Natural Language Processing

Project Presentation

NATURAL LANGUAGE INFERENCE USING LOGISTIC REGRESSION AND BERT MODELS

TEAM MEMBERS :

1) Siddharth Pullakhandam

2) Naralachetty Navitha

ABSTRACT :

This is a task based on Natural Language Inference(NLI). We are given two sentences for this task, referred to as the premise and the hypothesis. If the "premise" is correct, then we must decide whether the "hypothesis" is true (entailment), untrue (contradiction), or indeterminate (neutral). We used the Stanford Natural Language Inference (SNLI) dataset for this experiment.

INTRODUCTION :

Natural Language Inferencing (NLI) task is one of the most important subsets of Natural Language Processing (NLP) which has seen a series of development in recent years. There are standard benchmark publicly available datasets like Stanford Natural Language Inference (SNLI) Corpus, Multi-Genre NLI (MultiNLI) Corpus, etc. which are dedicated to NLI tasks. Few state-of-the-art models trained on these datasets possess decent accuracy. Natural language inference studies whether a hypothesis can be inferred from a premise, where both are a text sequence. In other words, natural language inference determines the logical relationship between a pair of text sequences.

Such relationships usually fall into three types:

- Entailment: the hypothesis can be inferred from the premise.
- Contradiction: the negation of the hypothesis can be inferred from the premise.
- Neutral: all the other cases.

TOOLS AND DATASET :

- Programming Language – Python
- Libraries – Tensorflow

- Dataset :

Stanford Natural Language Inference (SNLI) dataset :

Stanford Natural Language Inference (SNLI) Corpus is a collection of over 500000 labeled English sentence pairs (Bowman *et al.*, 2015).

MODELS : LOGISTIC REGRESSION

- Logistic Regression classifier using TF-IDF features :

Logistic regression model was trained using TF-IDF (Term Frequency-Inverse Document Frequency) features obtained using sklearn python library. The feature vector used to train the model is obtained by concatenating the TF-IDF vectors of premise and hypothesis. The model is trained (fit) using L-BFGS (Limited memory - Broyden–Fletcher–Goldfarb–Shanno algorithm) solver with a maximum iteration limit of 1000.

MODELS : BERT

- BERT stands for Bidirectional Encoder Representations from Transformers.
- BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.

SIMPLE EXAMPLES :

PRECISE	RELATION	HYPOTHESIS
A turtle danced	Entails	A turtle moved.
Turtle	Contradicts	Linguist
Every reptile danced	Neutral	No turtle moved.

EXPECTED RESULTS :

Model	Accuracy
Logistic regression	63.38%
BERT	98.02%

FUTURE SCOPE :

REFERENCES :

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding