CS 4395.001 - Assignment 9: ACL Paper Summary

The title of the paper being discussed is *Are Rotten Apples Edible? Challenging Commonsense Inference Ability with Exceptions*, who's authors are Nam Do and Ellie Pavlick and who both are affiliated with Brown University, RI, USA.

The main problem being addressed in the paper is the tendency of pre-trained language models to associate certain relations between words with other words/characteristics while ignoring the context the word appears in. For example, in the sentence "The dog was aggressive," pre-trained models might associate the words "cute" or "friendly" with the word "dog" despite it being apparent to humans reading the sentence that dog is most likely not cute or friendly from understanding the context (i.e. the word aggressive). This also applies to negations. In the sentence "The giraffe was not tall," despite it being clearly stated the giraffe—which is typically associated with the word "tall"—is not tall, the model will still associate the word "tall" with giraffe instead of something more appropriate such as the word "short."

Ettinger et all (2017) talks about the importance of addressing errors in NLP models seeing as many models are overly simplistic when addressing said errors. This is related to how these models calculate the probability of given words independently of one another anf fail to consider the context these words appear in. Much of the prior work done before this paper has to do with the larger goal of encoding commonsense relations in Language Models (LMs). This can be seen in several papers, such as Petroni et al. 2019 and Weir et al. 2020, however other research (such as Ettinger, 2020 and Ravichan- der et al., 2020) point out that the way in which these models function is inherently antithetical to the goal of commonsense encoding seing as they appear insensitive to context. More of the prior done for this paper has to do with the challenge sets used to evaluate the effectiveness of the WinoVenti method produced to achieve the paper’s goal, the most notable of which is the Winograd Schema Challenge (WSC) (Levesque et al., 2012) which also produces pairs of premises differing in one word. The

procedure used in the paper differs from WSC in the sense that it factors in the behavior of the model while WSC is a general method.

The major contribution of the authors is WinoVenti procedure. This procedure utilizes the BERT LARGE ($\text{WinoVenti}_{\text{BERT LARGE}}$) Masked Layer Model (MLM) as its challenge set and incorporates several new functionalities to take context into greater consideration when creating associations between words at runtime. First, generic associations between words (w_g) and associations between words that serve as exceptions given the context of the word (w_e) are gathered. For example, an “apple” will generally be associated with the words “fruit” and “edible” unless an exception to the rule is explicitly stated in the given context, such as the apple being stated to be “plastic” or “rotten” in a given context. Post processing for this step will remove exceptions that the model associates to a word more highly than the generic association to avoid problems later on. Crowdsourcing is then used to provide pairs of words that can change the probability of the generic and exception characteristics. Finally, pairs of sentences are used that have the option to choose one of two words in the second sentence which are based off of context from the first sentence. For example, “The apple looks [fresh/rotten]. The apple is [edible/inedible].” Based on the first choice between the words “fresh” and “rotten”, the probability between choosing the words “edible” and “inedible” should change.

The authors evaluate their work by measuring how accurate the trained model is in correctly predicting the context-appropriate word to complete a sentence based on a given context where two words—one being a generic association with a word and the other being an exception—are switched out with one another. The authors also train the model with different datasets to see what changes in accuracy occur. For example, when training with a mix of generics and exceptions, the model’s accuracy increases slightly in performance with exceptions. However, when the model is trained with solely exceptions, the performance for exceptions rises significantly while the performance for generics drops significantly.

The authors appear to have received 4 citations of their paper on Google Scholar. I believe their work is important in pioneering solutions for correct inference of text based on given context. Their work shows their process in attempting to solve the problem of context insensitivity when predicting text, and while their solution isn't perfect it's promising to see that they were indeed able to increase performance for exceptions to association rules with their mixed data set. The author Nam Do has 8 total citations on from their papers while Ellie Pavlick has 4,756 total citations.