Natural Language Inferencing for Hindi

Mid Project discussion

P S V N Bhavani Shankar, 22M0743
Sanjeev Kumar, 214050008
Pavan Kumar Yalavarthi, 22M0777

November 8

Problem Statement

- We aim to implement a Natural Language Inferencing task for Hindi Text.
- Natural language inference is the task of determining whether a “hypothesis” is true (entailment), false (contradiction), or undetermined (neutral) given a “premise”.
- We plan to implement a GRU and an LSTM for the task and also compare our models with BeRT which we will finetune for this task.

Related works

- Aggarwal, D., Gupta, V., & Kunchukuttan, A. (2022). IndicXNLI: Evaluating Multilingual Inference for Indian Languages. *arXiv*. https://doi.org/10.48550/arXiv.2204.08776

- With the introduction of IndicNLPSuite (Kakwani et al., 2020) by AI4Bharat there has been has an increased interest and effort towards the research for Indic languages model.

Related Works (contd.)

- Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv*, 2018, https://doi.org/10.48550/arXiv.1810.04805

Dataset

- INDICXNLI Dataset
- INDICXNLI is like existing XNLI dataset in shape/form but focusses on Indic language family.
- INDICXNLI include NLI data for eleven major Indic languages that includes Assamese, Gujarat, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, Hindi, and Bengali.

Dataset

- It uses the XNLI English dataset(premises and hypothesis) and translate it into 11 Indic language using IndicTrans (Ramesh et al., 2021).
- We are only focusing on Hindi Language in the dataset.
- Train : 392,702 instances
- Validation: 2490 instances
- Test: 5010 instances

https://huggingface.co/datasets/Divyanshu/indicxnli

Sample Data

premise (string)	hypothesis (string)	label (class label)
"अवधारणात्मक रूप से क्रीम स्किमिंग के दो बुनियादी आयाम हैं-उत्पाद और भूगोल।"	"उत्पाद और भूगोल क्रीम स्किमिंग का काम करते हैं।"	1 (neutral)
"आप मौसम के दौरान पता है और मुझे लगता है कि अपने स्तर पर उह आप उन्हें अगले स्तर पर खो देते हैं अगर वे मूल टीम को याद करने का फैसला करने के लिए बहादुर ट्रिपल ए से एक आदमी को…	"अगर लोग याद करते हैं तो आप निम्नलिखित स्तर पर चीजें खो देते हैं।"	0 (entailment)
"हमारी संख्या में से एक आपके निर्देशों का बारीकी से पालन करेगी।"	"मेरी टीम का एक सदस्य आपके आदेशों को बहुत सटीकता के साथ निष्पादित करेगा।"	0 (entailment)
"तुम्हें कैसे पता? यह सब फिर से उनकी जानकारी है."	"यह जानकारी उनके पास है।"	0 (entailment)
"हाँ मैं आपको बताती हूँ कि अगर आप उन टेनिस जूतों में से कुछ की कीमत तय करते हैं तो मैं देख सकती हूँ कि क्यों अब आप जानते हैं कि वे सौ डॉलर की रेंज में उठ रहे हैं"	"टेनिस जूतों की कीमतों की एक श्रृंखला है।"	1 (neutral)
"मेरा वॉकमैन टूट गया तो मैं अब परेशान हूँ.... मुझे बस स्टीरियो को असली जोर से चलाना है"	"मैं परेशान हूं कि मेरा वॉकमैन टूट गया और अब मुझे स्टीरियो को जोर से चलाना पड़ रहा है।"	0 (entailment)
"लेकिन कुछ ईसाई मोज़ेइक एप्स के ऊपर जीवित रहते हैं-शिशु यीशु के साथ वर्जिन, दाईं ओर प्रधान स्वर्गदूत गैब्रियल (उनके साथी माइकल, बायीं ओर, अपने पंखों से कुछ पंखों को छोड़कर गायब हो ग…	"अधिकांश ईसाई मोज़ेइक मुसलमानों द्वारा नष्ट कर दिए गए थे।"	1 (neutral)
"(स्लेट के जैक्सन के निष्कर्षों को पढ़ने के लिए पढ़ें।)"	"जैक्सन के निष्कर्षों पर स्लेट की राय थी।"	0 (entailment)

Architecture Plan

- We will use FastText Model trained on IndicNLI corpus for the sake of word embeddings
- The model will have two input layers, one for premise and another for hypothesis followed by an embedding layer which converts inputs to embeddings.
- This is given as input to LSTM or GRU layers followed by Dense layers which would output a classification label.

Metrics to be used

- Precision
- Recall
- F Score