COLLOQUIUM REPORT

“SQUIRREL OR SKUNK ? NLP MODELS FACE THE LONG TAIL OF LANGUAGE”

SPEAKER’S NAME: NATHAN SCHNEIDER(Asst. Prof., Dept. of Linguistics and Computer Science, Georgetown University)

DATE : 3rd September, 2021

NAME : Agasthya Harekal

The talk was mainly about the performance of various NLP models especially BERT based models, on detection of “long-tail” phenomenon in language. Also, probabilistic tagging of words into syntactic categories was done in order to conduct experiments on calibration and re-calibration of results. Initially, there was mention of some words that have rare senses, such as ,“if” where some usage might have very rare sense/meaning where truth of first proposition is not affected by condition. There was also discussion of quantity modifiers of items that have very rare sense. This becomes a challenge in supervised machine learning where the detection of more rare sense/meaning becomes difficult.

NLP models detect the long-tail phenomenon where tails provide the right meaning which the speaker considers as squirrel’s tail whereas other tails that have the wrong sense/meaning or turned out to be skunk’s tail. BERT model was selected to detect the sense of a word in sentences and compare with other models. The BERT model is considered a very good model to detect the rare word senses. BERT model was trained to retrieve instances that are similar to the given word in the sentence , where co-sine similarity was used. If this approach worked well, it was to be considered that BERT has good geography of contextualised embedding over instances when compared to human evaluation. The process of evaluation is to get embedding vector from BERT for the required word for the particular instance and get embedding vectors for all instances of the word in the training data and extract the BERT vectors and rank vectors based on the co-sine similarity.

After experiments, it was found that BERT is able to distinguish the senses quite accurately, compared to other larger models such as Roberta-base, xlnet-base-cased, gpt2. BERT CWE(Contextualized Word Embedding) similarity rankings for sentence s with a particular word having same sense were found to be more precise compared to CWE rankings of other models. This might be due to the way the model is trained and the organisation of model. All models considered for comparison were generic models. It was concluded that retrieving word sense match of a query word in context sentence is very hard if the sense is rare.

Next part of the lecture was based on calibration. For this part, probabilistic tagging was done where syntactic tags were assigned to each word with certain probability. The tags designate the words to various syntactic categories .Using two datasets, evaluation was done of the trust-worthiness of the model. The model is evaluated to give over-confident, under-confident and well-calibrated results. The model for forecasting the rain was evaluated here. The model that is overshooting the prediction is set to be over-confident or if it is undershooting its prediction, it is set to be under confident and the model is well-calibrated otherwise. Using training data and test data accuracy of probability prediction was evaluated. Also, Re-Calibration of the model was considered, that adjusts the probability prediction of the model to be better calibrated model. Re-calibration caused decrease in error for both the datasets. The calibration error was considered for frequent tags, moderately rare tags, very rare tags and evaluation was done and graph was plotted. It was found that rarest of tags have more mis-calibration than others based on graph analysis. Re-calibration was done to decrease error in rare tags by taking tag frequency into consideration. For tagging purposes, Tree based TreeRNN model and AddrMLP model offer better generalisation to unseen/rare tags compared to conventional non-constructive and sequentially constructive models for tagging.

Hence using BERT based model, experiments proved that long-tail phenomenon or the rare sense of words could be accurately predicted when compared to other models. The BERT model is a very good transformer based model for next sentence prediction, Natural Language Generation where sentences can be generated from meaning representations and many other tasks. BERT uses a transformer based model, that understands contextual relations of words in a sentence. It uses a Masked Language Model where fifteen percent of the words of a particular sequence is replaced by MASK token. Then, the based on the context of other words of the sequence, the model goes on to predict the value of masked words, which makes it an accurate model for prediction. It has performed very well on GLUE(General Language Understanding Evaluation) task set, SQUAD(Stanford Question Answering Dataset) , SWAG(Situations with Adversarial Generations). Hence it is a very apt model to be used for evaluation of long-tail phenomenon in language.