CS7637 HW3

Chenxi Gao
cgao71@gatech.edu

QUESTION 1:

Unlike case-based reasoning, analogical reasoning involves mapping and transferring the relationship between the two models. Mapping involves finding the correspondence of the objects in the source and the target problem. Transfer involves finding relationships by abstracting the relationship between the objects in the source problem and then adapt the same relationship to the target problem.

Although none of the features or objects in the human brain and the anthill are the same, there are deep relationships within each model's objects that are very similar to each other. First, I will construct a simple model of a human brain and an anthill and then discuss what can be analogically mapped from one model to the other.

Human brain:

1. The neurons work together to perform tasks and achieve goals.
2. The collection of all neurons is called a brain.
3. Each neuron communicates to the surrounding neurons through electrical and chemical signals through a thin fiber called axon. (MIT, 2019)
4. Neurons are the most fundamental unit of the nervous system. (MIT, 2019)
5. Each neuron may have specialized tasks, such as memory or actions.
6. The brain can produce new neurons through a process called Neurogenesis, where some cells can divide to form new cells. (Queensland, 2017)
7. The death of several neurons insignificantly affects the overall performance of the brain.

Ant colony:

1. The ant colony makes decisions and work collaboratively.

2. The collection of all individual ants is called an ant colony.
3. Ants have different roles, and each role performs different tasks.
4. Ant communicates to each other through pheromones, sound, and touch. (Reynolds)
5. An ant is the fundamental unit of an ant colony.
6. The ant queens lay eggs to reproduce and expand the colony.
7. The death of several ants in the colony does not significantly impact the colony as a whole.

I decided that an ant is analogical to a neuron, and an anthill is analogical to the brain. In this case, many deep relationships can be analogically transferred from one model to the other or the other way around. The following are examples of such relationships.

1. The "unit" and "collection" relationship can be transferred to both models as follows: an ant is a unit of an ant colony, and a neuron is a unit of a brain. The collection of all ants is an ant colony, and the collection of all neurons is called a brain. This is a kind of structural similarity.
2. The "role" relationship can be transferred to both models. Ants have specialized roles: ant queens lay eggs, and ant workers build colony and find foods. Neurons have specialized roles as well: some are memory cells, some cells help acquire senses like vision and olfactory, some are muscular cells that controls your actions. This is a kind of structural similarity.
3. The "collaboration" relationship can be transferred to both cases. The brain cells work together to achieve certain goals like memory, sensing and action. The ants works together to achieve certain goals as well like building tunnels and finding food. This is a kind of pragmatic similarity.
4. The "reproduction" and "death" relationships are similar and can be transferred to both cases. Both anthill and the brain has method to reproduce offsprings by dividing cells and laying eggs, respectively. The death of individuals, an ant or a neuron, will not significantly impact the anthill or the brain, respectively. This is a kind of semantic similarity.

I do not think an anthill is conscious philosophically, although there are many similarities between an anthill and a brain. There are some obvious differences. For example, the ants can be moving around in the anthill, but the neurons are

at a fixed location in the brain. Biologically, each ant has a brain. Therefore, each individual in the ant colony has free will, and I conclude that each individual in the anthill is conscious because it has a brain. However, the anthill itself is not conscious because it does not have a sense of "free will" like a brain does. Therefore, a collection of conscious individuals is not conscious.

For example, the anthill is also pragmatically similar to an army, where the ants are the soldiers, and the anthill is the army. The army's goal is to defeat the enemy, and every soldier works on his/her part collaboratively toward the goal. Each soldier is conscious. He/she can decide what to do next, but the army itself is not conscious.

QUESTION 2:

The first paper I chose is *The Dark Side of Ethical Robots* by Dieter Vanderelst and Alan Winfield. This paper addresses some inherited limitations of emerging ethical robots. Building ethical robots can be easily corrupted because it is easy to modify an ethical robot to an unethical one within a few lines of code. In the end, it illustrates concerns about how unscrupulous actors may result in a high risk to affect the robot's decision making in real-world safety critical robots, and thus to cause serious outcomes. Below are a summary and some keynotes of the paper.

1. The development of ethical robots is crucial when the robot is making inevitable moral real-world decisions. For example, the self-driving car must morally decide what to do before an inevitable crash would happen.
2. "A Ethical Layer ensures robots behave according to a predetermined set of ethical rules by (1) predicting the outcomes of possible actions and (2) evaluating the predicted outcomes against those rules."
3. This paper records three experiments. The robot uses the ethical layer describe above to make decisions to assist the human or not.
4. The first experiment concludes that robots can behave ethically.
5. The second experiment concludes that robots can behave egoistically after changing a single line of code to maximize its own takings.
6. The third experiment concludes that it is very easy to turn an ethical robot to an aggressive or even malicious one by changing a couple lines of code.

7. Conclusion: developing an ethical robot inevitably opening a Pandora's box of unethical robots.
8. The fact that it is easy to turn an ethical robot into an unethical one is dangerous.

I agree with the author that it is initially hard to build an ethical robot, but it is relatively very easy to turn an ethical robot to a malicious one. As an ethical robot, it always tries to minimize the cost of the outcome if all potential outcomes are undesirable. However, changing the minimizing problem to a maximizing problem can be very easily done in a few lines of code. In this case, the robot performs maliciously to maximize the cost. When we apply ethical robots to some real-world applications, this becomes extremely dangerous when it involves the safety of humans. Some examples include self-driving cars, self-driving planes, and AI surgeon robots.

I came up with the follow-up ethical questions after reading this paper: Because it is easy to turn an ethical robot into a malicious one, how can we ensure the safety of the moral settings? How to properly set the moral standard?

To answer the first question, I propose that the system should have a sanity check that is hard-coded and unchangeable before it performs any tasks. In the second question, I think computer scientists should collaborate with philosophers with different backgrounds when setting up the moral standards to accommodate most people with different backgrounds and cultures.

The second paper I chose is *A Computational Model of Commonsense Moral Decision Making*. This paper introduces a new computational model of moral decision making and characterizes the social structures of individuals and groups as a hierarchical Bayesian model. Below are a summary and keynotes of the paper:

1. "Humans learn to make ethical decisions by acquiring abstract moral principles through observation and interaction with other humans in their environment." (KleimanWeiner, Saxe, and Tenenbaum, 2017)
2. In this theory, ethical decisions are measure as utility-maximizing choice over a set of outcomes.
3. The paper suggest the moral dilemmas as a utility function that computes the trade-offs of values perceived by humans in the choices of the dilemma.

4. This paper characterizes the social structure of individuals and groups as a hierarchical Bayesian model.
5. The model can rapidly infer moral principles of individuals from limited number of observational data.
6. Demonstrating the model's capacity to rapidly infer group's norms, characterized as prior over individual moral preferences. Inferring shared moral values of a group is an important step towards designing an AI agent that makes socially optimal choices.

This paper uses several models to classify and quantify morality to make binary-decisions. I agree with the article that robots need to use models to make moral decisions because the machine itself is not conscious. It cannot learn morally what is right and what is wrong. Therefore, robots need to evaluate the risks and costs when an accident is inevitable. In the self-driving car example, the model classifies humans and objects into different groups such as male or female, young or old, fat or fit, assigning a value to denotes the importance based on their group. After that, it use Hierarchical Moral Naïve Bayes approach to evaluate the corresponding values and thus predicts the outcomes based on the weights. Although I agree with using models to make decisions, these values should not be universal because people have different moral standards. Because ethnicity itself is ambiguous and controversial sometimes, it is clearly hard to define the initial group values or what should be prioritized. Therefore, I have the following question:

> Because different people from different country, race, cultural background may have a different standard of ethnicity, how to evaluate and quantize the morality in algorithms to accommodate this?

I think developers may allow users to customize the moral settings in a pre-defined range, similar to customize your phone settings. This might sound weird, but at least the customers should know what kind of ethical settings are embedded in the system before using the product.

QUESTION 3

The first paper I chose is *AI Stories: An Interactive Narrative System for Children* by Ben Burtenshaw. This paper proposes an interactive dialogue system, which lets the children co-create narrative worlds through the conversation with the AI agent. Below are the summary of this paper and some key points.

1. AI Stories is an applied Computational Creativity and Natural Language Processing project to develop an interactive dialogue system for children.
2. Language play through dialogue is a fundamental process in the development of a child's language skills, through the creation of fictional worlds children learn how to use language.
3. Children experience a rich and changing space of narrative, where they can actively engage in their own interests. AI Stories would encase this established process within a usable and friendly software.
4. A completed AI Stories system will use dialogue to create a congruent and multi-tonal story, and draw on online information to enrich the narrative.
5. One has to respond in a way that maintains their interest, but must also maintain the narrative trajectory.
6. This paper proposes five systems to create complex and interesting non-sporadic or boring conversations: subsystems, topic based system, context based system, poetry and humor based system, and the selector.

Here is what I found interesting about this paper:

1. The paper proposes that AI technologies can take the role of parents to teach children language by setting up conversations or telling stories to children.
2. One of the systems the paper propose is called the poetry and humor system, which uses uses template sentences, rhyme, and word definitions to generate humorous and entertaining responses to maintain the user's interest.
3. It has a selector system to decide which of the proposed sentences would maintain the conversation. It does this using a combination of reinforcement learning and lexical analysis.

Some potential weaknesses are:

1. This proposal seems working on adults, which provides a conversation and storytelling system to maintain user interest. However, this might still not work for children because they are in the stage of learning a new language. Additionally, it is hard to tell what a child likes or dislikes because the child is also in the stage of learning emotions and how to express his/her feelings. The selector might perform incorrectly to predict children's preferences.
2. AI storytelling agent might reduce the time for parents to communicate with their children, and I believe parents' interaction with the child is crucial to set up their moral standards and worldviews.

I think the next thing that comes after this research should be to investigate children's behavior during the storytelling or conversation with the AI Interactive Narrative System, and how such behaviors correspond to childrens' emotions and preferences. Also, the effectness of AI storytelling should be compared with the interaction with the parents in future research.

The second paper I chose is *Dialogue Act Classification with Context-Aware Self-Attention* by Vipul Raheja and Joel Tetreault. This paper leverage the effectiveness of a context-aware self-attention mechanism coupled with a hierarchical recurrent neural network. Below are a summary and keynotes of the paper.

1. Self-attentive representations encode a variable length sequence into a fixed size, using an attention mechanism that considers different positions within the sequence.
2. Text Classification and sequence labeling are two widely-used examples that use self-attentive mechanism.
3. The word representation layer is followed by a bidirectional GRU (Bi-GRU) layer. Concatenating the forward and backward outputs of the Bi-GRU generates the utterance embedding that serves as input to the utterance-level context-aware self-attention mechanism, which learns the final utterance representation.

Here is what I found interesting about this paper:

1. This paper evaluates the classification accuracy of our model on the two standard datasets used for DA classification: the Switchboard Dialogue

Act Corpus (SwDA) (Jurafsky et al., 1997) consisting of 43 classes, and the 5-class version of the ICSI Meeting Recorder Dialogue Act Corpus (MRDA) introduced in (Ang et al., 2005).

2. This paper uses a ELMO, which is a method a bidirectionary LSTM model.
3. The context-aware self attention mechanism significantly outperforms earlier models on the commonly-used SwDA dataset and is very close to state-of-the-art on MRDA.

Some potential weaknesses are:

1. The self-attention mechanism cannot address the problem when one single word or phrase has multiple different meanings.
2. The self-attention mechanism cannot address the problem of punctuation.

As addressed in the paper, the next thing that comes after this research should be experimenting with other attention mechanisms such as multihead attention directional self-attention, block self-attention, or hierarchical attention. More research should be conducted on words with different meanings and punctuations in texts.

QUESTION 4

Do humans have free will? To answer this question, we must first address what is free will. While the dictionary states that free will is the ability to act at one's own discretion, there are many ways people define free will philosophically or theologically. In this article, I will use Christian List's argument, where an agent must have three capacities to have free will:

1. The capacity to act intentionally;
2. The capacity to choose between alternative possibilities;
3. The capacity to control one's actions.

First, humans have the capacity to act intentionally. This is an obvious fact that governs our society today. Almost all actions made by human are driven by intention or goal-oriented. For example, the reason we acquire education is to

acquire more and better job opportunities, making more money or become famous, and have a better life in the future. The reason why we chose Tech to persue our degrees is because it is a high-ranked and reputated research institute, and it has the most cutting-edge technologies in the world. Even some simple everyday tasks are intentional. For example, we eat because we need energy; we drink because we are thirsty; we sleep because we need rest. Therefore, humans are rational beings who acts intentionally all the time.

Second, humans have the capacity to choose between alternative possibilities. continuing the previous discussion, we chose to join Tech among all the universities which admitted us because we believe Tech is better. Not only humans have the ability to choose, animals, plants, and almost all living things have the ability to choose between alternatives. For example, plants tend to grow toward the direction of more sunlight. Animals tend to choose more neutrious food or food that provides more energy if available. This can explain the phenomenon that why more people prefer meat to vegetables because meat provides higher energy.

Third, humans have the capacity to control his/her actions because it is an obvious fact that we can do whatever we want to do. Our brain controls our actions. Therefore, humans have free will.

However, I think it is not possible for artificially intelligent agents to also exhibit free will like humans because they cannot meet one or more of the criteria listed above. I will first briefly discuss the requirements it meets, then discusses what it cannot satisfy.

First, I acknowledge that AI agents can have the capacity to choose between alternatives. Additionally, AI agents can perform even better than humans when trying to find the best option. For example, today's AI systems have robust searching algorithms that can find the optimal solution to reach the goal within milliseconds given several alternative paths. Therefore, AI agents have the ability to choose wisely.

Second, AI agents can control their actions. The CPU of the AI agent acts like the brain, and the code is like the instructions. If the AI agent is programmed correctly, the CPU will adequately control the AI agent's action even when

surprising events occur. Just like how the human brain controls human actions, the CPU of an AI agent controls the agent's action.

However, I think AI agents lack the ability to act intentionally. AI agents must be given some tasks or goals in order to choose and perform actions. Without an initial goal, AI agent would halt. Some may argue that AI agents may be programmed to have random goals if no tasks are given. However, there is no true randomness in an AI program. All random numbers are pseudorandom. Additionally, the generation of random goals is meaningless because it is not beneficial to the agent itself. Comparing to humans, we act intentionally for our own sake. Therefore, AI agents cannot act intentionally.

REFERENCES

Brain, The. "How Do Neurons Communicate (so Quickly)?" MIT McGovern Institute, 21 Oct. 2019, mcgovern.mit.edu/2019/02/28/ask-the-brain-how-do-neurons-communicate/.

"Can You Grow New Brain Cells?" Queensland Brain Institute, 20 Dec. 2017, qbi.uq.edu.au/blog/2017/11/can-you-grow-new-brain-cells.

Reynolds, Matt. "Carpenter Ants 'Throw up' on Each Other to Say Hello." WIRED UK, WIRED UK, 29 Nov. 2016, www.wired.co.uk/article/ants-throwing-up-hormones-development.

Dieter Vanderelst and Alan Winfield, 2018, The Dark Side of Ethical Robots.

Richard Kim Max Kleiman-Weiner Andres Abeliuk Edmond Awad Sohan Dsouza Josh Tenenbaum Iyad Rahwan, 2017, A Computational Model of Commonsense Moral Decision Making.

Ben Burtenshaw, Nov. 2020, AI Stories: An Interactive Narrative System for Children.

Vipul Raheja, Joel Tetreault, 2019, Dialogue Act Classification with Context-Aware Self-Attention.