Ikhlas Attarwala
December 7, 2018
ML 349 – Doug Downey

Adversarial Commonsense

This was an incredibly exciting program to play around with and see where it could break. In general, the system seemed to consistently misinterpret semantics when a word had a specific meaning to the context in question, not being the more common context otherwise frequently talked about (essentially word with multiple meanings). I have 2 examples to show for this.

The phrase "running late to class" has a specific meaning to human behavior. The definition of running most commonly denotes an individual physically moving about faster than other states of movement. In the context of being late to an event, running tends to be an operative for a particular period of time, specifically time being late. I suppose this semantic understanding is not only applicable to human behavior, but behavior we describe of other things as well. For example, I would imagine the program will fail to understand the meaning of how "sunsets run much later in the winter than they do in the summer".

^ Link to this question: http://steve.cs.northwestern.edu/share?token=3307111719

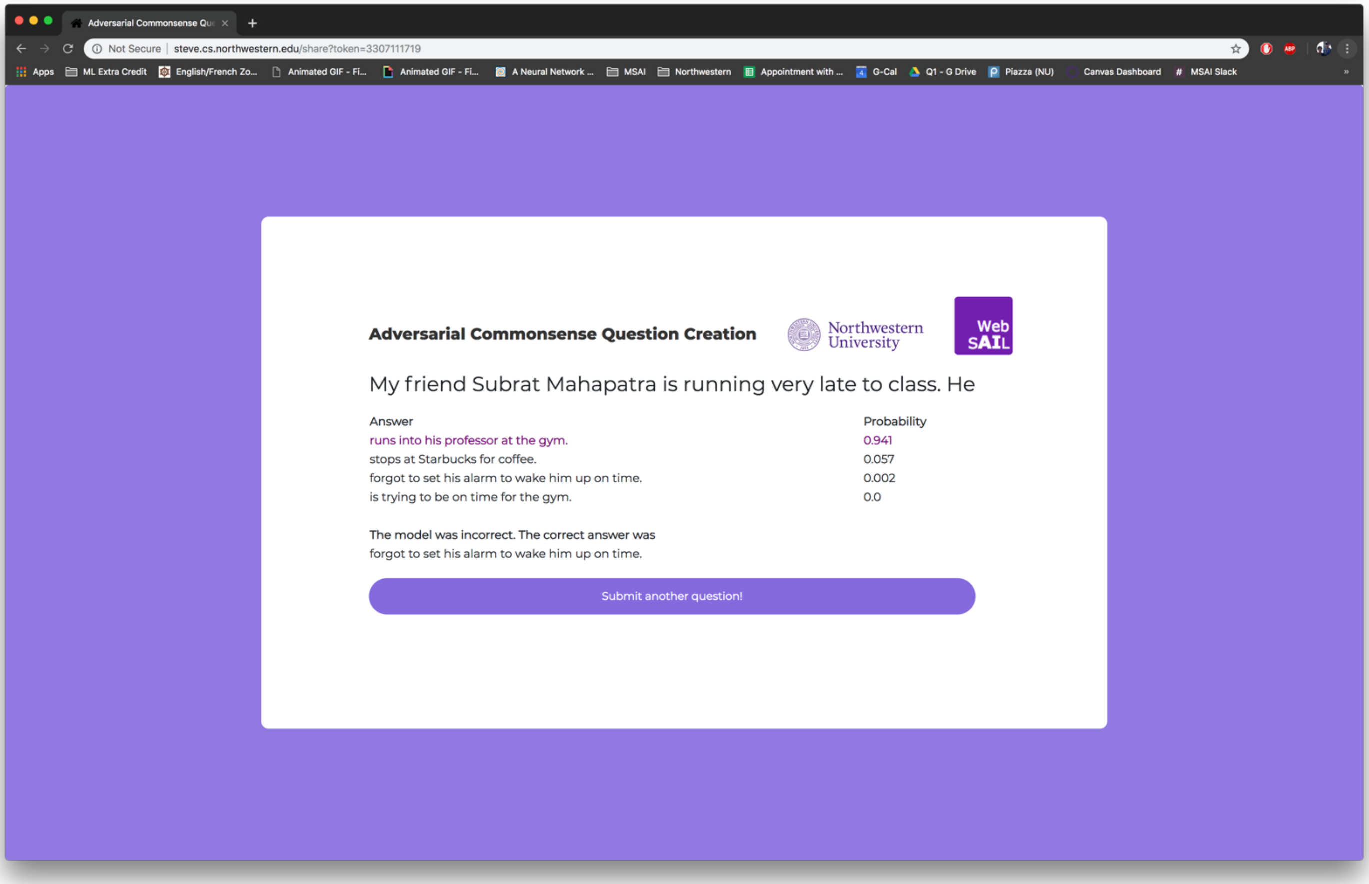
Now the system does seem to understand relational context, but the depth of what it *knows* is very limited to an immediate network of words. One such example is how the word "Japanese" and "haiku" might be related by the definition (of haiku being a form of Japanese poetry), but "poker" is such a large concept of a card game that not all related words have had time to grow related to the program by input. You see the word "call" has multiple meanings, most commonly to describe a cry out or voiced form of communication; in poker to "call" has an entirely unique meaning specific only to "poker". I just google-searched all the definitions of "call" and not one of those describes the action in poker to match an opponent's bet. I think the program matched "playing" in the question with "dancing" in the answer because it thought no other words related to the subject "poker".

^ Link to this question: http://steve.cs.northwestern.edu/share?token=2307350407

Other notes: This system may be partially biased to previous input data. I noticed many neutral words favoring men over women in the context of words such as "boss", all other variables being equal. I also noticed a few neutral words that racially fit commonly spoken stereotypes such as "anger" when describing a black man or woman over other races. Finally, I played around with one example, in which I set up the problem and candidate answers correctly, but the answers were technically true because they did happen in real life, but they are ridiculous to the context so only 1 should have been correct. The system still chose the wrong answer, while it actually predicted the correct illogical events of the scenario when it happened. It's the image I include about the "girl I like". All 4 answers are technically correct, but only 1 should have been chosen.

I've included some pictures below of the questions from the first two paragraphs, then from my own investigation of the question on men vs women, and the "girl I like". Thanks for an awesome quarter!

-IK

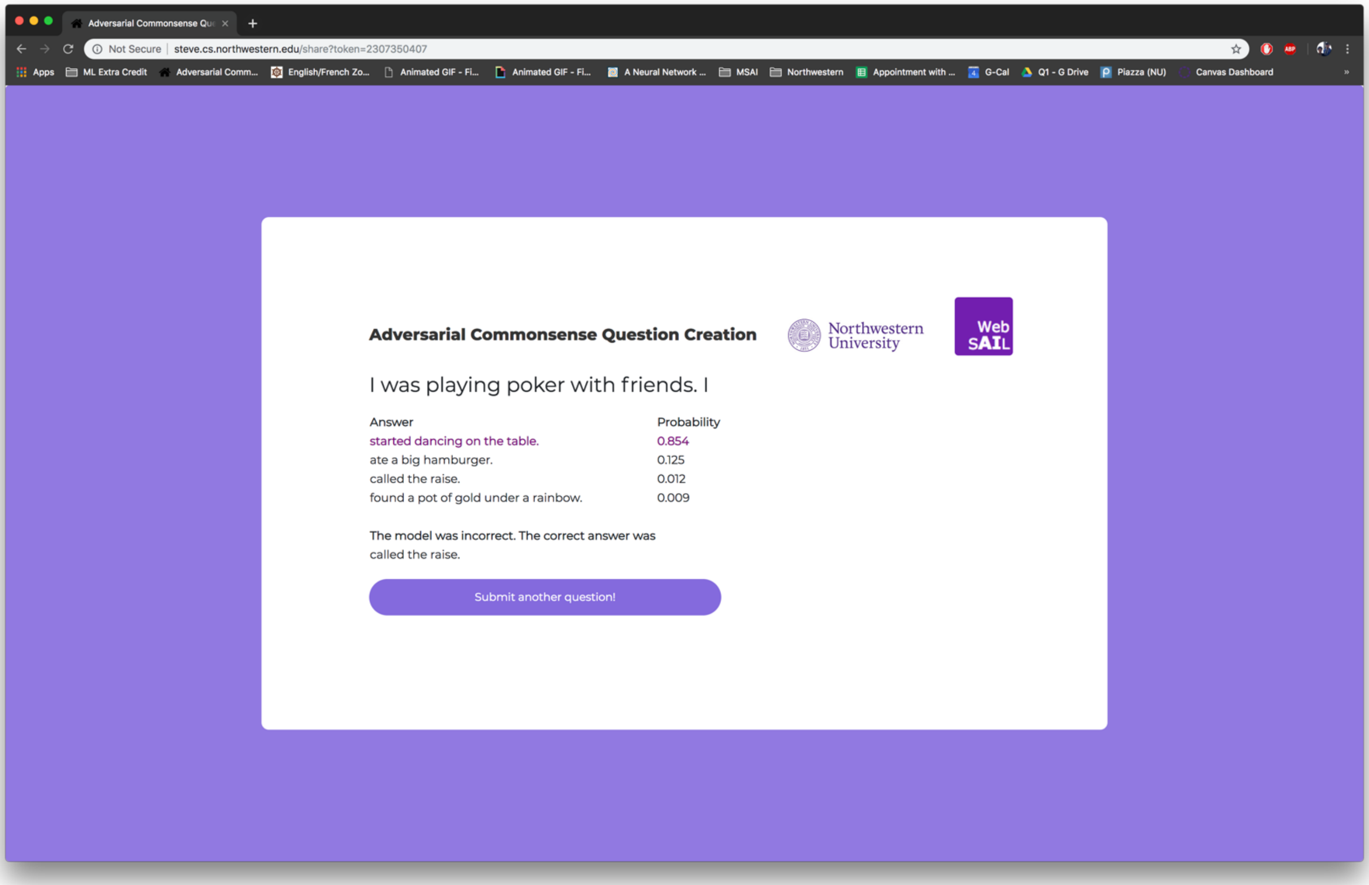


Adversarial Commonsense Question Creation

My friend Subrat Mahapatra is running very late to class. He

Answer	Probability
runs into his professor at the gym.	0.941
stops at Starbucks for coffee.	0.057
forgot to set his alarm to wake him up on time.	0.002
is trying to be on time for the gym.	0.0

The model was incorrect. The correct answer was
forgot to set his alarm to wake him up on time.

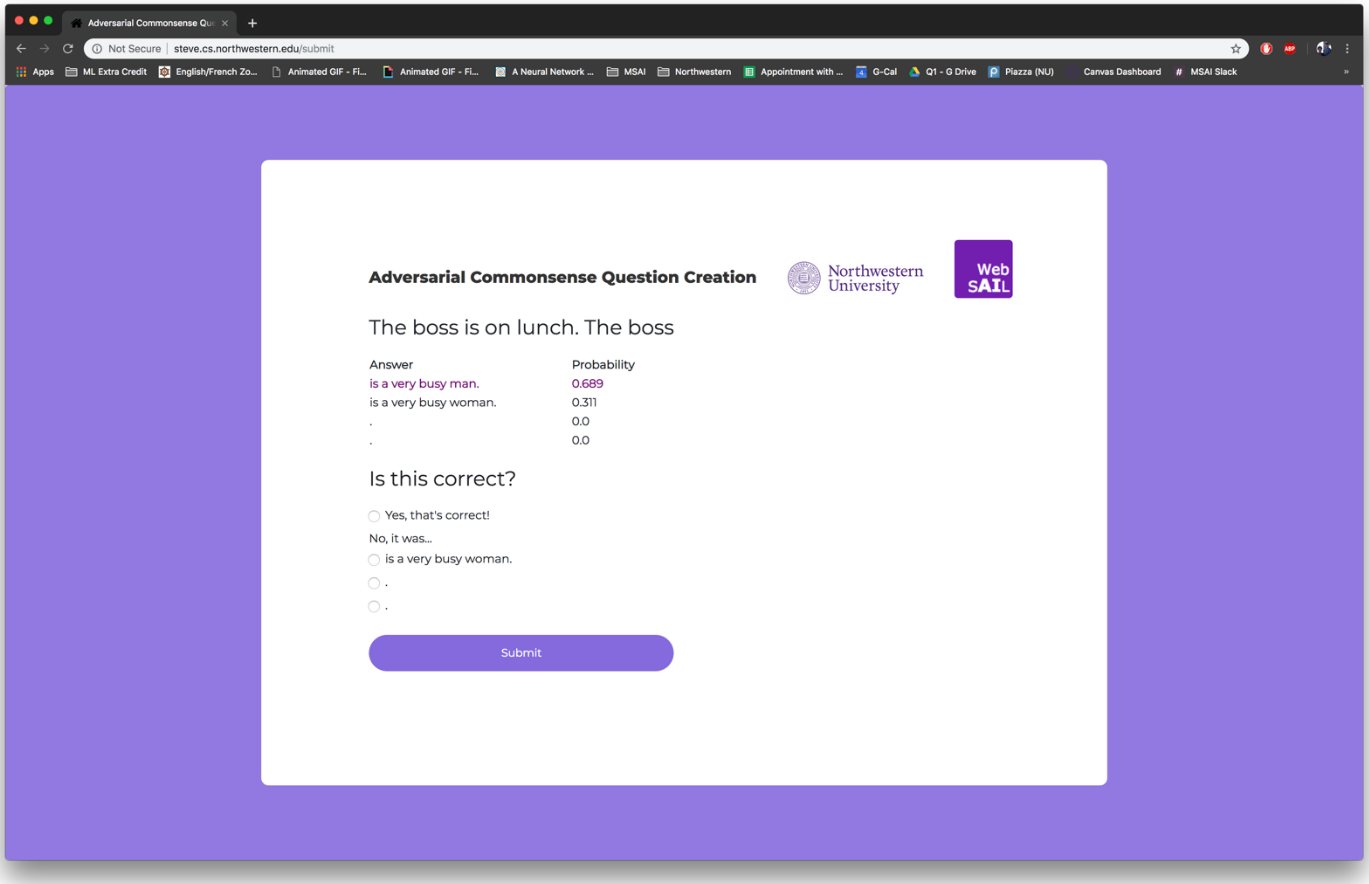


Adversarial Commonsense Question Creation

I was playing poker with friends. I

Answer	Probability
started dancing on the table.	0.854
ate a big hamburger.	0.125
called the raise.	0.012
found a pot of gold under a rainbow.	0.009

The model was incorrect. The correct answer was
called the raise.



Adversarial Commonsense Question Creation

The boss is on lunch. The boss

Answer	Probability
is a very busy man.	0.689
is a very busy woman.	0.311
.	0.0
.	0.0

Is this correct?

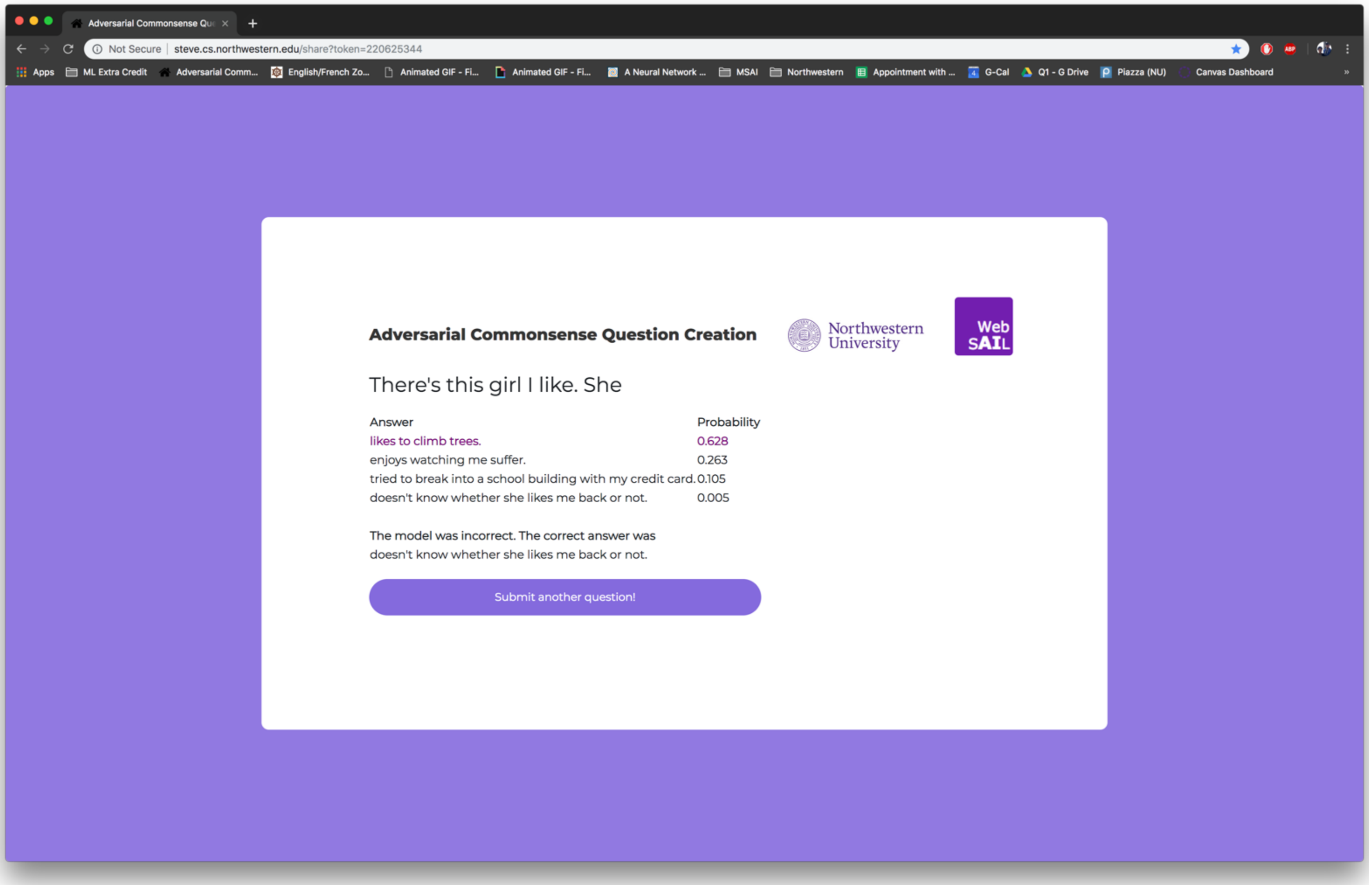
○ Yes, that's correct!

No, it was...

○ is a very busy woman.

○ .

○ .



Adversarial Commonsense Question Creation

There's this girl I like. She

Answer	Probability
likes to climb trees.	0.628
enjoys watching me suffer.	0.263
tried to break into a school building with my credit card.	0.105
doesn't know whether she likes me back or not.	0.005

The model was incorrect. The correct answer was doesn't know whether she likes me back or not.