TELUGU PARAPHRASE DETECTION USING SIAMESE NETWORK

M Rohith; M Jaswanth Venkat; P Venkata Akhil; Deepa Gupta
Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.

Abstract

- Paraphrasing methods will generate, identify or extract the phrases or sentences that convey the same meaning.
- The ability to detect similar sentences is crucial for several applications, such as text summarization, and plagiarism detection.
- In our project, we will identify a given pair of Telugu sentences if they are Paraphrased or Non-paraphrase.
- We have used a Siamese architecture for calculating the amount of similarity between pairs of sentences.

Introduction

- Paraphrases are phrases which convey a similar meaning are but used with different wording.
- Paraphrase detection is an NLP Classification Problem. In our model, we are identifying paraphrases for the Telugu language.
- we have created our dataset which contains 1032 pairs of sentences, out of which 516 are paraphrase pairs and 516 are non-paraphrase pairs
- The sentences are selectively chosen from the available Telugu corpus and some Telugu articles.
- To create a paraphrase pair the corresponding paraphrase is generated by using the paraphrasing tool and crosschecked by us.
- And for non-paraphrase pairs, the second sentence is selected randomly such that it is not in a similar meaning and labelled accordingly.

Here are a few examples of Paraphrase and Non-Paraphrase

నేను అబద్ధం చెబుతున్నాను.
I am Lying
నేను నిజం చెప్పడం లేదు.
I am not saying the Truth
This is a paraphrased pair

మీ ఆరోగ్యం త్వరలో కుదుట పడాలని కోరుకుంటున్నాను
I want you to recover soon
మీ ఆరోగ్యం వెంటనే క్షీణించాలని కోరుకుంటున్నాను
I want your health to get deteriorated
This is a Non-paraphrase pair

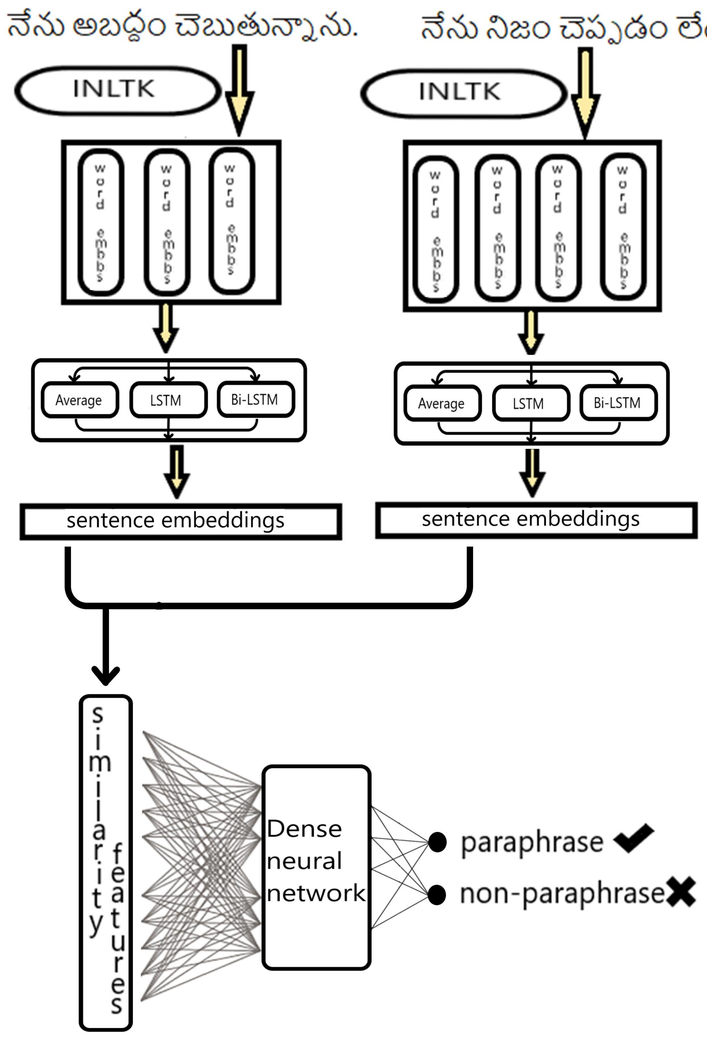
	Training data	Avg no of words	Test dataset	Avg no of words
Paraphrase	417	4.86	113	5.53
Non-paraphrase	407	5.87	94	5.02
Total	824	5.36	207	5.3

Table 1: Information about the dataset which we created

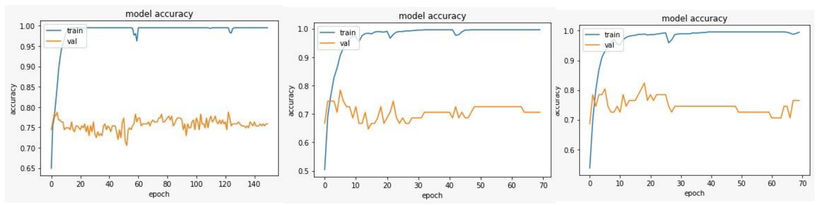
Methods and Materials

- Siamese architecture is used for our project, in which both the sentences go through the same processing and extraction before comparison.
- In the model an embedding layer is created to extract embeddings of a given word, the layer requires an embeddings file to fetch the embeddings.
- We used INLTK embeddings for Telugu to get to create the embeddings file for our dataset.
- Then we are calculating sentence embedding from word embeddings using LSTM, and Bi-LSTM and averaging the word embeddings.
- Then the sentence embeddings are sent as input for similarity feature extraction.
- The similarity features are extracted from the sentence embeddings and passed to a fully connected network to obtain the similarity score.

Proposed Model Architecture



Results



Average Model Performance | LSTM Model Performance | BI- LSTM Model Performance

Chart 1. Performance Measure for our Models.

Discussion

- The model is limited to a small dataset and can be extended further. With large and more cleaned data the models will definitely tune better than the current results.
- we have created three models and compared their performance for our dataset The Bi-LSTM model performs better compared to others.

Conclusions

- In the proposed work, the main aim is to identify the paraphrases from the dataset. We used the INLTK sentence embedding tool to calculate the sentence embedding for our dataset.
- We use the Siamese Network, LSTM and Bi-LSTM for calculating the similarity score between both the sentences and predict whether it is Paraphrase or Non-Paraphrase.
- For Indian Languages paraphrase detection is very less explored so we can explore more and try with another Indian language for the paraphrase Detection.

References

1. El Desouki, M. I., & Gomaa, W. H. (2019). Exploring the Recent Trends of Paraphrase Detection. International Journal of Computer Applications, 975, 8887.
2. A., Hwang, Y.-S., Sumita, E.: Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005), pp. 17–24 (2017)
3. Deepa Gupta ,Vani K, ASE@DPIL-FIRE2016: Hindi Paraphrase Detection using Natural Language Processing Techniques & Semantic Similarity Computations
4. D. Aravinda Reddy, M. Anand Kumar and K. P. Soman, Paraphrase Identification in Telugu Using Machine Learning
5. Basant Agarwala,b,∗ , Heri Ramampiaroa , Helge Langsetha , Massimiliano Ruoccoa,c ; “A Deep Network Model for Paraphrase Detection in Short Text Messages” In Information Processing & Management Journal (IPM), 54(6), pp. 922-937. Elsevie
6. AL-Smadi, Mohammad, Jaradat, Zain, Al-Ayyoub, Mahmoud, Jararweh, Yaser, “Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features”
7. Chi, Xiaoqiang, Yang Xiang, and Ruchao Shen. "Paraphrase Detection with Dependency Embedding." 2020 4th International Conference on Computer Science and Artificial Intelligence. 2020.
8. Yuan, Zhao, and Sun Jun. "Siamese Network cooperating with Multi-head Attention for semantic sentence matching." 2020