0 and 1. For example, if ground truth is "He is playing chess in the backyard" and output sentences are S1: "He is playing tennis in the backyard", S2: "He is playing badminton in the backyard", S3: "He is playing movie in the backyard" and S4: "backyard backyard backyard backyard backyard backyard backyard". The score of S1, S2 and S3 would be 6/7,6/7 and 6/7. All sentences are getting the same score though information in S1 and S3 is not same. This is because BELU considers words in a sentence contribute equally to the meaning of a sentence which is not the case in real-world scenario. Using combination of uni-gram, bi-gram and n-grams, we can to capture the order of a sentence. We may also set a limit on how many times each word is counted based on how many times it appears in each reference phrase, which helps us prevent excessive repetition.

b) GLUE (General Language Understanding Evaluation) score: Previously, NLP models were almost usually built to perform effectively on a unique job. Various models such as LSTM, Bi-LSTM were trained solely for this task, and very rarely generalized to other tasks. The model which is used for named entity recognition can perform for textual entailment. GLUE is a set of datasets for training, assessing, and comparing NLP models. It includes nine diverse task datasets designed to test a model's language understanding. To acquire a comprehensive assessment of a model's performance, GLUE tests the model on a variety of tasks rather than a single one. Single-sentence tasks, similarity and paraphrase tasks, and inference tasks are among them. For example, in sentiment analysis of customer reviews, we might be interested in analyzing ambiguous reviews and determining which product the client is referring to in his reviews. Thus, the model obtains a good "knowledge" of language in general after some generalized pre-training. When the time comes to test out a model to meet a given task, this universal "knowledge" gives us an advantage. With GLUE, researchers can evaluate their model and score it on all nine tasks. The final performance score model is the average of those nine scores. It makes little difference how the model looks or works if it can analyze inputs and predict outcomes for all the activities.

Considering these metrics in mind, it helps to evaluate the performance of an NLP model for a particular task or a variety of tasks.

5.2 Challenges

The applications of NLP have been growing day by day, and with these new challenges are also occurring despite a lot of work done in the recent past. Some of the common challenges are: Contextual words and phrases in the language where same words and phrases can have different meanings in a sentence which are easy for the humans to understand but makes a challenging task. Such type of challenges can also be faced with dealing Synonyms in the language because humans use many different words to express the same idea, also in the language different levels of complexity such as large, huge, and big may be used by the different persons which become a challenging task to process the language and design algorithms to adopt all these issues. Further in language, Homonyms, the words used to be pronounced the same but have different definitions are also problematic for question answering and speech-to-text applications because they aren't written in text form. Sentences using sarcasm and irony sometimes may be understood in the opposite way by the humans, and so designing models to deal with such sentences is a really challenging task in NLP. Furthermore, the sentences in the language having any type of ambiguity in the sense of interpreting in more