CSE424/CSE713
Pattern Recognition

Course Instructor:

Annajiat Alim Rasel

Research Assistant:

Sania Azhmee Bhuiyan

Student Tutor:

Ehsanur Rahman Rhythm

Team-18:

23366009 Ripa Sarkar

20101188 Mirza Raiyan Ahmed

21101103 Khandker Samia Rahman Pranti

20101557 Mohammod Tareq Aziz Justice

Project Title

Develop a possible dataset for Bangla locution using ML and NLP techniques

Table of Contents:

Introduction

- Locution refers to the study of speech sounds in a language.
- Developing a dataset for Bangla locution can aid in various applications, such as speech recognition, text-to-speech synthesis, and language learning.
- ML and NLP techniques can be utilized to create a comprehensive dataset.

Literature Review

Paper 1

Analysis of Bangla Transformation of Sentences Using Machine Learning.

The researchers developed a method to correctly identify a sentence (sorol, jotil, jougik). The model accepts a Bangla sentence, detects the sentence structure type, and outputs the sentence type. The most popular and well-known six supervised machine learning methods were used to classify three forms of phrase formation: Sorol Bakko (simple sentence), Jotil Bakko (complex sentence), and Jougik Bakko (compound sentence). They trained and tested their 2727-number dataset from multiple sources. Accuracy, precision, recall, f1-score, and confusion matrix were calculated from the dataset. The decision tree classifier achieves 93.72% accuracy.

Paper 2

An Emperical Framework of Idioms Translator From Bengali to English: Rule Based Approach.

This study proposes a Bengali-to-English translation framework.Parsing grammar rules are context-sensitive. The top-down algorithm parses sentences. They developed a sentence-idiom translation algorithm. The method is tested with 15000 sentences. The system performs well with 85.33% accuracy. Their paper translates idiomatic Bangla sentences to English.

Literature Review

Paper 3

UDDIPOK: A reading comprehension based question answering dataset in Bangla language.

The system uses Proper Noun, Acronym, Abbreviation, Phrase preposition, and Idiom dictionaries in addition to ordinary lexicons and various Example bases. Tables with English proper names, acronyms, abbreviations, phrase prepositions, and idioms and their Bangla translations were also employed. Idioms are translated using a literal and pattern example base.

Paper 4

ANUBAAD – A Hybrid Machine Translation System from English to Bangla.

UDDIPOK, a novel Bangla reading comprehension dataset, is presented in this work. 270 reading passages, 3636 questions, and answers from textbooks, middle and high school exams, newspapers, etc. Passages, questions, and replies make up this CSV dataset. Thus, machine learning researchers can quickly process data.

Ideas

- Data Collection and Annotation
- Data Preprocessing
- Feature Extraction
- Model Architecture
- Training and Testing

Challenges

- Limited Existing Resources
- Phonetic Variations
- Data Preprocessing

Future Work

- Accumulate More Data
- Incorporate New Techniques

Conclusion

- Developing a dataset for Bangla locution is crucial for advancing various applications in speech processing and language technology.
- ML and NLP techniques enable the creation of comprehensive datasets and robust models.
- The dataset can facilitate research, development, and innovation in Bangla speech technology.

Bibliography

[1] R. K. Das, S. S. Sammi, K. Kobra, M. R. Ajmain, S. A. Khushbu & S. R. H. Noori. (2022). Analysis of Bangla Transformation of Sentences Using Machine Learning. International Conference on Deep Learning, Artificial Intelligence and Robotics, Key Digital Trends in Artificial Intelligence and Robotics, pp 36–52.

[2] A. Khatun, , M. G. Hussain, , M. J. Islam, S. Kabir & M. Mahin. (2020). An Emperical Framework of Idioms Translator From Bengali to English: Rule Based Approach. 2020 IEEE Region 10 Symposium (TENSYMP), 5-7. DOI: 10.1109/TENSYMP50017.2020.9230738.

[3] T. T. Aurpa, M. S. Ahmed, R. K. Rifat, M. M. Anwar & A.B.M. S. Ali. (2023). UDDIPOK: A reading comprehension based question answering dataset in Bangla language. Data in Brief, 2023, Volume 47, https://doi.org/10.1016/j.dib.2023.108933.

[4] S. Naskar, D. Saha & S. Bandyopadhyay. (2004). ANUBAAD – A Hybrid Machine Translation System from English to Bangla.

Thank you