2025 EAAI Mentored Undergraduate Research Challenge:
Playing Word Association Games

Richard G. Freedman (Smart Information Flow Technologies (SIFT); rfreedman@sift.net)
DOI: 10.1145/3655032.3655035

Abstract

The topic for EAAI 2025’s Mentored Undergraduate Research Challenge is Playing Word Association Games. What does that mean? Where are the applications? How can you get started? We break down the topic, discuss applications, and explore project ideas in this column.

Introduction

The EAAI Mentored Undergraduate Research Challenge invites undergraduate students to team up with research mentors and participate in an artificial intelligence (AI) research project from start to finish. For 2025, the specific challenge is for students to research and develop a bot that can play a word association game similar to Apples-to-Apples™ and a number of other board games. The recent innovations in large language models (LLMs) has created quite the buzz, but there has been a long history of research in AI for natural language understanding and information retrieval. Furthermore, games with a social component such as word association games introduce additional challenges that involve understanding other players beyond their strategies. How can computational systems use AI to think about others’ preferences and interests? How should a bot make decisions that align with its own artificial persona? These are just a few project ideas that relate to the challenge topic, but they are far from all the possibilities.

Not sure where to start? Do not worry, this column has you covered! We break down the buzzwords, clarify the concepts, and dig into the details so that you can identify your perfect research project for the challenge. After describing the word association game, we give a brief overview of related research areas that might be helpful for getting started. In particular, we go over some natural language and information retrieval methods to introduce possible starting approaches for making word associations. Then, we discuss some cognitive modeling efforts to get you thinking about how your bot can play appropriately with other people and/or bots. This column ends with more details about forming a team and registering for the 2025 EAAI Mentored Undergraduate Research Challenge. We look forward to seeing all the cool and creative projects you come up with for the challenge—happy researching!

What are Word Association Games?

Word association games typically involve players choosing a word that has some relation to a specified target word, phrase, image, etc. (we will use the term “target”). Improvisation activities and ice-breakers at events lack structure and simply challenge people to say whatever comes to mind that makes sense. However, a number of board games have added structure to this social activity in the form of cards containing words and targets. Players take turns in the role of judge selecting a card with a target (usually at random) and deciding a winner for the “best” associated word to the chosen target. The other players have a collection of cards with words in their hand, and each player must choose one of those cards to submit to the judge as their associated word for the target. In an attempt to keep submissions anonymous, the players submit their cards face-down and the judge shuffles the submissions so that they do not know who submitted each word card. The choice of “best” is in quotes because the player in the current judge role has no rules about how to select a winner for the turn, which means players submitting a card with a word on it need to consider what the judge would label a good choice.

Regardless of the role a player has each turn, they will select one word card from an array of word cards to be the “best” word association with the target card. For the players not in the judge role, their personal hand of word cards...
represent that array. All those players’ chosen word cards serve as the array for the player in the judge role; the judge does not interact with cards from their hand. The greatest difference between the roles is the criteria for selecting the “best” word card because the player in the judge role defines “best.” Each player not in the judge role needs to figure out what the judge will define as “best,” and the player in the judge role needs to apply their definition of “best.” If the players all know each other, then understanding each player’s definition of “best” for word associations might come easily from past experiences and familiarity with their personality—an inside joke might become the “best” word association even if the target card and word card typically have little to do with each other. When the players are less familiar with each other, then it is important to observe which word cards the player selects while they are judge—do the judge’s decisions imply any preferences, biases, or trends besides the association of the words? If you are the judge that turn, then what criteria do you apply when selecting the “best” word association for the target?

For teams who want to see their bots play together in this year’s challenge, we will use word cards and target cards that resemble the word association game *Apples-to-Apples™* (Mattel, 2008). Our word cards will list one noun (no proper nouns) per card, and each target card will list one adjective. The list of available words and targets could vary between games similar to how different editions of a game will have its own unique cards, but we will provide a default list for teams who wish to have their bot play with other teams’ bots in case their word association selection method requires a precomputed ontology, knowledge base, or machine learning model. What do these have to do with making decisions about word associations? Keep reading onwards to find out!

### How Could AI Associate Words?

Information retrieval studies how to generate and apply knowledge, and its applications are all around us from search engines (Croft, Metzler, & Strohman, 2009) to finding related works (Bela & Jörn, 2009). Natural language understanding studies how our textual content translates to concepts for the purposes of knowledge, communication, and more. If you are wondering how these relate to playing word association games, it comes down to handling the word association tasks more than playing the game itself. Say the target word is “sticky;” what sort of words go with sticky? How did you make those connections between the words? How could a machine choose those words and/or make similar connections?

One way to think about it is how frequently the words appear together because adjectives are usually near the nouns they describe: a sticky situation, some sticky glue, a tree branch with sticky sap, and so on. *N-grams* count how often certain words appear together out of all possible sequences of *n* words (Jurafsky & Martin, 2000). Despite how simple this sounds, text prediction for autocompletion has had success using *n*-grams (Yazdani, Safdari, Golkar, & Niakan Kalhori, 2019). You can also consider multiple values of *n*, but keep in mind that words that are nearby do not always imply correlation. Negation terms such as ‘not’ might appear in your *n*-grams, but the context of negation does not count against those words being nearby.

Another thing to consider is whether the words share a common theme. A popular collection of methods for this during the 2000s and early 2010s was topic modeling (Blei, Ng, & Jordan, 2003), and more recent approaches consider variational autoencoders (Srivastava & Sutton, 2017). In general, these methods aim to group words that appear together frequently in related text documents or snippets. For example, one group might contain sticky alongside tape, glue, attach, and blob; another group might contain sticky alongside mud, sap, honey, goo, and guano. The former group could be about arts and crafts while the latter seems to focus on natural things from the great outdoors. We assume this from our experiences with the words and their definitions, but it is not guaranteed to be accurate because the machine lacks the definitions and themes of the source material. Topic models assume that the theme is missing information that justifies the identified group of words.

In fact, LLMs have much in common with these *latent semantic analysis* approaches
Landauer, 1998) given the patterns that these methods try to find both involve hidden variables. Like a known unknown, we define the hidden variables as “something is responsible for a correlation that we observe, but we do not quite know what that something is to directly observe it.” As humans designing these model structures, we might have a hunch from our experiences, like figuring out the topics describing each cluster of words above. However, the statistical methods and/or learned models lack the same context such that the most meaning these hidden variables have to machines are values whose assignment contribute to the improved conditional likelihood, utility function, etc. The people who wrote the original documents knew the context influencing what they wrote, but the learning algorithm only has access to the text output without the authors’ thought process and inspirational context. Thus, the machine invents its own patterns to mimic the observations regardless of the true process that generated the original (Shanahan, McDonell, & Reynolds, 2023)—this is why some call LLMs “parrots” (Bender, Gebru, McMillan-Major, & Shmitchell, 2021) rather than thinking for themselves (Wei et al., 2022).

Explicit semantic analysis aims to address the concern of no context by generating definitions. The earliest work of which the author is aware used the titles of Wikipedia articles as features, creating a vector measuring how frequently each word appears in the respective article (Gabrilovich & Markovitch, 2009). Even if the machine does not know what the titles’ words mean semantically, there is a correlation between grounded concepts that we can inspect and understand as people. If all the words in a sentence have a reasonable amount of presence in one of the Wikipedia articles, then we can associate what those words have in common through that feature name (Wikipedia article’s title) rather than with a mysterious random variable and its assigned values. Other examples include query facets (Kong & Allan, 2013) and Bayesian Case Models (Kim, Rudin, & Shah, 2014).

**How Could AI Consider Others and Itself?**

In the earlier days of AI research, game-playing research considered traditional strategy games where everyone is an adversary competing to win through making “better” decisions. The adaptation of game theoretic methods from economics was a key approach because playing these types of games satisfied many of the assumptions underlying game theory, especially that everyone plays rationally to optimize their outcome given that everyone else is also optimizing their outcome (Tadelis, 2013). Most the research considered state representations, heuristics, algorithms, and shortcuts that could handle a computer’s limited resources when playing games live because these applications were not off-line studies of behavior: time limits and memory constraints matter (Russell & Wefald, 2003). More recent game playing research in AI for Poker (Moravčík et al., 2017) and Hanabi (Bard et al., 2020) add strategies that deal with the complexity of reasoning over partially observable information and nondeterminism not just on the game state, but also within the other players’ minds for their strategies and private knowledge about the game state.

The majority of these studied games favor computational power as sufficient criteria for scaling good performance. That is, increasing resources to make more calculations over less time often improves game-playing performance. Moravec’s paradox (Moravec, 1990) points out that computers can outperform people in such reasoning tasks because they are designed to compute things (what a surprise), but people are more competent than machines at many other tasks that do not have such formulaic solutions. This is why vision, signal processing, and natural language have been popular areas of study in AI with so much potential for discovery! We take our abilities for granted, but machines still struggle to do these tasks well even with the rapid performance improvements that came from deep learning (Waldrop, 2019). With this motivation, let’s consider why game theory might not help us as much with word association games.

The biggest issue is defining the outcomes of actions because they are not payoffs in the traditional sense. We cannot simply assign point
values to every card combination because the word association validity is one of two criteria in scoring. The changing judge introduces the other criteria because each judge has a personal evaluation for every word association, and their perspective likely differs from yours. This comes from many factors including cultural background (Kirby, Dowman, & Griffiths, 2007), life experiences, knowledge of vocabulary, and other qualitative factors. There is no universal quantitative measure for which words go better with certain targets. Someone might associate sticky with dog because of a childhood memory where their pet dog rolled in the mud and got leaves stuck to its fur, but another person might find no reason for a dog to be sticky without such a moment in their past. Worst case, a judge who wishes to be a troll and violate the players' expectations might select the card whose word is least associated with the target or even pick a card at random—yikes!

If we know the judge's personal evaluation of each word compared to the target word, then we can start to make some informed decisions. The optimal move is to play the card from your hand whose word has that judge's greatest score with respect to the target because the judge will perceive it to have the greatest value among your hand's options. Next, to break ties when hands have multiple cards that are each worth the greatest score for that judge, you might consider which of those words is more flexible in general. The expected value of a word's association with possible targets and possible judges implies whether that card is niche or likely usable for later rounds of play. In such a case when the immediate value is relatively equal, the long-term value might indicate a card to keep for later so that it is rational to play the more niche card that has the lesser expected payoffs for future situations. Of course, we could delve deeper into considering which cards were already played to know when you can safely play a less associated niche card because it would score enough to win without the competition, etc. However, the cards that other players have around the table are random with no obvious clues regarding the words against which your cards' words are competing. You also do not always know how familiar other players are with each other to have an advantage on knowing the current judge's personal evaluation of words compared to the target.

Regardless of whether your bot tries to do all these complex computations for playing a word association game, in which ways could the bot consider how much a particular judge evaluates the card values? Preference modeling (Rossi, Venable, & Walsh, 2011) and theory of mind (Langley, Cirstea, Cuzzolin, & Sahakian, 2022) each study how an agent thinks about what other agents do and want. There are many ways for one agent to represent what it believes is inside another agent's mind. When an agent like your bot observes another agent make a choice, then your bot can store information about that choice and what that choice implies. For example, if the player in the judge role chooses 'candy' for sticky, does your bot decide that the judge simply believes candy is a better match than the other words? Does this particular judge prefer sweets and confectionaries compared to the other cards' words' categories? Until there is more information, the bot can keep track of various hypotheses (Hutson, 2023). Bayesian updates on a prior distribution is one way to keep track of these hypotheses through accumulating the observations that support each one (McCann, 2020).

Flipping the tasks around, what should your bot choose to do when it plays the role of judge? While it might sound strange to design a personality with preferences for your bot because it is a computer program, this adds an important factor for a social game like this one. In fact, the chatbot Eugene Goostman (Aamoth, 2014) was the first of its kind to pass a formal Turing Test because the content Eugene generated aligned with the specified personality. How did choppy English sentences (this was before LLMs) pass as reasonable behavior? The developers made the bot's origin be from a country where English is not the native language, and they set the bot's age young enough to not have a chance of learning to speak English with fluent proficiency. Training the bot on content that aligns with the specified interests and preferences further cemented what it knew when engaging in conversations. When Brian Christian participated in the Turing Test as one of the humans convincing judges that he was not a chatbot, he realized these sort of properties,
from personality to genuine experiences, were what made him different from the machines (Christian, 2012). This is not advocacy for your word association game-playing bot to go compete in the Turing Test, but it does provide context for things you can consider that will make your bot unique when it is the judge—what do other bots and human players need to figure out about your bot in order to choose favorable cards from their hand? How will your bot communicate these trends based on the cards other players present for association with the target?

One Last Thing: Being Transparent!

Although we can throw together some code and use a few AI algorithms to create a bot that plays word association games, it is important to keep in mind that there are additional considerations. Even if the bot you create can make word associations and consider other people’s preferences, how do you know that it is playing the word association game well and consistently? Do you trust that your algorithms, data (if using machine learning-based methods), and knowledge representations are working consistently? If so, then why? Is it good enough that you believe it works because of the code or some extensive testing? Will that justification be good enough for other people?

Many of them will not be able to read or write code themselves, and simply claiming that our systems work is not the most comforting reassurance. In fact, biases in datasets (Mehrab, Morstatter, Saxena, Lerman, & Galstyan, 2021), the risks of human error in programming, and malicious actors taking advantage of exploits in knowledge representation models (Microsoft, 2016; Ren, Zheng, Qin, & Liu, 2020) give users even fewer reasons to be comfortable with intelligent systems that act like black boxes. From our own perspective as AI researchers, if we do not fully understand what is going on, then can we guarantee consistent performance? If not, when will our bot fail through making a “wrong” decision? Everyone makes mistakes, but understanding our mistakes is what allows us to improve—our intelligent systems currently do not have such capabilities in most the traditional AI algorithms. While reinforcement learning can give out negative rewards to dissuade repeating the same bad decision in the future, this assumes the reward function is accurate and properly defined (Amodei et al., 2016; Griffith, Subramanian, Scholz, Isbell, & Thomaz, 2013). Overall, everyone has stakes in reliable AI systems, and we need a means of providing evidence that supports reliability.

Transparency is a growing area of interest in AI to provide that evidence, and it is a responsibility we should keep in mind as we develop intelligent systems that others will have to use in their everyday lives. Two of the more popular means of transparency include explainability (Gunning & Aha, 2019) and interpretability (Rudin, 2019). Explanations are more post-hoc such that the AI system provides a human-understandable justification for the decisions it makes; this answers questions such as “why did you do that?” Right or wrong, a justification gives people the opportunity to assess how the algorithms, data, and knowledge play a role in generating output. In contrast, interpretability is more pre-hoc such that the AI system’s information and processes are human-understandable for inspection before use; this answers questions such as “if this were to happen, what would you do?” Interpretable models and methods that are sufficiently summarized through text and/or visualizations enable people to explore the capabilities and edge cases of the system in advance of general use. Which of these approaches do you prefer for getting to know how your bot actually plays word association games?

Transparency is by no means restricted to just explanations and interpretable models, though. One method that IBM’s Watson used while playing Jeopardy (Ferrucci et al., 2010) is a confidence score. The author of this column remembers being an undergraduate student just learning about AI research at this time, and Watson was a huge inspiration as it correctly answered questions consistently and faster than the human champions. Alongside its response, Watson listed the top three answers ranked with a confidence score (greater implies more confidence in that response). This mattered most on the Final Jeopardy challenge asking about a city in the United States because Watson replied “What is Toronto?,” which is a city in Canada (IBM, 2011). Although we do not know what infor-
Watson consulted nor how it used that information, we do know that Watson’s confidence in this response was absurdly low and just barely higher than the next-most-confident responses. Suddenly, we have a way to represent “I do not know” even if the AI system is not programmed to acknowledge that it is struggling to find the answer, and a person can accept “I do not know” as an indicator to ignore this particular output. This is more reassuring than thinking that Watson is confident in its incorrect response.

Other Project Considerations

In addition to the research challenges described above, there are a few alternative projects that teams may consider that align with this year’s challenge. Special thank you’s to those listed in the acknowledgments below for their thoughtful contributions on these. First, because LLMs are very popular right now, what roles can they play? Relying on prompt engineering for the bot’s decision-making during gameplay is a rather dangerous practice given the lack of transparency behind how they work, bias in training documents (Acerbi & Stubbersfield, 2023; Kotek, Dockum, & Sun, 2023), and potential to hallucinate incorrect information or justifications (Smith, 2023; IBM, 2023). However, using them as a text generator can be a beneficial means of dataset creation or augmentation. If you need example sentences with ‘sticky’ in order to establish its n-grams, generate topic models, or retrieve some other information from natural language, then the LLM can produce those sentences based on the documents it received for training that included ‘sticky’ or related words. Whether the generated sentence is an accurate statement or not does not matter as much when you care about the combination of words in the sentence. Second, there are plenty of other word association games besides Apples-to-Apples. Most of them continue to use the described structure above of “choose a word from a list of options that best relates to a target, and then the judge decides which of those choices is the most related,” but they can have some nuances to the list of options, targets, and judges that lead to different games. If your team identifies an alternative word association game that introduces different challenges for developing a bot that can play it, then feel free to share your experiences on creating a bot to play that word association game!

Participating in the Challenge

The EAAI 2025 Mentored Undergraduate Research Challenge invites teams of students and mentors to work together on a research project involving playing word association games—the goal is to complete a feasible project and submit a paper about the research to EAAI 2025. As these projects can become ambitious, it is important that students focus on one idea of interest and think of a simple task within that idea. Mentors are expected to be involved as guides for the students to evaluate feasibility, provide tips and ideas, and teach the research pipeline from observation and ideation, to the scientific method, to presenting results in a full paper. To provide a starting point for teams, we plan to make available code for a game server and a simple bot client in several programming languages so that students can focus on the AI research even if they are not familiar with game-playing bots or a specific programming language. If you create a bot client in another programming language, we will be happy to include it for others who might consider that language as well. Such resources and up-to-date information about the challenge will be available at https://www.yetanotherfreedman.com/resources/challenge_pwag.html.

If you are interested in AI techniques that think about finding the best word matches based on the players around the table, then we encourage you to consider participating in this challenge. Please make sure to form a team that meets the following requirements:

- At least one undergraduate researcher who has not completed a post-secondary education degree. Students in community college are also eligible for this role. Students in this category are expected to play a significant role in the project.
- At least one mentor actively engaged in research who either received a Ph.D. or works in a position involving undergraduate/community college student mentorship. This can be a faculty member at an aca-
academic institution, a postdoctoral fellow, or a researcher in industry who has experience training undergraduate students in research. Mentors in this category are expected to be involved with the students regularly to guide them along their journey.

- As long as the above two roles are satisfied, additional team members are allowed. Additional members may include graduate students as long as the undergraduate researchers are actively involved in the research process (ideas, experiment design, paper writing, etc.). Graduate students may also provide additional mentorship to the undergraduate students, but they cannot serve as a substitution for the mentor’s participation.

Once you form a team, please contact the author of this column with the names, e-mail addresses, and roles (mentor, undergraduate student, etc.) of all team members to register your team in the challenge.

There are no limits to team sizes or number of teams per institution. However, due to conference logistics, there will be a limit on the number of accepted papers for publication and presentation at EAAI 2025. All submitted manuscripts will undergo peer review for writing quality, evidence of quality research at the undergraduate level, and relevance to the topic of playing word association games. We look forward to seeing your exciting and creative research on this topic!

Acknowledgments

The author would like to thank Todd W. Neller for establishing the Mentored Undergraduate Research Challenge and all his efforts organizing past challenges; his time and guidance while preparing for the challenge since 2023 has been invaluable. The author also thanks Calin Anton, Justin Stevens, and Jason R. Wilson for their discussions and feedback on this year’s topic and challenge details.

References


Christian, B. (2012). The most human human:
What artificial intelligence teaches us about being alive. USA: Anchor Books.


Richard G. Freedman is a Researcher at SIFT. In his quest to understand how machines can more naturally interact and play with people, Freedman studies the intersection of artificial intelligence and human-computer/robot interaction. Freedman enjoys sharing opportunities that inspired him, introducing STEAM to K-12 and research to undergraduates.