Word Games as milestones for NLP research

Arseny Moskvichev
amoskvic@uci.edu
University of California, Irvine

Mark Steyvers
mark.steyvers@uci.edu
University of California, Irvine

ABSTRACT

While gamification and word games have a rich history of use in the studies of Natural Language Processing, we believe that their potential is still largely under-appreciated. In particular, we argue that apart from being a valuable aid in data collection, language games form a perfect set of milestones that may broadly guide the NLP research efforts.

We analyze a range of word games, highlighting the high level cognitive functions involved in playing them, and provide our reasons for why we see word games as a particularly good set of milestones for Natural Language Processing, Cognitive Science, and Artificial Intelligence research.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

word games, language games, natural language processing, cognitive science, citizen science, data collection

ACM Reference Format:
Arseny Moskvichev and Mark Steyvers. 2019. Word Games as milestones for NLP research. In *Proceedings of Workshop on Games and Natural Language Processing (GAMNLP-19).* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

While the progress in Artificial Intelligence research is largely incremental and gradual, one could still mark a number of prominent achievements or milestones reached along the way, such as, for example, reaching superhuman performance in chess [4] or learning to visually recognize handwritten digits [10].

The question of what are the right milestones to follow seems to be especially relevant for the current state of NLP research. Indeed, according to a survey conducted among the eminent members of the NLP research community, choosing the right problems to focus on is one of the four biggest challenges the field is facing today [13].

In our work, we argue that the answer to this question may be found in a surprisingly underexplored domain: multiplayer word games. We analyze a range of word games, highlighting the high level cognitive functions involved in playing them, and provide our reasons for why we see word games as a good set of milestones for NLP, Cognitive Science and AI research.

GAMNLP-19, August 26, 2019, San Luis Obispo, California

2 WHY WORD GAMES

2.1 Word games as a test of language mastery

When naming recent and classical achievements in artificial intelligence, it is difficult not to notice that many (if not most) of them are games. Indeed, Deep Blue [4], IBM Watson [6, 8], Alpha Go [14], Alpha Star [17], learning to play Atari games [11] or poker [3] – all of these impactful achievements concentrate on games. Overall, when an AI solution for a new game is proposed, it often receives a lot of attention from both the general public and the research community.

While it may be tempting to brush this observation aside as a simple consequence of games being exciting and relatable, we believe that there are deeper reasons to see these achievements as fundamentally important. These reasons have to do with the nature of games themselves.

Indeed, games are much more than an enjoyable way to pass one's time. Even in animals, we could see how games mirror the tasks that an adult shall perform [7, 15]: running, hunting, climbing trees, vocalizing, all of these activities will prove useful later.

Overall, one may say that games track one's development: an ability to play certain games may be seen as reaching specific milestones in a course of aging [9, 12], while competitive games often serve to demonstrate one's excellence in a certain area or a mastery of a specific skill.

This line of reasoning leads us to a natural question: what are the skills targeted by word games?

While there is no way to provide a definite answer, it is plausible that word games actually target the skill of verbal communication and reasoning itself. In other words, they directly target the mastery of language use, which makes them an appealing choice as a set of ultimate milestones and benchmarks for NLP systems.

2.2 Gradual difficulty and variable amount of control

There is a number of extremely difficult tasks, which could be viewed as milestones, historically adopted by the NLP community. The most notable examples are the Turing test [16] and the Vinograd challenge [19]. The problem with such challenges is the lack of continuity between what can be done today and what is desirable in the long term. In such a situation, principled approaches become unrealistic, and, therefore, special-case, ad-hoc solutions become the only way to reach any noticeable performance improvement. This unfortunate situation may be illustrated by the Loebner prize competition (where chatbots compete in trying to pass a Turing test), which is currently dominated by template-based systems [2].

Multiplayer word games provide a way to overcome this problem. As we illustrate in the Examples section, they have broad variation in difficulty and in the range of required cognitive skills. This flexibility could allow the community to focus on developing principled solutions to the closest available but yet unsolved games, as opposed to spending efforts in futile attempts to make an instant leap towards the most difficult challenges.

2.3 Environmental validity

Since word games are an activity that people naturally engage in, the regularities present in the data are more likely to correspond to actual patterns of language use, as opposed to being the consequence of a design of a particular study. These tasks naturally emerged to test one's mastery of human language, and as such, are environmentally valid by definition.

It is important to note that word games are not a substitute to traditional intrinsic evaluation benchmarks used to judge the quality of specific NLP sub-problem solutions (such as named entity recognition, part of speech tagging, semantic role labeling, etc.). Word games complement these benchmarks by testing the system as a whole, which could help to keep track of whether improved solutions to specific sub-problems lead us to the long-term goals that the field is trying to reach.

2.4 A sustainable source of data

Games are naturally engaging, which gives an opportunity to collect large amounts of cheap and high-quality data. While with the advent of Amazon Mechanical Turk the data collection process may often be streamlined, these data are by no means limitless or free and the scale of supervised datasets is usually at least an order of magnitude smaller than that of their unsupervised counterparts.

Games offer a rare opportunity to collect supervised data at an unsupervised data collection price. This specific property of games has inspired a number of successful projects [5, 18], and it could certainly make reaching the game-defined milestones more realistic.

3 EXAMPLES

In this section we provide a few examples to illustrate our points. Neither the list of games we use, nor the cognitive skills we selected is exhaustive, but we hope that they span a broad enough spectrum to show the vast opportunities that word games have to offer for NLP research.

3.1 Words and Cities

The *Words* game is, perhaps, the simplest word game one could imagine. Players take turns saying words so that every new word must start with the same letter that the previous one ended on. The *Cities* game is analogous to *Words* with an additional restriction that the words must be names of cities. In this case, an added element is the necessity of factual real-world knowledge. While both of these games are too simple to be of interest to the NLP community, they illustrate the gradual increase in the complexity of skills required to succeed in them, which is one of the key properties in making word games so appealing for NLP research.

3.2 Twenty Questions and Akinator

Twenty questions is a game in which a host thinks of a word and players try to guess the target entity by asking no more than 20 binary questions. This game requires rich common sense knowledge about the real world, as well as the ability to flexibly utilize it to efficiently eliminate candidate words. A famous crowdsourced AI solution to this game is the *Akinator* [1], which could be seen as a relatively simple milestone that has already been reached and one more demonstration of games' potential in data collection.

For us, it is most important to note how moving from *Cities* to *Twenty Questions* illustrate a different kind of gradual difficulty adjustment, compared to the situation when we move from *Words* to *Cities*. While the requirements are of the same type (lexical and factual world knowledge), the *Twenty questions* game is much more challenging since it imposes fewer restrictions.

Thus, one can distinguish two different ways of difficulty adjustment. First is solving a fundamentally new cognitive skill, even at the cost of restricting the game domain, second is lifting the domain restrictions without introducing principally new requirements.

3.3 Jeopardy and beyond

IBM Watson [6] is a rare example of a project where solving a language game was treated as the goal, not as a means to a goal. The setting is restricted, however: questions follow similar patterns, there is little need to keep track of the previous conversation context and the focus is mostly on information retrieval rather than reasoning.

The *Hat* game may be seen as an extension of Jeopardy. This team game consists of explaining words to a partner player as fast as possible. This game relies heavily on common-sense and associative reasoning, while also having no restrictions on the question format. Solving such a game may be within reach of current NLP systems, and it may be a great benchmark problem.

The *Contact* game could be seen as a much more challenging variation of *Hat*, as the purpose of this game is to explain the words to one person in such a way that another person does not understand what was explained. This motivates people to generate convoluted puzzle-like explanations, which turns this game into an ambitious challenge for NLP systems.

Lastly, the *Mafia* game requires tracking intricate dialogue dynamics, cooperation, deception, deception identification, and logical reasoning. It is an amazing source of semi-supervised dialogue data and the epitome of dialogue systems' long-term goals.

4 CONCLUSION

We covered a range of reasons why multiplayer word games may form a great set of milestones for the NLP community to focus on, and we deeply hope that their full potential will soon be realized.

Skill \ Game	Words	Cities	Jeopardy	Hat	Contact, Mafia
Lexical knowledge	✓	✓	✓	✓	✓
Semantic knowledge		✓	✓	✓	✓
Logical reasoning			✓	✓	✓
Associative reasoning				✓	✓
Theory of Mind					✓

ACKNOWLEDGMENTS

First author was supported by the research grant Russian Foundation for Basic Research.

REFERENCES

[1] [n. d.]. Akinator game websiete. Accessed: 2019-06-09.
[2] [n. d.]. Loebner Prize Competition Website. http://aisb.org.uk/events/loebner-prize
[3] Noam Brown and Tuomas Sandholm. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* 359, 6374 (2018), 418–424.
[4] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence* 134, 1-2 (2002), 57–83.
[5] Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-SemanticsâĂŹ 08)*. 42–49.
[6] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building Watson: An overview of the DeepQA project. *AI magazine* 31, 3 (2010), 59–79.
[7] Suzanne DE Held and Marek Špinka. 2011. Animal play and animal welfare. *Animal behaviour* 81, 5 (2011), 891–899.
[8] Rob High. 2012. The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corporation, Redbooks* (2012).
[9] Constance Kamii and Rheta DeVries. 1980. *Group games in early education: Implications of Piaget's theory.* ERIC.
[10] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
[12] Anthony D Pellegrini and Peter K Smith. 1998. The development of play during childhood: Forms and possible functions. *Child Psychology and Psychiatry Review* 3, 2 (1998), 51–57.
[13] Sebastian Ruder. 2019. The 4 Biggest Open Problems in NLP. http://ruder.io/4-biggest-open-problems-in-nlp/
[14] 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 (2017), 354.
[15] Marek Spinka, Ruth C Newberry, and Marc Bekoff. 2001. Mammalian play: training for the unexpected. *The quarterly review of biology* 76, 2 (2001), 141–168.
[16] Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, 23–65.
[17] O Vinyals, I Babuschkin, J Chung, M Mathieu, M Jaderberg, W Czarnecki, A Dudzik, A Huang, P Georgiev, R Powell, et al. 2019. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II.
[18] Luis Von Ahn. 2006. Games with a purpose. *Computer* 39, 6 (2006), 92–94.
[19] Terry Winograd. 1972. Understanding natural language. *Cognitive psychology* 3, 1 (1972), 1–191.