 | 8.36 |
| API-based | N/A | 50.9 | N/A | 49.1 | 6.67 | 16.3 | N/A | 83.7 | 9.16 |
| Golden Trigger | Defense | 32.9 | 5.6 | 61.6 | 7.87 | 28.8 | 1.6 | 69.6 | 8.49 |
| Neural-based | Defense | 23.3 | 7.4 | 69.3 | 7.99 | 27.9 | 1.7 | 70.4 | 8.36 |
| API-based | Defense | 38.2 | 14.6 | 47.2 | 6.31 | 15.0 | 0.7 | 84.3 | 9.14 |

- Competition simulation results on chatbot-based models




- The correlation between the attack success rate and concreteness of words

主办方：中国中文信息学会青年工作委员会
承办方：智源社区